

Arabic Fake News Detection: A PRISMA-Based Systematic Review of Methods, Datasets, and Challenges

Mohammed Fikery Ghemas^{1*}, Hesham Ahmed Hassan², Wafaa Samy El-Kassas³

Faculty of Computers and Information Technology, The Egyptian E-Learning University (EELU), Giza, Egypt¹

Faculty of Computing & Artificial Intelligence, Cairo University, Giza, Egypt²

Faculty of Computers and Information Technology, The Egyptian E-Learning University (EELU), Giza, Egypt³

Abstract—Despite rapid progress in Arabic fake news detection, reported advances remain difficult to interpret because existing studies vary substantially in datasets, language coverage, model architectures, evaluation protocols, and explainability practices. To address this fragmentation, this study conducts a PRISMA-based systematic review of 75 peer-reviewed studies published between 2019 and April 2026. The methodological quality of each included study is assessed with an eight-item checklist. The review provides an integrated synthesis of datasets, data sources, annotation practices, preprocessing strategies, feature representations, classical machine learning, deep learning, transformer-based, hybrid, and emerging large language model approaches, as well as explainability methods and evaluation practices. A unified taxonomy is developed to organize the methodological evolution of Arabic fake news detection, while frequency analysis and a dataset-stratified performance comparison are used to characterize the adoption and reported performance of different model families. Besides descriptive cataloguing, the study critically evaluates the comparability and the evidential strength of the reported model performance. It shows that the apparent performance improvements cannot be interpreted independently of the characteristics of the dataset, the class balance, the coverage of dialects, the evaluation protocols, and the limited cross-dataset and cross-domain validation. These findings provide an evidence-based research map of the field and identify key priorities for standardized benchmarks, robust and reproducible evaluation, cross-dataset and cross-dialect validation, reliable explainability, and the development of more generalizable Arabic fake news detection systems.

Keywords—Systematic literature review; Arabic fake news detection; Arabic NLP; transformer models; hybrid models; text classification; Arabic datasets

LIST OF ABBREVIATIONS

Abbreviation	Full Form / Description		
AFND	Arabic Fake News Dataset	KNN	K-Nearest Neighbors
FND	Fake News Detection	DT	Decision Tree
NLP	Natural Language Processing	XGBoost	Extreme Gradient Boosting
ML	Machine Learning	CNN	Convolutional Neural Network
DL	Deep Learning	RNN	Recurrent Neural Network

LLM	Large Language Model	LSTM	Long Short-Term Memory
XAI	Explainable Artificial Intelligence	BiLSTM	Bidirectional Long Short-Term Memory
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses	GRU	Gated Recurrent Unit
RQ	Research Question	BiGRU	Bidirectional Gated Recurrent Unit
MSA	Modern Standard Arabic	BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words	AraBERT	Arabic BERT
TF-IDF	Term Frequency-Inverse Document Frequency	MARBERT	Arabic social-media BERT model
POS	Part-of-Speech	mBERT	Multilingual BERT
NER	Named Entity Recognition	CAMELBERT	Contextualized Arabic Model based on BERT
SVM	Support Vector Machine	AraELECTRA	Arabic ELECTRA model
RF	Random Forest	QARIB-BERT	Arabic BERT model for Arabic NLP
LR	Logistic Regression	AUC	Area Under the Receiver Operating Characteristic Curve
NB	Naïve Bayes	TP	True Positive
FN	False Negative	TN	True Negative
FP	False Positive		

I. INTRODUCTION

The advent of social media platforms has facilitated real-time human interactions and the rapid dissemination of news. In addition to traditional news sources, individuals increasingly rely on online platforms to express themselves and gather news from around the world [1]. Social networks generate vast amounts of data, including opinions, sports, and political news, and rumors and fake news created by Internet users [2]–[5]. While offering instant access to information, these platforms also facilitate the dissemination of unchecked data that can inadvertently mislead users. The spread of fake news on social media, often designed to maximize engagement through fear-based and persuasive language, is a growing concern that amplifies its societal impact [4], [6]. Fake content poses a unique

challenge because it can distort public perception and be exploited for political and commercial gain.

Among the most prominent challenges facing modern societies is the proliferation of Arabic fake news in the Arab world, which spreads misinformation and indirectly influences public opinion. Despite significant progress in fake news detection using natural language processing (NLP) techniques such as BERT, the Arabic language presents unique challenges [7]–[9]. Arabic is characterized by many dialects and linguistic styles in different countries and regions, and rich morphology and complex grammatical structures. These features make the understanding, analysis, and classification of Arabic texts more difficult than in many other languages. Moreover, the unavailability of high-quality Arabic linguistic resources and dedicated tools hinders the development of robust fake news detection models. Hence, there is an urgent need to develop models and frameworks that take into consideration the linguistic and cultural aspects of Arabic.

This survey examines feature extraction techniques for Arabic fake news detection that rely on linguistic and multimodal representations designed to capture the complexity of Arabic texts and media. Traditional approaches use TF-IDF, Bag-of-Words (BoW), and n-gram models to identify lexical features from Modern Standard Arabic and various dialects [10]–[12]. Morphological and syntactic features, such as part-of-speech tagging, dependency parsing, and named entity recognition (NER), enhance contextual understanding. Semantic embed-

ding techniques have dominated recent studies, with Word2Vec, FastText, and transformer-based models such as AraBERT and mBERT providing deep contextual representations of Arabic text. Several studies have employed hybrid architectures, such as CNN-LSTM and BiLSTM-BiGRU, to capture both spatial and sequential dependencies in linguistic patterns [13]–[15]. For visual and multimodal datasets, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used to extract image authenticity features and evaluate visual-textual coherence [16], [17]. Previous studies have also explored the fusion of textual and visual embeddings with metadata-based features, such as publication date, author credibility, and source reliability, to improve fake news detection performance. This combination of feature extraction strategies has substantially advanced Arabic fake news detection.

Several surveys have reviewed Arabic fake news detection from different perspectives. Some focused mainly on datasets, while others emphasized machine learning techniques, feature extraction methods, or general fake news detection, with only partial attention to Arabic. Although these surveys provide valuable contributions, many of them are limited by restricted time coverage, a small number of reviewed studies, insufficient coverage of transformer-based and hybrid approaches, limited analysis of dataset characteristics, or inadequate discussion of advanced dimensions such as metadata, temporal features, multimodal content, and explainability [8], [9], [18]–[22]. Table I summarizes The main differences between existing surveys and the present survey.

TABLE I. COMPARISON OF EXISTING ARABIC FAKE NEWS DETECTION SURVEYS

papers	Year	Period Covered	No. of Papers	Arabic Focus	Dataset Analysis	Explanation methods	Model Taxonomy	Research Gaps
Althabiti et al. [18]	2023	2019–2022	27	Yes	Limited	NO	ML/DL	Moderate
Albtoush et al. [19]	2025	2020–2024	Not reported	Yes	Moderate	Limited	ML/DL /Transformers	Detailed
Touahri & Mazroui – [20]	2024	Not specified	Not specified	Yes	Moderate	NO	ML	Some
Yousef et al. [21]	2024	2016-2023	29	Yes	Moderate	Limited	ML/DL /Transformers	Some
Muaad et al. [8]	2023	2017–2023	48	Partial	Moderate	Limited	ML/DL	Moderate
Amjad et al. [22]	2022	2015-2022	20	partial	Limited	NO	ML	Some
Sellami et al.* [9]	2026	2019–2025	Not specified	Yes	Detailed	NO	ML/DL /Transformers	Detailed
This review	2026	2019–2026	75	Yes	Comprehensive	Detailed discussion	ML / DL / Transformer / Hybrid	Detailed

a. * Published concurrently with the present review (2026); included for completeness rather than as a prior survey.

Furthermore, most previous surveys were conducted before the rapid adoption of large language models and advanced Arabic transformer architectures. Therefore, an updated and comprehensive survey is required to provide a unified view of the field and identify future research directions.

As shown in Table I, the present survey extends previous work by providing a broader and more recent synthesis of Arabic fake news detection research. Rather than treating model architectures, datasets, evaluation practices, and explainability as isolated aspects, this study provides an integrated and critical evidence-based synthesis that examines how these dimen-

sions jointly shape the interpretation of progress in Arabic fake news detection, particularly in terms of generalizability and reproducibility [8], [9], [18]–[22]. It reviews 75 studies published between January 2019 and April 2026 and analyzes the field from multiple complementary perspectives, including datasets, preprocessing, feature representation, model families, evaluation metrics, metadata, temporal signals, multimodal content, explainability, and open challenges.

The main contributions of this survey are summarized as follows:

- An integrated PRISMA-based synthesis of 75 peer-reviewed studies published between 2019 and April 2026, providing a unified taxonomy of Arabic fake news detection approaches and mapping the evolution of the field from classical machine learning to deep learning, transformer-based, hybrid, and emerging large language model approaches.
- A dataset-centered and methodological analysis that examines how dataset characteristics, annotation practices, dialectal coverage, class balance, and evaluation protocols affect the interpretation and comparability of reported model performance.
- A critical synthesis of the evidential strength and generalizability of existing research, highlighting the limitations of direct model comparisons caused by inconsistent datasets, evaluation settings, and limited cross-dataset, cross-domain, and cross-dialect validation.
- An evidence-based research agenda that integrates unresolved challenges in explainability, reproducibility, multimodal detection, and standardized evaluation, and identifies priorities for developing more reliable and generalizable Arabic fake news detection systems.

The remainder of this study is organized as follows: In Section II (Background and Definitions), the main definitions and terminology are reviewed. The main research questions and search strategy of this survey are described in Section III (Research Methodology). Section IV (Detection Techniques) provides a detailed taxonomy of the approaches for preprocessing, feature extraction, and classification. The main datasets used for Arabic fake news detection are studied, and the characteristics and limitations of the datasets are discussed in Section V (Datasets). Section VI (Evaluation Metrics) explains the most common evaluation metrics used in fake news detection and the importance of each metric. Section VII (Synthesis of Findings, Research Gaps, and Future Directions) identifies current research gaps and proposes future research directions. Finally, Section VIII presents the conclusions.

II. BACKGROUND AND DEFINITIONS

Fake news detection in Arabic has emerged as a critical research area within misinformation studies. This section provides an overview of fake news definitions, terminology, and its multifaceted impact on Arab societies across social, political, economic, and health domains.

A. Fake News Definition

Fake news lacks a universally accepted definition, and it is defined differently depending on the purpose of the research [1], [19]. For example, one study defines fake news as intentionally fabricated or misleading information that imitates legitimate news content in form and presentation but aims to deceive readers, manipulate opinions, or serve political and economic interests [20]. Fake news has also been characterized as news intended to deceive rather than inform, which leads to systematic differences in language and communicative purpose compared with authentic news. A previous study provided and defined a set of terms related to fake news, including hoaxes,

rumors, spam, misinformation, disinformation, clickbait, satire, and propaganda [21].

Fake news is associated with various related terms, including false or misleading information, misinformation, and fabricated information. It is typically spread for political or commercial gain through social media platforms and mobile applications [6], [19], [22], [23]. It consists of fabricated information deliberately created to appear authentic. Due to its potential for highly complex and large-scale dissemination, it often causes significant societal and financial harm. Rumors are unverified information initiated by potentially untrustworthy sources and circulated among users before their veracity can be established. Spam refers to irrelevant or unsolicited messages sent by individuals or groups on social media, such as advertisements, malicious links, and poor-quality content. Misinformation is false information spread unintentionally, for example, through honest mistakes or outdated knowledge, without intent to mislead. Disinformation is a type of false information spread among users with the main goal of misleading others for some purpose, such as deception [6].

Arabic fake news refers to information shared on social media platforms with the intent to deceive audiences, manipulate opinions, or achieve undisclosed objectives. It imitates credible news styles while concealing its deceptive purpose and often spreads rapidly across online communities. Detecting Arabic fake news is challenging due to the linguistic complexity and dialectal diversity of the Arabic language. The existence of multiple dialects, rich morphology, and variations in word forms complicates text representation and classification. Furthermore, limited annotated datasets and inconsistencies between Modern Standard Arabic (MSA) and dialectal usage reduce model performance. Ambiguities in word meanings and inconsistent spelling across regions also hinder the use of natural language processing tools [7], [24]. This necessitates the development of language-specific approaches to effectively identify and mitigate the spread of fake news in Arabic contexts.

B. Impact of Spreading Fake News

Arab countries in the Middle East have been significantly affected by the proliferation of fake news. A substantial number of people across the Arab world actively use social media platforms. Social media shapes daily life by influencing public opinion, identity, and cultural norms, thereby enabling political mobilization and expression [4]. Arabic fake news significantly influences public perceptions of economic, political, social, and religious issues in the Arab world. The rapid spread of false information on social media exacerbates societal divisions, increases social tension, and erodes trust in legitimate news sources.

Arabic fake news influences users by promoting biased or false beliefs, thereby altering their responses to authentic news stories. Several authors have demonstrated that such misinformation, spread for political or commercial gain, can significantly sway elections and public opinion. The spread of fake news extends beyond political and economic domains to other sectors. During the COVID-19 pandemic, substantial fake news and rumors proliferated within the health domain [2], [25]–[27]. Fake news contributes to unrest and social instabil-

ity by altering social perspectives, provoking public reactions, and has been linked to civil unrest in several Arab countries. Fake news undermines information integrity and public confidence, posing a significant challenge to trust in the Arab media landscape. Fake news causes harm at multiple levels, including individual, community, political, and financial dimensions.

III. RESEARCH METHODOLOGY

This study adopts a systematic literature review and taxonomy-based mapping approach to identify, screen, and synthesize research on Arabic fake news detection. The methodology was designed to increase transparency, reproducibility, and coverage of the field by explicitly reporting the information sources, the search strings, the publication period, the eligibility criteria, the screening procedure, the data extraction process, and the study-selection flow. The reporting structure follows the logic of PRISMA 2020 [28]. The synthesis aims to build a technical taxonomy of Arabic fake news detection approaches, including datasets, preprocessing techniques, feature representations, model families, evaluation metrics, and research gaps. All supplementary materials referred to in this review (Supplementary Tables S1–S5) are publicly available at <https://doi.org/10.5281/zenodo.23474152>.

A. Research Questions

The following research questions guided the review:

RQ1. What models have been employed in Arabic fake news detection, and how do they compare in terms of reported evaluation performance?

RQ2. How does Arabic fake news detection perform with transformer-based architectures compared to traditional machine learning and deep learning models?

RQ3. What are the common preprocessing, feature extraction, and representation methods for Arabic text, dialectal content, and social media data?

RQ4. How are Large Language Models (LLMs) being utilized in Arabic fake news detection, and what opportunities and limitations do they introduce compared with traditional encoder-based transformer models?

RQ5. How have explainable AI techniques, particularly SHAP, LIME, and attention visualization, been used to interpret Arabic fake news detection models, and what gaps remain in transparency, explanation reliability, and practical usability?

RQ6. To what extent are dataset characteristics, including size, source, annotation quality, class balance, dialectal coverage, and accessibility, associated with variations in reported model performance?

RQ7. What are the primary limitations of current Arabic fake news datasets, and what dataset-related deficiencies should future research rectify?

B. Search Strategy and Information Sources

A systematic literature search was conducted to find studies on Arabic fake news detection. Peer-reviewed studies published from January 2019 to April 2026 were searched. This period was chosen to cover both early studies on Arabic fake news detection and recent advances with deep learning, transformer-based models, and large language model approaches. The literature search was last conducted on 30 April 2026. During the revision, the search was updated with additional terms (rumor, misinformation, disinformation, fact-checking, and claim verification) within the same time window. This identified 17 further potentially eligible studies; the full texts of five could not be retrieved, and two did not meet the inclusion criteria, so ten studies were added [16], [29]–[37]. In parallel, all previously included studies were re-verified against the eligibility criteria, which led to the exclusion of ten records (two duplicates, six studies without Arabic data, one study on product reviews rather than news, and one non-peer-reviewed preprint). The review therefore includes 75 studies in total (65 from the original search and 10 from the updated search).

The main academic databases and digital libraries used were Scopus, Web of Science Core Collection, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Wiley Online Library, SAGE Journals, and MDPI. Google Scholar was used as a supplementary search engine to increase recall and identify potentially missed records. ResearchGate was used only as a supplementary source for locating full texts or author-uploaded versions; it was not treated as a primary bibliographic database unless the work could be verified as a published journal or conference paper.

Backward and forward snowballing were also performed by checking the reference lists of key studies and using citation tracking to identify further relevant papers.

C. Exact Search Strings

The search strings combined Arabic-language scope terms, fake-news-related terms, and detection/modeling terms. The generic search string was adapted to the syntax of each database, as shown in Table II.

D. Eligibility Criteria

The inclusion and exclusion criteria were defined before full-text screening to ensure consistency and reduce selection bias, as shown in Table III.

TABLE II. SEARCH STRATEGY AND DATABASE-SPECIFIC QUERY STRINGS USED FOR RETRIEVING ARABIC FAKE NEWS DETECTION STUDIES

Database / Source	Exact search string
Generic Boolean string	((Arabic OR Arab*) AND ("fake news" OR "false news" OR "true news" OR "real news") AND (detection* OR classification* OR identification OR "fake detection") AND ("machine learning" OR "deep learning" OR transformer* OR BERT OR AraBERT OR MARBERT OR CAMELBERT OR NLP OR "natural language processing"))
Scopus	TITLE-ABS-KEY ((Arabic OR Arab*) AND ("fake news" OR "false news" OR "true news" OR "real news") AND (detection* OR classification* OR identification OR "fake detection") AND ("machine learning" OR "deep learning" OR transformer* OR BERT OR AraBERT OR MARBERT OR CAMELBERT OR NLP))

Web of Science	TS=((Arabic OR Arab*) AND ("fake news" OR "false news" OR "true news" OR "real news") AND (detection* OR classification* OR identification OR "fake detection") AND ("machine learning" OR "deep learning" OR transformer* OR BERT OR AraBERT OR MARBERT OR CAMELBERT OR NLP))
IEEE Xplore	("All Metadata": "Arabic" OR "All Metadata": Arab) AND ("All Metadata": "fake news" OR "All Metadata": "false news" OR "All Metadata": "true news" OR "All Metadata": "real news" OR "All Metadata": "fake" OR "All Metadata": "false" OR "All Metadata": "true" OR "All Metadata": "real") AND ("All Metadata": "detection" OR "All Metadata": "classification" OR "All Metadata": "verification" OR "All Metadata": "machine learning" OR "All Metadata": "deep learning" OR "All Metadata": "transformer" OR "All Metadata": BERT)
ACM Digital Library	[[All: Arabic] OR [All: Arab]] AND [[All: "fake news"] OR [All: "false news"] OR [All: "true news"] OR [All: "real news"]] AND [[All: detection] OR [All: classification] OR [All: verification] OR [All: "machine learning"] OR [All: "deep learning"]] OR [All: transformer] OR [All: BERT]
ScienceDirect / SpringerLink / Wiley / Taylor & Francis / SAGE / MDPI	(Arabic OR Arab*) AND ("fake news" OR "false news" OR "true news" OR "real news") AND (detection* OR classification* OR identification OR "fake detection") AND ("machine learning" OR "deep learning" OR transformer* OR BERT OR AraBERT OR MARBERT OR CAMELBERT OR NLP)
Google Scholar	"Arabic fake news detection" OR "Arabic false news detection" OR "Arabic true news classification" OR "Arabic real news" AraBERT OR "classification" OR "Arabic rumor detection" OR "Arabic fake news" AraBERT OR "Arabic false news" "machine learning"

TABLE III. ELIGIBILITY CRITERIA USED FOR SELECTING STUDIES

Inclusion criteria	Exclusion criteria
<ol style="list-style-type: none">1. Published between January 2019 and April 2026.2. Peer-reviewed journal articles and conference papers with sufficient methodological detail.3. Studies focused on Arabic fake news, misinformation, disinformation, real news, fact-checking, claim verification, or news detection.4. Studies using Arabic text, Arabic social-media content, Arabic news data, Arab-world fact-checking data, or multilingual datasets with a clearly separable Arabic component.5. Studies proposing or evaluating datasets, preprocessing methods, feature representations, machine learning models, deep learning models, transformers, and hybrid models.6. Studies reporting enough information to extract at least the dataset/source, method/model, evaluation metric, and main finding.	<ol style="list-style-type: none">1. Studies unrelated to Arabic content or without a separable Arabic component.2. General studies about fake news not related to detection, classification, verification, or computational analysis.3. Editorials, opinions, posters, abstracts, theses, lecture slides, non-scholarly blog posts, or papers without full text.4. Duplicate records; when multiple versions existed, the most complete peer-reviewed version was retained.5. Studies with insufficient methodological details, missing evaluation results, or unclear dataset descriptions.6. Studies focusing only on non-textual misinformation unless Arabic textual/multimodal detection was part of the method.

E. Screening and Selection Process

All retrieved records were exported to a reference manager or spreadsheet. Duplicate records were removed using DOI, title, author names, publication year, and venue information. The screening process was conducted in three stages.

First, the titles and abstracts of the retrieved records were screened by the first author, and the two co-authors reviewed the study selection decisions to ensure consistency with the predefined inclusion and exclusion criteria. Second, the full texts of potentially relevant papers were assessed against the inclusion and exclusion criteria. Third, the final included studies were checked again during data extraction to ensure that each paper provided sufficient information for synthesis. Reasons for full-text exclusion were recorded, including absence of Arabic data, product reviews rather than news, non-peer-reviewed or unverifiable venues, scope outside fake news detection, and duplicate publication versions. Full-text exclusions with reasons are listed in Supplementary Table S4.

F. Data Extraction

A structured extraction sheet was created to ensure that the same information was collected from each included study. The extracted fields included bibliographic information, research objective, fake-news definition or task formulation, data source, dataset name, dataset size, language variety or dialect, labeling strategy, preprocessing steps, feature extraction or representation method, model family, specific algorithms or architectures, train-test split or validation strategy, evaluation metrics, reported results, limitations, and future work. The ex-

tracted characteristics and reported results of all 75 included studies, including studies that are not discussed individually in the text [38]–[44], are provided in Supplementary Table S1.

Additional fields were extracted to support the survey's taxonomy, including whether the study used classical machine learning, deep learning, transformer-based models, or hybrid models. This extraction process enabled both descriptive synthesis and structured comparison across studies.

G. Quality Assessment

The quality checklist examined the following key aspects of each study: 1) the fake news detection task was clearly defined, 2) the dataset source, size, and labelling procedure were adequately described, 3) the preprocessing and feature representation methods were sufficiently explained, 4) the classification model or architecture was clearly presented, 5) appropriate evaluation metrics were reported, 6) comparisons with baseline approaches were provided, 7) limitations and threats to validity were discussed, and 8) sufficient experimental details were reported to support reproducibility.

The quality assessment criteria were defined to examine the methodological rigor and reporting quality of the included studies. These criteria address dataset transparency, evaluation practices, reproducibility, and experimental reporting. Each included study was scored on the eight criteria (Yes = 1, Partly = 0.5, No = 0; maximum score = 8) by the first author, and the scores were verified by the co-authors. Studies scoring 6 or more were classified as high quality, 4–5.5 as medium, and below 4 as low quality. Of the 75 included studies, 61 (81%) were rated high, 11 (15%) medium, and 3 (4%) low quality

(mean score = 6.38 out of 8). The least satisfied criteria were the reporting of preprocessing details (Q3, fully met by 21% of the studies) and reproducibility (Q8, 24%), followed by the use of evaluation metrics beyond accuracy (Q5, 49%). These reporting gaps limit the comparability of results across studies and are taken into account in the synthesis (Section IV). Per-study scores are provided in Supplementary Tables S2 and S3.

H. Data Synthesis and Taxonomy Development

The included studies were synthesized using a narrative and taxonomy-based approach. First, the studies were categorized by their main contribution, such as feature engineering, model development, transformer-based detection, multimodal detection, explainability, or analysis of fake news spread. Second, detection methods were categorized into classical machine learning, deep learning, transformer-based models, hybrid approaches, and large language model approaches, where applicable. Third, datasets and data sources were mapped according

to source type, including social media platforms, online news websites, and manually constructed corpora.

Because reported performance values often reflect different datasets, label definitions, sample sizes, and evaluation protocols, the results were compared cautiously. Accuracy, precision, recall, F1-score, and AUC were extracted when available, but direct model ranking was avoided unless studies used the same dataset and comparable experimental settings.

I. PRISMA Flow Diagram

The study-selection process is summarized using a PRISMA 2020 flow diagram. The diagram reports the number of records identified from databases and other sources, records removed before screening, records screened, records excluded, full-text reports assessed for eligibility, full-text reports excluded with reasons, and studies finally included in the review, as shown in Fig. 1.

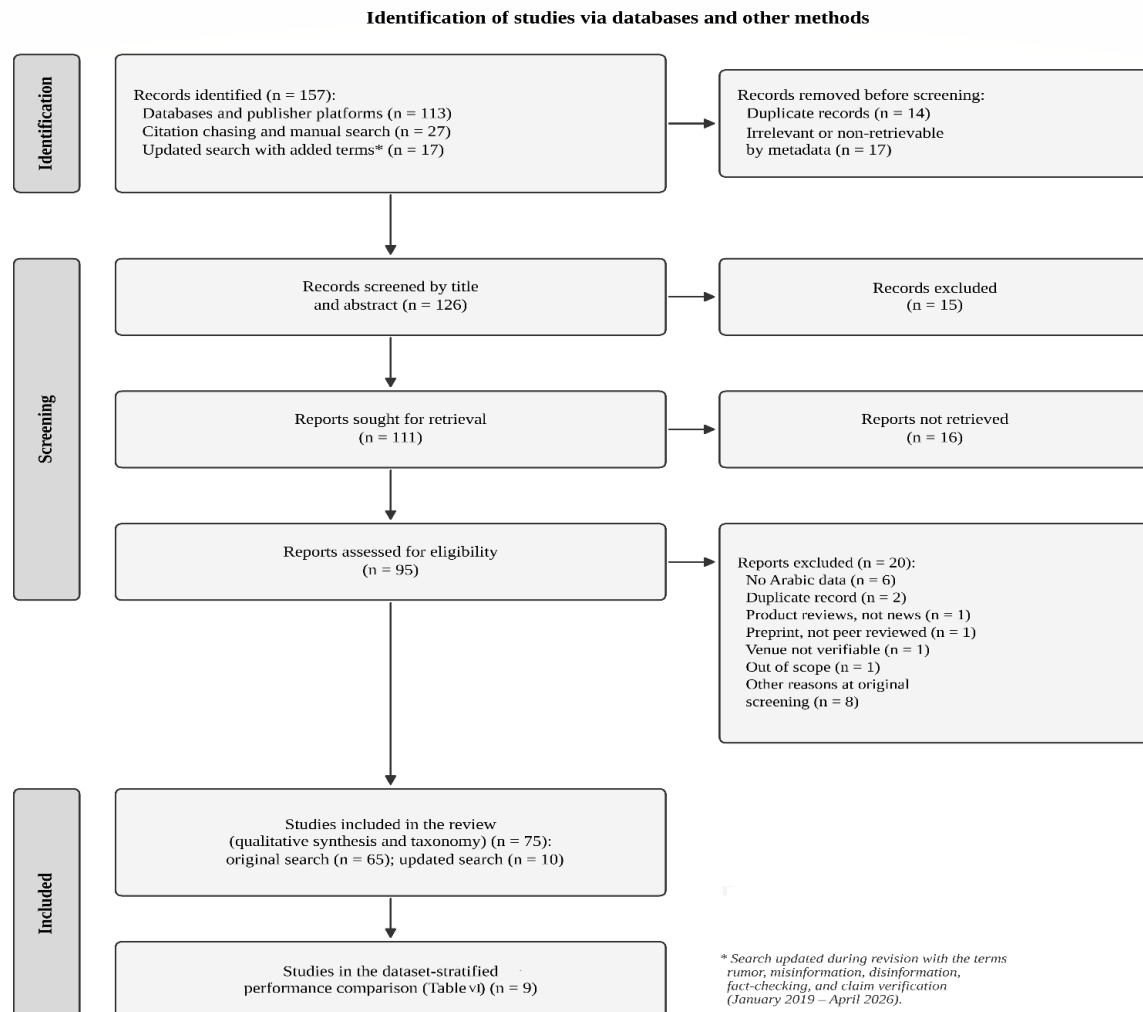


Fig. 1. PRISMA-based study selection flow for Arabic fake news detection studies included in the qualitative synthesis and taxonomy development; the dataset-stratified performance comparison (Table VI) covers a subset of these studies.

IV. DETECTION TECHNIQUES: A COMPREHENSIVE TAXONOMY

The classification process is one of the most important stages in detecting fake news in Arabic, as it includes several steps, such as preprocessing, feature extraction, and classification models, which are detailed in this section.

A. Preprocessing

Preprocessing is a key step in the Arabic fake news detection process because Arabic textual data is prone to noise, spelling variation, informal writing, duplicated content, URLs, hashtags, emojis, and inconsistent orthographic forms, especially when collected from online news platforms and social media [5], [24], [45]–[48]. The preprocessing techniques of the reviewed studies are classified into four main categories: general text cleaning, Arabic-specific normalization, linguistic preprocessing, and dataset-level preparation. General cleaning includes the removal and/or replacement of punctuation marks, special characters, URLs, numerals, emojis, hashtags, duplicated postings, and repeated characters. The goal of Arabic-specific normalization is to reduce orthographic diversity by unifying distinct forms of Alef (ا → آ, إ, إ), normalizing Alef Maqsurah and Ta Marbuta (ة → ة, ي → ى), deleting diacritics, and eliminating tatweel [7], [47], [49]. The language preprocessing includes stemming, tokenization, lemmatization, and rooting. At the dataset level, preprocessing includes approaches such as removing duplicates and balancing classes (e.g., by oversampling or undersampling).

Fig. 2 shows the basic preprocessing steps performed on the raw Arabic text data before feature extraction. These procedures involve text cleaning, duplication removal, data balancing, normalization, word segmentation, and root identification to get clean texts for fake news detection models.

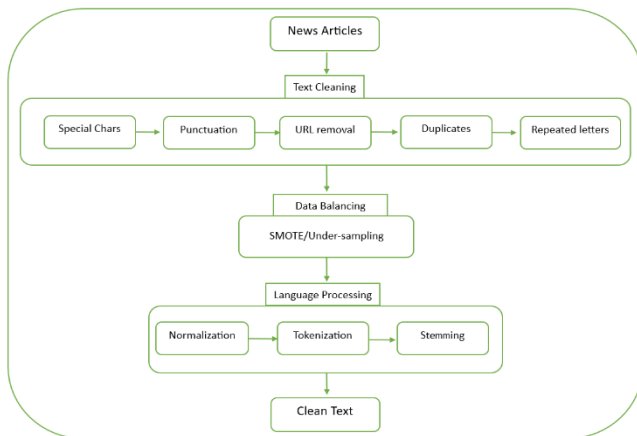


Fig. 2. Main preprocessing steps applied to Arabic news text before feature extraction and classification.

A key observation from the reviewed literature is that preprocessing is often reported as a fixed preliminary step rather than being systematically evaluated. Many studies describe cleaning and normalization procedures, but fewer studies investigate how individual preprocessing decisions affect classification performance [47], [50]. This limitation is particularly important because preprocessing requirements differ across

model families. Traditional machine learning models generally depend more heavily on explicit cleaning, normalization, and feature engineering, whereas transformer-based models may benefit from lighter preprocessing that preserves contextual and stylistic information [49]. Therefore, this survey analyzes preprocessing not only as a technical preparation step, but also as a methodological factor that affects feature representation, model performance, and the comparability of reported results across Arabic fake news detection studies.

Overall, the reviewed studies show that preprocessing is essential for reducing textual noise and orthographic variation in Arabic fake news datasets. However, the lack of standardized preprocessing protocols and the limited reporting of the impact of preprocessing make it difficult to fairly compare results across studies. In future work, it is recommended to report the preprocessing configurations more explicitly, evaluate their effect using ablation experiments, and distinguish between the preprocessing strategies that are suitable for classical machine learning models and those that are suitable for transformer-based Arabic language models.

B. Feature Extraction

Feature extraction is the process of transforming cleaned Arabic text into meaningful numerical representations that capture linguistic and semantic information. It involves using classical techniques such as TF-IDF, Bag-of-Words, and n-grams, static embeddings such as Word2Vec, FastText, and GloVe-Arabic, and transformer-based embeddings such as AraBERT to encode textual features, as shown in Fig. 3. These extracted features enable machine learning, deep learning, and hybrid models to accurately detect and classify fake Arabic news. Feature extraction in Arabic is particularly challenging due to the morphological richness of the language, derivational complexity, dialectal variation, and context-dependent meanings [51], [52]. Therefore, an appropriate feature representation method is needed to improve the ability of detection models to distinguish between fake and real Arabic news. As shown in Fig. 4, the reviewed studies mainly employ traditional sparse features, static word embeddings, and contextual transformer-based embeddings.

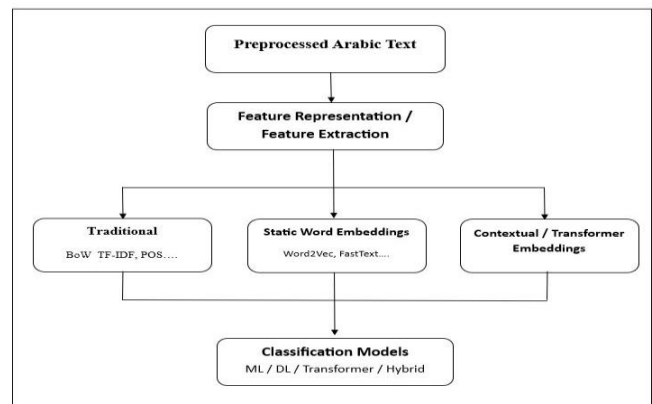


Fig. 3. Taxonomy of feature representation methods for preprocessed Arabic text in Arabic fake news detection.

1) *Traditional feature extraction*: Traditional feature extraction refers to classical methods of converting raw text

data into numerical representations using statistical, frequency-based, and linguistic approaches that do not capture the semantic relationships between words. These representations are frequency-based, sparse, context-independent, interpretable, and vocabulary-dependent. The main traditional techniques are:

- 1) Bag of Words (BoW): Represents text as a vector of word frequencies or binary presence. It loses order and context, but is simple and effective for small datasets [49].
- 2) Term frequency-inverse document frequency (TF-IDF): This technique assigns less weight to common words and more weight to rare but important words in the text. When the vocabulary is controlled, it works well for Arabic news or sentiment classification; its basic formulation is given in Eq. (1) [53]. It is often used with machine learning models [9]–[12], [54].

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

where, $\text{TF}(t, d)$ is the frequency of term t in document d , N is the total number of documents in the corpus, and $\text{DF}(t)$ is the number of documents that contain term t . The inverse document frequency (IDF) component assigns lower weights to frequently occurring terms and higher weights to rarer terms. This equation represents the basic TF-IDF formulation; smoothing and vector normalization may vary depending on the implementation.

2) *Static word embeddings*: Static word embeddings are dense representations of words in a continuous semantic space. Unlike traditional sparse features, these representations can encode semantic relations between words. However, each word is assigned a fixed vector independent of its context. AraVec is a pre-trained Arabic word embedding model developed for Arabic text, including Modern Standard Arabic and dialectal Arabic. It has been used in Arabic NLP tasks to represent semantic relationships between words [55]–[58]. GloVe-Arabic is based on the Global Vectors for Word Representation approach. It generates word vectors by using word co-occurrence statistics from large Arabic corpora, allowing models to capture global semantic patterns in text [54], [59]–[61]. Other static embedding techniques have been applied to Arabic fake news detection, such as Word2Vec and FastText. FastText can be very useful for Arabic as it takes into account subword information that can help in representing morphologically rich words [54], [59]–[64].

3) *Contextual and transformer-based embeddings*: Contextual embeddings generate different representations for the same word depending on its surrounding context. This makes them more suitable for Arabic text, where the meaning of a word can vary based on the sentence structure, morphology, and dialect. Recently, transformer-based models such as AraBERT, MARBERT, CAMELBERT, and ArabicBERT have been increasingly used in Arabic fake news detection, since they can capture deep contextual and semantic information [25], [65], [69]. These models are better suited

than traditional and static embedding approaches to address ambiguity, contextual meaning, and complex linguistic patterns in Arabic text.

4) *Summary of feature extraction usage*: Fig. 4 shows the distribution of feature extraction techniques used in the reviewed studies for Arabic fake news detection. Traditional feature extraction methods (TF-IDF, BoW, POS, n-grams) were used in 25 studies, while 18 studies employed static word embeddings (Word2Vec, FastText, AraVec, GloVe-Arabic). The most dominant category was transformer-based embeddings, used in 50 studies. This distribution shows that feature extraction for Arabic fake news detection has moved from simple, sparse representations to semantic and contextual representations.

It is important to note that these frequencies represent the number of included studies using each category of feature representation. Since a single study may employ multiple feature extraction methods, the categories are not mutually exclusive.

This trend reflects a clear methodological shift in the field. Classical methods remain useful because they are easy to interpret and compatible with classical machine learning algorithms, but they do not consider specific issues of the Arabic language such as rich morphology, ambiguity of words, dialectal variation, and contextual meaning. Static embeddings address some of these limitations by embedding words in dense semantic spaces, but still have limited contextual awareness. The widespread use of transformer-based embeddings suggests that recent studies tend to utilize Arabic-aware contextual models such as AraBERT, MARBERT, CAMELBERT, and other BERT-based models to capture deeper linguistic and semantic cues. Therefore, this analysis helps not only to show the frequency of feature extraction techniques, but also to show the evolution of Arabic fake news detection research towards more context-aware and semantically rich representations.

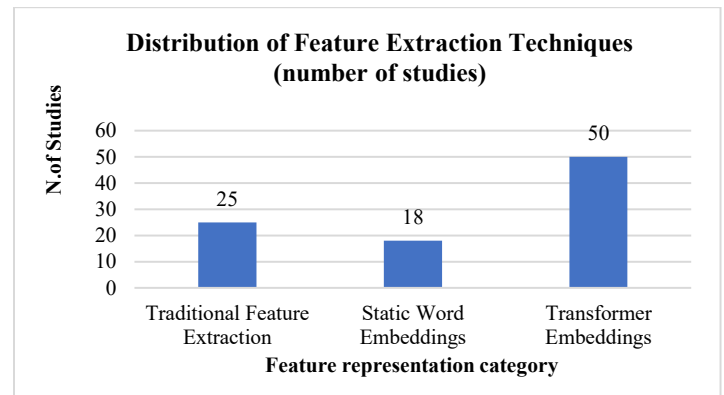


Fig. 4. Distribution of feature extraction and representation techniques reported in Arabic fake news detection studies.

C. Classification Approaches

Classification approaches constitute the central stage in Arabic fake news detection, where extracted textual representations are mapped into final prediction labels. As shown in Fig. 5, the general detection pipeline takes Arabic textual inputs (news articles, tweets, posts, and comments) and transforms them using feature extraction techniques such as TF-IDF, Bag-

of-Words, word embeddings, and contextual representations. The representations are then fed into various classification model families, including classical machine learning, deep learning, transformer-based models, and hybrid-based approaches to classify whether the content is fake or real. This section, therefore, reviews and analyzes the main classification approaches adopted in the studies of Arabic fake news detection, focusing on the models used, their role in the methodology, the trends in reported performance, and their implications in the face of the linguistic and contextual complexity of Arabic misinformation.

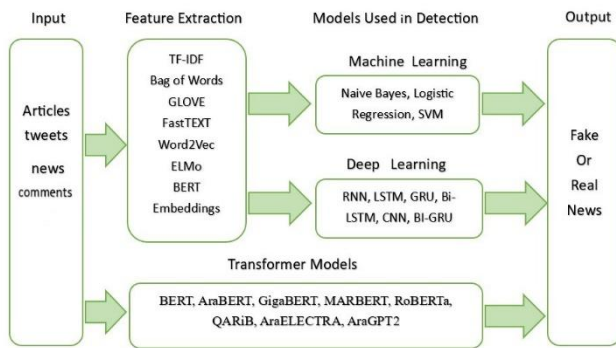


Fig. 5. General pipeline of Arabic fake news detection from textual input to classification output.

1) *Classical machine learning models*: Classical machine learning models are one of the basic approaches to fake news detection in Arabic, and they are still important baselines for evaluating more advanced architectures. These models usually depend upon explicit feature representations such as TF-IDF, Bag-of-Words, n-grams, part-of-speech features, and other lexical or linguistic features to represent Arabic text as numerical vectors for classification. Classical classifiers, unlike deep learning and transformer-based models, do not learn contextual representations directly from the raw text. Hence, their effectiveness is highly dependent on the quality of preprocessing, normalization, tokenization, and feature engineering. This dependency is of particular importance in Arabic due to the language's rich morphology, orthographic variation, dialectal diversity, and context-sensitive meanings.

In the reviewed studies, several classical machine learning algorithms have been applied for the classification of fake vs. real Arabic news, such as Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbors (KNN), XGBoost, and Decision Tree (DT). As shown in Fig. 6, RF and SVM were the most used classical classifiers, followed by NB and LR, and KNN, XGBoost, and DT were less frequently used [10]–[12], [54], [70], [71]. The continued use of these models shows that classical machine learning is still relevant for the detection of Arabic fake news, especially when datasets are relatively small, computational resources are limited, or interpretable baseline models are needed. Moreover, these models serve as a good baseline to assess whether the performance gains by deep

learning or transformer models are due to better representation learning or dataset-specific features.

Despite their practical advantages, classical machine learning models present several limitations in Arabic fake news detection. Their reliance on sparse and handcrafted representations limits their ability to capture semantic relations, long-range dependencies, pragmatic cues, and dialect-specific meanings. Furthermore, the performance of these models is often sensitive to the choice of pre-processing steps such as stemming, stop-word removal, diacritic removal, and normalization, which are not consistently reported across studies. Another major limitation is that many studies evaluate classical models on a single dataset, which makes it difficult to evaluate their generalizability across domains, dialects, platforms, and annotation schemes [9]. Therefore, future research should continue to use classical machine learning models as transparent and reproducible baselines, but with fair comparisons with deep learning and transformer-based models under the same settings of preprocessing, feature configurations, datasets, and evaluation protocols.

The distribution of classical machine learning models further exposes the methodological preferences in Arabic fake news detection research. The common use of SVM and RF suggests that researchers tend to favor models that are robust with sparse text representations like TF-IDF and Bag-of-Words, and can perform reasonably well on relatively small or medium-sized datasets. LR and NB also continue to be popular because of computational efficiency, ease of implementation, and appropriateness for creating transparent baseline results, as shown in Fig. 6. The lower frequency of XGBoost, KNN, and DT may be a reflection of their narrower use in Arabic fake news studies, a sensitivity to feature dimensionality, or a trend in recent years to skip traditional classifiers and jump directly to deep learning and transformer-based architectures. However, use frequency does not necessarily indicate superior performance but represents methodological trends, dataset limitations, and experimental preferences found in the reviewed literature.

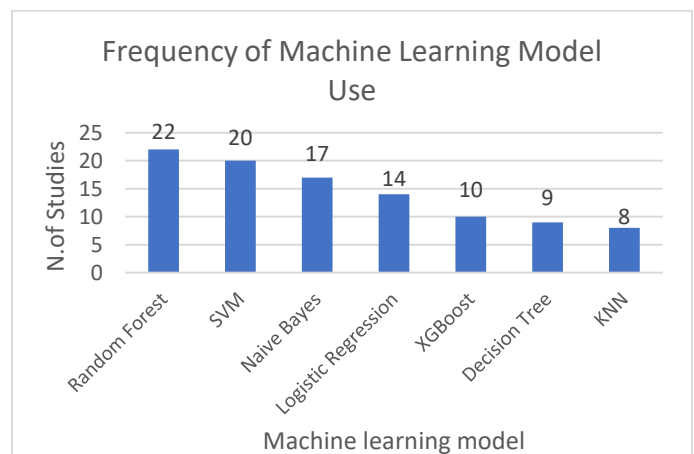


Fig. 6. Distribution of commonly used machine learning models in Arabic fake news detection studies.

2) *Deep learning models*: Deep learning models represent a major methodological advancement in Arabic fake news detection because they reduce the dependence on manually engineered features and enable models to learn more complex textual representations from data. Deep learning models can capture semantic, syntactic, and sequential patterns in Arabic text through distributed representations and neural architectures, as opposed to classical machine learning approaches that generally rely on sparse representations such as TF-IDF and Bag-of-Words. This ability is important for Arabic fake news detection since the language has rich morphology, flexible word order, dialectal variation, and context-dependent meanings. Therefore, deep learning methods are often used to model more profound linguistic structures and identify nuanced deceptive clues that are hard to capture well by traditional feature-based methods.

Various deep learning architectures have been employed for the classification of Arabic fake news in the reviewed studies, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), Gated Recurrent Units (GRUs), Bidirectional LSTM (BiLSTM), and Bidirectional GRU (BiGRU) [57], [61], [72]–[74]. As shown in Fig. 7, CNN, BiLSTM, and LSTM were the most frequently used deep learning models, followed by RNN, BiGRU, and GRU. This distribution indicates a tendency for researchers to prefer architectures that can capture either local textual patterns or sequential dependencies. CNN-based models can be used to detect local n-gram-like patterns and salient textual features, while LSTM and GRU models are good at capturing temporal and sequential dependencies in text. BiLSTM and BiGRU are bidirectional models that are designed to better understand the context by processing the input sequence in both forward and backwards directions, thus providing a more accurate representation of the meaning of the sentence in Arabic, where the context can be based on the surrounding words.

However, deep learning models also have several limitations for Arabic fake news detection despite their advantages. First, they usually need larger and more representative labelled datasets than classical machine learning models, which remains a challenge given the scarcity and imbalance of Arabic fake news datasets. Second, deep learning models may be sensitive to dataset size, domain variation, hyperparameter settings, and embedding quality, making the reported performance hard to compare across studies [59], [60]. Third, although these models learn richer representations than classical classifiers, they may still struggle with dialectal variation, implicit meanings, sarcasm, and cross-domain generalization when trained on limited or domain-specific datasets. In addition, many studies do not provide sufficient ablation experiments to isolate the contribution of model architecture, word embeddings, preprocessing, or dataset characteristics. Therefore, future work should investigate deep learning models using standardized benchmarks, cross-dataset validation, well-defined experimental protocols, and comparative studies with traditional machine learning and transformer-based methods.

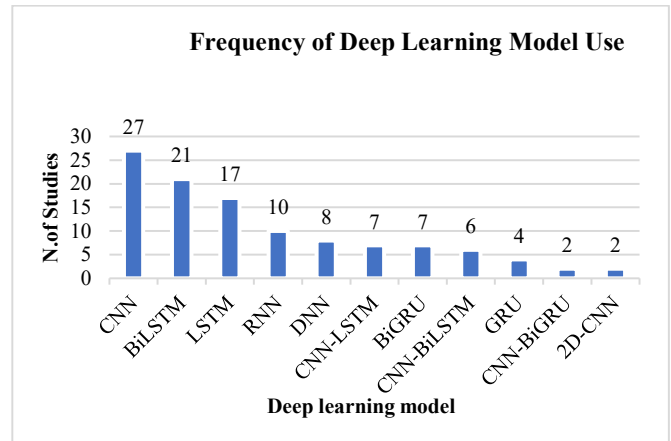


Fig. 7. Distribution of commonly used deep learning models in Arabic fake news detection studies.

The high frequency of CNNs indicates the importance of detecting salient phrase-level and n-gram-like features, while the high use of LSTM and BiLSTM indicates the need to model word order, contextual dependencies, and long-range linguistic relations in Arabic text. The relatively low usage of DNN, GRU, RNN, BiGRU, and 2D-CNN may reflect dataset-driven experimental choices, the weaker ability of simple recurrent models to capture long-range dependencies, or the increasing trend of recent studies to move towards hybrid and transformer-based models. Less frequent are hybrid deep learning models (e.g., CNN-LSTM, CNN-BiLSTM, and CNN-BiGRU), but their presence indicates a developing trend that aims to combine local feature extraction and sequential context modelling. However, the frequency of use of the models does not necessarily imply better performance, but rather reflects the methodological choices, dataset limitations, and experimental procedures followed in the literature reviewed.

3) *Transformer models*: Transformer-based models have become a major direction in Arabic fake news detection due to their ability to generate deep contextual representations of text. Unlike recurrent deep learning models that process text sequentially, and conventional machine learning approaches that use handcrafted features, transformers use self-attention mechanisms to capture semantic relationships between words across the whole input. This ability is important for detecting fake news in Arabic, because deceptive content often depends on context and relies on implicit meanings, emotional language, and subtle linguistic cues. Thus, transformer-based models provide a better basis for representing the complexity of Arabic text and for better discriminating fake news from real news.

The importance of transformer-based models is even more evident in Arabic due to the linguistic complexity of the language. Arabic is morphologically rich, syntactically flexible, orthographically varied, dialectally diverse, and semantically ambiguous. Pre-trained Arabic and multilingual transformer models such as BERT, AraBERT, MARBERT, ARBERT, CAMeLBERT, AraELECTRA, QARiB-BERT, and GigaBERT have been used to tackle these issues by learning deep contextual representations from large-scale corpora and fine-

tuning them for fake news classification [65], [66], [68], [69], [75]. This transfer learning helps in detecting Arabic fake news because the labelled datasets of Arabic are usually small, imbalanced, or domain-specific. Transformer models leverage pre-trained knowledge of language to better understand sentence context, semantic connections, dialectal expressions, and word interactions in Arabic news and social media texts, as shown in Fig. 8.

The reviewed studies show that transformer-based models have benefited Arabic fake news detection, since they enable classifiers to shift from surface lexical patterns to deeper semantic and contextual analysis. These models can capture deceptive cues not only at the lexical level but also within broader sentence and discourse contexts. In this sense, they are particularly useful for identifying fake news that relies on linguistic manipulation, deceptive framing, or subtle contextual distortion. Moreover, the transformer-based models give a strong basis for advanced directions such as hybrid architectures, contextual feature extraction, cross-domain detection, multilingual transfer learning, and explainable fake news detection [68], [69], [76]–[79]. However, their effectiveness still depends on dataset quality, domain coverage, dialectal representation, and evaluation design. Therefore, future research should evaluate transformer models using standardized Arabic benchmarks, cross-dataset validation, dialect-aware testing, and explainability methods to ensure that reported improvements reflect genuine generalization rather than dataset-specific performance.

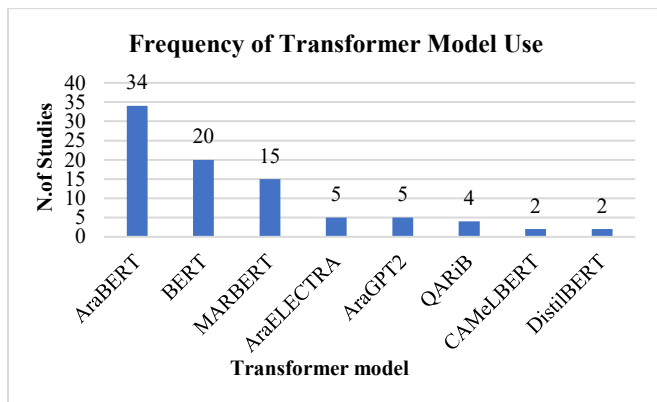


Fig. 8. Frequency distribution of transformer-based models reported in the reviewed Arabic fake news detection studies.

Overall, transformer-based models represent a key turning point in Arabic fake news detection by shifting the field from feature-engineering-based and sequential modeling approaches toward contextual representation learning. Their main contribution lies not only in improving classification performance, but also in enabling more Arabic-aware and context-sensitive detection of deceptive content, especially when misinformation depends on subtle framing, emotional language, or implicit meanings. However, their effectiveness remains constrained by several limitations, including high computational cost, limited interpretability, sensitivity to fine-tuning settings, and insufficient evidence of generalization across datasets, domains, dialects, and time periods. Future research should prioritize fair comparative assessment, standardized Arabic benchmarks, cross-dataset and cross-dialect validation, explainable trans-

former-based models, and the integration of textual representations with metadata, temporal signals, and multimodal features.

4) *Large language models for Arabic fake news detection:* Large Language Models (LLMs) represent an emerging direction in Arabic fake news detection. Unlike encoder-based transformer models such as AraBERT, MARBERT, CAMELBERT, and QARIB-BERT, which are usually fine-tuned for a specific classification task, LLMs are large-scale generative models that can perform classification, reasoning, instruction following, and explanation generation through natural-language prompts. This makes them particularly relevant to Arabic fake news detection, where the classification decision may depend on context, dialectal variation, implicit meaning, cultural references, and the ability to produce a human-understandable justification. Therefore, LLMs can be viewed as an extension of the transformer-based direction, moving from fixed supervised classification toward instruction-based, generative, and evidence-aware detection.

In Arabic fake news detection, LLMs can be used through different strategies. First, they can be applied in zero-shot settings, where the model is asked to classify an Arabic claim or news item as fake or real without task-specific training [33], [37]. Second, they can be used in few-shot prompting, where a small number of labeled examples are included in the prompt to guide the model's decision. Third, LLMs can support explanation generation by producing a natural-language rationale for why a piece of news is likely to be fake or real [80]. Fourth, they can be used for data augmentation, where synthetic Arabic fake or real news examples are generated to reduce data scarcity [3], [81]. Finally, LLMs may be combined with retrieval-augmented generation, where external evidence from trusted sources is retrieved and then used to support the final classification decision. These uses are especially important for Arabic because annotated fake news datasets remain limited, imbalanced, and often restricted to specific domains or dialects.

Among LLM-related approaches, GPT-based models are the most directly relevant to the Arabic fake news detection literature reviewed in this survey. GPT models are generative transformer models that learn to predict the next token in a sequence and can generate fluent text from short prompts. In the Arabic misinformation context, GPT-based models have been used both as a source of automatically generated misleading text and as a component in detection or data augmentation pipelines [69], [81]. For example, Arabic GPT-2 has been used to generate artificial Arabic tweets, and a transformer-based detector was then trained to distinguish human-written Arabic text from GPT-2-generated Arabic text [69]. This direction is important because it shows that fake news detection is no longer limited to identifying manually written misinformation, but must also address machine-generated Arabic misinformation. Other Arabic studies have also used AraGPT2 or LLM-based augmentation to enrich Arabic fake news datasets and improve classification robustness [3], [46], [81]. These studies suggest that GPT-style models can contribute to Arabic fake news detection by generating additional training data, simulating machine-generated misinformation, and supporting prompt-based classification and explanation.

However, the current Arabic fake news detection literature is still limited in its use of other large language model families such as LLaMA [82], Falcon [83], and Jais [84]. Such models are important as open or Arabic-oriented alternatives to closed GPT-based systems. LLaMA-style models are useful for future research because they can be adapted locally and fine-tuned using parameter-efficient techniques. Falcon models are also promising because they are open large language models developed in the Arab region and may support future Arabic misinformation detection experiments. Jais is particularly relevant because it is an Arabic-centric bilingual LLM designed to support Arabic and English, including instruction-tuned variants that can better follow Arabic prompts. Nevertheless, these models have not yet been widely evaluated in Arabic fake news detection compared with AraBERT, MARBERT, and other encoder-based transformer models. Therefore, future studies should investigate LLaMA, Falcon, Jais, and Arabic instruction-tuned LLMs for Arabic fake news detection using standardized datasets, zero-shot and few-shot prompting, fine-tuning, retrieval-based verification, and cross-dialect evaluation.

Among the 75 included studies, 10 (13%) used large language models or generative language models. Most of them used GPT-2/AraGPT2 models or zero-shot transfer for text generation, data augmentation, detection, or as baselines [3], [46], [60], [67], [69], [72], [81], [85], and only two directly evaluated a recent instruction-following LLM (GPT-4) [33], [37]. In the ArAIEval 2023 disinformation task, GPT-4 prompting reached a micro-F1 of 79.01% on the development set, compared with 76.83% for an ensemble of Arabic BERT models [33], whereas the factual accuracy of GPT-4 on fact-checked Arabic health claims expressed in four dialects was only 54.0–55.6% [37]. This limited and mixed evidence indicates that LLM-based Arabic fake news detection remains at an early stage.

Despite their potential, LLMs introduce several limitations. The first limitation is hallucination, where the model may generate unsupported claims, fabricated evidence, or confident but incorrect explanations. This is highly problematic in fake news detection because a fluent explanation may appear trustworthy even when it is not factually grounded. The second limitation is explainability. Although LLMs can generate natural-language rationales, these rationales are not necessarily faithful to the model's internal decision process. Therefore, generated explanations should not be treated as complete evidence of model transparency. The third limitation is computational cost, since fine-tuning or deploying large models requires considerable memory, processing power, and inference time. Additional challenges include prompt sensitivity, inconsistency across runs, limited dialectal coverage, and the lack of large, balanced, and publicly available Arabic fake news benchmarks. Overall, LLMs are promising tools for Arabic fake news detection, but they should be used with caution and supported by evidence grounding, human validation, transparent evaluation protocols, and explainability methods.

5) *Hybrid-based models*: In this survey, hybrid-based models are defined as cross-paradigm approaches that combine components from different model families or

representation strategies under one unified Arabic fake news detection framework. Hybrid models in this survey refer to those models that combine transformer-based representations with machine learning classifiers, transformer models with deep learning architectures, or deep learning feature extractors with classical machine learning models. These models are distinct from standalone deep learning architectures such as CNN, LSTM, BiLSTM, GRU, or RNN discussed under deep learning models. Examples include AraBERT or MARBERT embeddings with SVM, RF, LR, or XGBoost classifiers; BERT-based representations with CNN, LSTM, or BiLSTM layers; and neural feature extraction with traditional classifiers [13], [15], [58], [64], [81], [85], [86]. This classification provides a more explicit taxonomy by separating single-family models from cross-family hybrid approaches, and points to how hybrid models are striving to merge the semantic power of transformers, the representation-learning capacity of deep learning, and the classification robustness of classical machine learning.

Hybrid-based models are one of the most important directions in Arabic fake news detection, where 24 papers reported the application of hybrid architectures or cross-paradigm combinations, according to the reviewed studies. In many of these studies, the hybrid models achieved the best results reported in their respective experimental settings, suggesting that the complementary modelling strategies can improve the classification effectiveness [13], [58], [64], [76], [86]. From a comparative perspective, they aim to compensate for the limitations of standalone approaches. Classical machine learning models are efficient and interpretable but rely heavily on handcrafted features; deep learning models capture richer local and sequential patterns but may struggle with limited data and generalization; transformer-based models offer strong contextual representations but are computationally expensive and less interpretable. Hybrid models attempt to leverage these strengths using lexical clues, contextual embeddings, sequential dependencies, dialectal expressions, and classifier-level decision boundaries. This is especially useful in detecting fake news in the Arabic language, where the deceptive content may incorporate morphological complexity, implicit meanings, emotional framing, writing style, and interpretation based on context.

However, the strong reported performance of hybrid models should be interpreted cautiously because the reviewed studies differ in datasets, preprocessing strategies, feature representations, evaluation metrics, and validation protocols. In addition, hybrid models have several limitations, such as high computational cost, increased architectural complexity, reduced interpretability, and overfitting risk when used with a small or imbalanced dataset of Arabic fake news. Furthermore, many studies do not perform sufficient ablation analysis to ascertain if the performance enhancement is attributable to the transformer representation, the deep learning component, the classical classifier, or the dataset-specific characteristics. Future studies should evaluate hybrid models on standard Arabic benchmarks, cross-dataset and cross-dialect validation, component-level ablation studies, explainability methods, and fair comparisons with standalone machine learning, deep learning, and transformer-based models.

6) *Model usage frequency analysis*: To complement the taxonomy of classification approaches discussed in the previous subsections, this section presents a frequency-based analysis of the models used across the reviewed Arabic fake news detection studies. The models were grouped into four main families: machine learning, deep learning, transformer-

based, and hybrid-based approaches. Table IV reports the number of studies in which each model was explicitly used. Since several studies employed more than one model, the reported values represent the number of studies using each model family, and the families are not mutually exclusive.

TABLE IV. FREQUENCY DISTRIBUTION OF MODEL FAMILIES AND SPECIFIC MODELS REPORTED IN ARABIC FAKE NEWS DETECTION STUDIES

Type	Model	Other models	No. of studies
Machine learning	Random Forest (RF), SVM, Logistic Regression (LR), Naïve Bayes (NB)	XGBoost (XGB), KNN, Decision Tree (DT), Other ML models	34
Deep Learning	CNN, LSTM, BiLSTM	DNN, RNN, GRU, BiGRU, Composite DL models, Other DL models	50
Transformer-based	AraBERT, MARBERT, BERT	AraELECTRA, AraGPT2, QARiB, CAMeLBERT, DistilBERT, Other Transformer models	50
Hybrid-based	ML + DL, ML + Transformer, DL + Transformer, ML + DL + Transformer	—	24

Note: Values are the number of included studies ($n = 75$) using at least one model of each family; since a single study may use models from several families, the values are not mutually exclusive.

Table IV provides empirical evidence of the methodological evolution of Arabic fake news detection research. Classical machine learning models appear in 34 studies, where their role is mainly to provide robust and interpretable baselines rather than the most advanced modelling direction. The most frequently used classical classifiers are Random Forest (22 studies) and Support Vector Machine (20 studies), followed by Naïve Bayes (17), Logistic Regression (14), XGBoost (10), Decision Tree (9), and KNN (8). They are computationally efficient, transparent, and effective when combined with sparse feature representations such as TF-IDF, Bag-of-Words, and n -grams; therefore, they are still prevalent. The high frequency of deep learning models, used in 50 studies, points to a large shift towards automatic feature extraction and learning of dense representations. The prevalence of CNN, LSTM, Bi-LSTM, and composite deep learning models indicates that researchers are increasingly interested in modelling not only surface lexical cues but also sequential dependencies, contextual patterns, and semantic structures in Arabic fake news.

Another major direction is represented by transformer-based models, used in 50 studies. The most used transformer model was AraBERT (34 studies), followed by other BERT-based variants, MARBERT, AraELECTRA, AraGPT2, QARiB, CAMeLBERT, and DistilBERT, as summarized in Fig. 8. Recent work on Arabic fake news detection has increasingly used pretrained contextual language models to tackle the linguistic complexity of Arabic, including its morphology, dialectal variation, semantic ambiguity, and context-dependent meanings [65], [68], [75], [76]. Although hybrid models are less frequent (24 studies), their presence is methodologically important as it indicates an increasing trend of combining complementary modelling paradigms such as machine learning, deep learning, and transformer-based representations in unified detection frameworks.

Overall, the table contributes to the survey by not only describing the models but also providing a structured quantitative synthesis of the evolution of Arabic fake news detection research. These findings suggest that the field has not simply moved from classical machine learning to deep learning or transformers, but has instead evolved into a layered methodological landscape where traditional classifiers remain relevant for benchmarking, deep learning models facilitate automatic representation learning, transformers offer enhanced contextual understanding, and hybrid models aim to leverage these combined strengths. Thus, this frequency analysis offers a more informed basis for comparing model families, identifying dominant research trends, and highlighting the need for fair evaluation protocols, standardized Arabic benchmarks, cross-dataset validation, and ablation studies to determine whether reported performance improvements are due to the model architecture, feature representation, dataset characteristics, or experimental design.

7) *Comparative analysis and research implications*: The comparative analysis of the surveyed studies indicates a clear methodological evolution in Arabic fake news detection from classical machine learning models to deep learning, transformer-based architectures, and hybrid approaches [9], [58], [64], [85]. This evolution indicates a transition from handcrafted features to more context-aware and semantically richer representations. Classical machine learning models continue to provide value as interpretable and computationally efficient baselines. Deep learning models enable automatic feature learning, transformer-based models improve contextual understanding, and hybrid models aim to combine the complementary strengths of different modelling paradigms. This section thus contributes by reframing model comparison as a broader methodological assessment of interpretability, data requirements, contextual capacity, computational cost, generalization, and practical applicability, rather than a simple ranking of accuracy scores as shown in Table V.

TABLE V. COMPARATIVE ANALYSIS OF CLASSIFICATION MODEL FAMILIES IN ARABIC FAKE NEWS DETECTION

Aspect	ML	DL	Transformer	Hybrid
Feature Engineering	High	Medium	Low	Medium
Training Time	Low	Medium	High	Very High
Computational Cost	Low	Medium	High	Very High
Data Requirement	Low–Medium	Medium–High	High	High
Contextual Understanding	Low	Medium	Very High	Very High
Arabic Morphology Handling	Weak–Moderate	Moderate	Strong	Strong
Arabic Dialect Handling	Weak	Moderate	Strong	Strong
Interpretability	High	Low	Low	Low–Medium
XAI Evidence	Moderate	Limited	Limited	Limited
Generalization Evidence	Limited	Limited	Promising but under-tested	Promising but under-tested
Robustness to Domain Shift	Low	Medium	Medium–High	Medium–High
Reported Performance Trend	Moderate	High	Very High	Often highest within single studies
Main Limitation	Manual features	Data-hungry	Cost and low interpretability	Complexity and weak ablation evidence

Classical machine learning models like SVM, RF, LR, NB, and XGBoost still have an important role in Arabic fake news detection since they offer transparent, reproducible, and efficient baseline systems, as reported in several of the reviewed studies. The main advantage of these approaches is that they are effective on small to medium-sized datasets and can be combined with traditional feature representations, such as TF-IDF, Bag-of-Words, n-grams, lexical cues, and linguistic indicators [9]–[12], [49], [54]. However, their dependence on handcrafted features hinders their ability to capture semantic ambiguity, implicit meanings, long-range dependencies, dialectal variation, and pragmatic cues. Thus, classical machine learning should be preserved in future research as a necessary benchmark, but not as a sufficient one for modelling the linguistic and contextual complexity of Arabic misinformation.

Two major advances over traditional feature-based classification are deep learning and transformer-based models. Deep learning architectures such as CNN, LSTM, BiLSTM, GRU, and BiGRU reduce the need for manual feature engineering and allow models to learn local, sequential, and contextual patterns directly from data. CNN models are effective for the detection of phrase-level features, and LSTM and bidirectional models are more effective for modelling contextual dependencies in Arabic text [59], [61], [63], [64], [73], [74]. Deep learning models, however, require larger and more representative datasets, which remains a challenge for Arabic fake news detection due to the scarcity, imbalance, and limited dialectal coverage of datasets. Transformer-based models further enhance the contextual representation by self-attention mechanisms and pre-trained Arabic language models such as AraBERT, MARBERT, ARBERT, CAMeLBERT, AraELECTRA, QARiB-BERT, and GigaBERT [65], [68], [69], [76], [86]. These models offer improved semantic understanding but are costly in terms of computation, sensitive to fine-tuning parameters, and often poorly evaluated over datasets, dialects, domains, and periods of time.

Hybrid models emerge as one of the most promising directions in Arabic fake news detection because they combine complementary strengths from multiple model families. These models may integrate transformer-based embeddings with classical classifiers, deep learning architectures with machine learning classifiers, or transformer representations with CNN, LSTM, or BiLSTM layers [13], [15], [17], [76], [86]. Such combinations allow hybrid systems to benefit from semantic representation, automatic feature learning, sequential modeling, and robust classification boundaries. This is particularly relevant for Arabic fake news detection, where deceptive content may involve rich morphology, dialectal variation, emotional framing, stylistic manipulation, and implicit meanings. However, the strong performance reported by many hybrid models should be interpreted cautiously, as many studies do not provide sufficient ablation analysis to determine whether the improvement comes from the embedding model, the neural architecture, the classifier, preprocessing decisions, or dataset-specific factors.

The comparative results also highlight important gaps in the current research literature. While most studies still mainly focus on accuracy, precision, recall, and F1-score, less attention is paid to cross-dataset validation, explainability, temporal robustness, and multimodal detection. A model that performs well on one dataset may not generalize to another dataset collected from a different platform, topic, dialect, or time period. Similarly, high-performing deep learning, transformer-based, and hybrid models often operate as black-box systems, which limits their transparency and practical adoption. Moreover, most Arabic fake news detection studies remain text-centered, although misinformation in real online environments often includes images, videos, captions, metadata, user behavior, source credibility, and propagation patterns [19]. Therefore, future research should move beyond accuracy-centered evaluation toward robustness-centered evaluation by adopting standardized benchmarks, cross-dataset and cross-dialect testing,

temporal validation, ablation studies, explainability methods, and multimodal frameworks.

To provide a quantitative perspective on reported model performance, Table VI presents selected experimental results across four Arabic fake news datasets: AFND, ArCOV19-Rumors, ANS, and AraNews. The comparison includes representative machine learning, deep learning, transformer-based, and hybrid approaches, reporting their evaluation splits, accuracy, and F1-scores. These results were selected as representative examples from studies reporting accuracy and/or F1-score on these four datasets; the datasets, evaluation settings, and reported results of all 75 included studies are provided in Supplementary Table S1. Given the substantial methodological diversity among the reviewed studies, this focused comparison illustrates dataset-specific performance variations rather than a definitive ranking of model architectures.

As shown in Tabel VI, reported performance varies considerably across datasets and model architectures. For AFND, accuracy ranges from 90.45% to 98.42% across the selected studies, while results on ANS and AraNews show different performance patterns [14], [15], [54], [66], [74], [76], [86]. For example, AraBLA reports 95.09% accuracy on AFND, compared with 77.39% on ANS and 90.13% on AraNews [76]. However, these differences should not be interpreted as direct evidence of model superiority or cross-dataset generalization, since studies employ different data splits, dataset subsets, pre-processing procedures, and evaluation protocols. Moreover, differences in F1-score definitions further restrict comparability. These findings highlight the importance of standardized evaluation settings, transparent reporting, and cross-dataset validation when assessing progress in Arabic fake news detection. The limited generalization is illustrated by a study in which MARBERT reached an accuracy of 0.97 within one Arabic rumor dataset but only about 0.25–0.53 when tested on COVID-19 datasets from a different domain [32].

TABLE VI. COMPARATIVE PERFORMANCE OF SELECTED ARABIC FAKE NEWS DETECTION STUDIES ACROSS FOUR DATASETS

First Author (Year)	Dataset	Model	Split	Accuracy	F1-score
Suliman (2026) [15]	AFND	AraBERT-CNN	90/10	90.45%	89%
Alrayzah et al. (2026) [76]	AFND	AraBLA	80/10/10	95.09%	95.49%
Merzah et al. (2025) [74]	AFND	Dual BiGRU	NR	91.92%	92.66%
Al Mamlook et al. (2025) [54]	AFND subset	Optimized SVM	80/20	98.20%	98.1%
Turki et al. (2025) [86]	AFND	WaraBERT-BiLSTM	75/25	93.83%	NR
Aboulola & Umer (2024) [14]	AFND	ELMo-CNN-LSTM	80/20	98.42%	98.93%
Haouari et al. (2021) [25]	ArCOV 19-Rumors	MARBERT	80/20	75.70%	74%
Zamzami et al. (2026) [87]	ArCOV 19 sub-set	EMBL	NR	92.51%	90.89%

Alrayzah et al. (2026) [76]	ANS	AraBLA	80/10/10	77.39%	70.3%
Merzah et al. (2026) [66]	ANS	CNN-BiLSTM	80/20	71.63%	65.4%
Alrayzah et al. (2026) [76]	AraNews	AraBLA	80/10/10	90.13%	86.6%
Turki et al. (2025) [86]	AraNews	WaraBERT-BiLSTM	80/20	81.25%	NR

Overall, this comparative synthesis shows that no single model family can fully address the challenges of Arabic fake news detection. Classical machine learning models are essential for interpretability and baseline comparison, deep learning models improve automatic representation learning, transformer-based models enhance contextual understanding, and hybrid models provide a promising direction by integrating complementary modeling strengths. The main implication is that future progress should not depend only on achieving higher classification scores, but also on developing models that are explainable, reproducible, temporally stable, generalizable across domains and dialects, and applicable to real-world multimodal misinformation environments.

8) *Explainable fake news detection*: Explainable fake news detection has recently become an important direction in fake news research, particularly because high classification performance alone is not sufficient for trustworthy decision-making. In Arabic fake news detection, most existing studies have mainly focused on improving predictive performance using machine learning, deep learning, transformer-based models, and hybrid architectures. However, relatively few studies have explicitly incorporated explainability methods to clarify why a model classifies a given Arabic news item as fake or real. This limitation is particularly critical in Arabic because linguistic complexity, dialectal variation, morphology, emotional framing, and context-dependent meanings affect fake news detection [50]. Therefore, explainability is a research direction that is under-explored but necessary for Arabic fake news detection [88]. It can assist researchers, fact-checkers, and end users in understanding the degree to which the model is relying on meaningful linguistic and contextual cues or dataset-specific artefacts. Explanations can be generated using post hoc methods such as SHAP and LIME, or through model-based visualization techniques such as attention visualization in BERT- and AraBERT-based models.

SHAP is one of the most popular explanation techniques for interpreting machine learning and deep learning predictions. It estimates the contribution of each attribute, word, or symbol to the final classification decision based on Shapley values. In the fake news detection domain, SHAP is applied to identify textual, contextual, or metadata features that cause a model to classify a news item as fake or true [88], [89]. This allows researchers to examine whether the model detects deceptive cues (emotional language, suspicious claims, source citations, language inconsistencies, etc.). SHAP can provide local interpretations for individual news items and comprehen-

sive interpretations that summarize the most influential features across a dataset [88], [89]. However, its application to Arabic fake news detection remains challenging, as Arabic texts often exhibit rich morphological structure, dialectal variations, and subword segmentation in transformer-based models. By contrast, SHAP-based interpretation has been applied more widely in English-language fake news detection [89]. SHAP is also computationally expensive, especially for neural networks or large-scale transformer-based architectures.

LIME is another important post-hoc explainability method that has been applied in fake news detection research. Unlike SHAP, which estimates feature contribution using a game-theoretic framework, LIME explains a specific prediction by perturbing the input text and observing how the model prediction changes. In text classification, this usually means removing or modifying words and then estimating which words have the strongest influence on the model's decision. LIME is particularly useful for fake news detection because it provides simple and intuitive local explanations that show the words or phrases contributing to a fake or real classification [14], [88], [90]. In Arabic fake news detection, LIME can help reveal whether the model focuses on relevant Arabic linguistic signals or on superficial dataset-specific patterns. Nevertheless, LIME has limitations. Its explanations are local rather than global; it is sensitive to the perturbation strategy, and its generated explanation may not always faithfully represent the internal reasoning of complex models. These limitations become more important when dealing with Arabic transformer models, where words may be split into subword units that are harder to interpret directly.

Attention visualization provides another form of interpretability, especially for attention-based deep learning and transformer models. Transformer-based architectures are based on the self-attention mechanism to model the relationships among tokens. Visualizing attention weights can help researchers inspect which parts of the input text the model attends to most during classification [88]. In fake news detection, attention visualization may show whether the model focuses on emotional expressions, named entities, claim-related phrases, source indicators, or suspicious linguistic patterns. This can be useful as a diagnostic tool for understanding the behavior of BERT-based and AraBERT-based models. However, attention weights should be interpreted cautiously because attention is not always equivalent to explanation. A token receiving close attention does not necessarily mean that it is the true causal reason behind the model's decision. Therefore, attention visualization should be used as a supporting interpretability technique rather than as complete evidence of model transparency.

Explainable BERT-based fake news detection is important because BERT and its variants have achieved strong results in many NLP classification tasks, but are often difficult to interpret due to their multi-layer transformer architecture and contextual embedding mechanisms [88]. In fake news detection, explainable BERT approaches attempt to make these models more transparent by using techniques such as LIME, SHAP, Anchors, attention maps, token attribution, and hidden-state visualization. These methods can help identify the words, tokens, or contextual patterns that influence the prediction of a BERT-based fake news detector. For Arabic, this direction

becomes even more important with models such as AraBERT, MARBERT, CAMeLBERT, QARib-BERT, and AraELECTRA. Explainable AraBERT aims to clarify how Arabic-specific transformer models classify fake and real Arabic news. However, Arabic explainability faces other challenges, including dialectal variation, normalization effects, morphology, code-switching, sarcasm, and subword tokenization. For instance, a model may emphasize subword fragments over complete Arabic words, which can lead to a less comprehensible explanation for human users. Therefore, explainable AraBERT requires Arabic-aware explanation visualization, subword aggregation, and evaluation by Arabic-speaking annotators or fact-checkers.

Based on the reviewed Arabic fake news detection studies, explainability remains limited compared with classification performance evaluation. Most previous studies have concentrated on enhancing accuracy, precision, recall, and F1-score via machine learning, deep learning, transformer-based models, and hybrid architectures. However, only 3 of the 75 included studies (4%) explicitly incorporated XAI methods, using LIME and/or SHAP [14], [90], [91], to explain why a model classifies an Arabic news item as fake or real. This indicates that explainable Arabic fake news detection is still at an early stage. The main challenges are: the limited adoption of XAI methods, the lack of standardized evaluation of the quality of explanations, the difficulty of interpreting Arabic subwords, the computational cost of explanation methods, and the absence of human-centered evaluation with Arabic users or professional fact-checkers. Therefore, future work must move beyond accuracy-focused evaluation and embrace explainability as a central dimension of model evaluation. Future directions include applying SHAP and LIME to AraBERT, MARBERT, CAMeLBERT, and LLM-based Arabic fake news detectors, developing Arabic-specific explanation visualizations, evaluating explanations with human fact-checkers, combining textual explanations with metadata, temporal and multimodal features, and assessing whether explanations are stable across datasets, dialects, domains, and time periods.

V. DATASETS

This section presents an overview of the datasets used in the literature of Arabic fake news detection. It discusses the datasets' sources, domains, linguistic coverage, size, balance, and how they were used in the reviewed studies. It also discusses the most significant dataset-related limitations affecting model performance, comparability, and generalization, and highlights future directions for building more reliable Arabic fake news datasets.

A. Importance of Datasets in Arabic Fake News Detection

Datasets are essential for Arabic fake news detection as they directly impact model learning, evaluation, and generalization. In the reviewed studies, datasets were used to identify, classify, and analyze fake and real Arabic news from different sources and domains [13], [74], [79]. However, the availability of Arabic fake news datasets is still limited and unevenly developed, due to several factors: the diversity of Arabic dialects, the limited availability of verified sources, and the variation of misinformation topics across political, health, social, and news

domains. These factors influence the construction of datasets and the reliability and comparability of the model results.

The value of a dataset is therefore not limited to its size. A reliable dataset should provide sufficient labeled examples, clear source information, transparent class distribution, and linguistic diversity in Arabic fake news detection [7], [8], [77], [92]. These characteristics are particularly important because Arabic has rich morphology, extensive vocabulary, and substantial variation between Modern Standard Arabic (MSA), dialectal Arabic, Arabizi (Arabic written in Latin script), and informal social media writing. When datasets do not capture this linguistic and contextual diversity, models can achieve high scores on a single benchmark but perform poorly when deployed to new topics, platforms, or dialects.

B. Dataset Sources, Domains, and Linguistic Coverage

The reviewed datasets were collected from heterogeneous sources, including online news websites, newspapers, Twitter/X, Facebook, YouTube comments, fact-checking platforms, crowdsourced corpora, and translated or auxiliary benchmark datasets [5], [21], [26], [93]. This diversity is useful because fake news does not appear in one fixed format or platform. News articles tend to contain formal MSA and longer textual structures, whereas social media datasets often contain short posts, dialectal expressions, informal spelling, hashtags, emojis, user interactions, and temporal propagation patterns. As a result, the source of the dataset can strongly influence the linguistic patterns learned by a model.

Domain coverage is another important dataset feature. Several Arabic fake news datasets focus on a specific topic, such as COVID-19, vaccination, political rumors, Syrian crisis content, Hajj-related news, healthcare claims, or financial misinformation [2], [3], [25], [27], [92], [94]. Although domain-specific datasets are valuable for targeted misinformation detection, they may limit model generalizability. The models are trained on a narrow domain and learn topic-specific vocabulary rather than more general deception signals. Therefore, to evaluate whether a model can generalize beyond the dataset it was

trained on, multi-domain datasets and cross-domain evaluation are required.

Another major challenge is dialect coverage. Arabic varies considerably across regions and registers, including Gulf, Egyptian, Levantine, and Maghrebi dialects, Classical Arabic, and MSA [7], [77], [95]. The studies show that transformer-based models, such as AraBERT and MARBERT, can enhance contextual understanding, especially when they are trained or fine-tuned on data containing different dialectal varieties. However, models trained mainly on MSA or formal news sources tend to perform poorly on dialectal or social media text. Hence, dialect-aware dataset construction is key for robust Arabic fake news detection.

C. Dataset Landscape in the Reviewed Studies

The dataset landscape for Arabic fake news detection research consists of a mix of large public news datasets, social media corpora, task-specific benchmarks, and smaller manually annotated datasets. The most utilized and influential among them is the Arabic Fake News Dataset (AFND) in the reviewed studies. AFND is a collection of 606,912 Arabic news articles from 134 public Arabic news websites in 19 Arab countries [96]. It contains three types of credibility: credible, not credible, and undecided. Despite its large scale and wide geographic coverage, its three classes are unevenly distributed. Moreover, AFND labels reflect the credibility of the publishing source rather than the verification of individual claims, so models trained on it may learn source-specific style rather than veracity, which limits generalization.

Table VII summarizes The main datasets identified in the primary studies. The table reports dataset size, source, class balance, and usage count across the reviewed papers. The usage count is the number of included studies that used each dataset; because a study may use several datasets and many studies rely on custom datasets, the counts are not expected to sum to 75. This overview supports a comparative analysis by showing which datasets dominate the literature, which sources are most frequently used, and where the field still depends on small, imbalanced, or domain-specific datasets.

TABLE VII. SUMMARY OF ARABIC FAKE NEWS DETECTION DATASETS IDENTIFIED IN THE REVIEWED STUDIES ACCORDING TO SIZE, SOURCE, CLASS DISTRIBUTION, AND REPORTED USAGE FREQUENCY

Dataset Name	Records	Dataset Source	Balanced?	Usage Count
Arabic Fake News Dataset (AFND / AFN) [96]	606,912	134 public Arabic news websites across 19 Arab countries	imbalanced	17
AraNews Dataset [55]	123,219	Arabic news articles from newspapers	imbalanced	11
ArCOV19-Rumors [25]	9,414	Twitter/X	imbalanced	10
Arabic News Stance Corpus (ANS) [97]	4,546	Arabic news stance	imbalanced	6
COVID-19-Fakes [98]	4,954	Twitter/X	imbalanced	5
JoNewsFake	50,000	Facebook posts	imbalanced	3
AraCOVID19-MFH [27]	10,828	Twitter/X	imbalanced	2
Kaggle Arabic Fake and Real News Articles	12,790	Kaggle Arabic fake/real news article dataset.	imbalanced	1
Arabic Satirical Fake News Dataset [39]	6,895	Satirical articles	balanced	3
Translated Kaggle Fake-News Dataset	20,000	English Kaggle fake-news dataset translated	balanced	2
AFND-balanced	60,000	Derived subsets from AFND.	balanced	3

D. Recurrent Dataset Limitations

In addition to listing datasets, this survey highlights recurring dataset-level limitations that affect the validity, comparability, and reproducibility of Arabic fake news detection results [3], [67], [79], [99]. The most frequent issues are related to limited dataset size, class imbalance, narrow domain coverage, weak dialectal representation, insufficient multimodal/contextual information, and incomplete reporting. These

limitations explain why high performance in one study may not necessarily indicate robust performance across Arabic misinformation contexts. Table VIII summarizes these recurring limitations and connects each dataset-level problem to its likely impact on model performance and to a concrete direction for future dataset construction. This positioning moves the discussion from a descriptive dataset landscape to a critical comparative analysis of the structural problems that repeatedly appear across Arabic fake news datasets.

TABLE VIII. KEY LIMITATIONS OF ARABIC FAKE NEWS DATASETS AND THEIR IMPLICATIONS

Main Dataset Limitation	Common Problem	Impact on Model Performance	Suggested Direction
Limited dataset size	Many Arabic datasets contain a relatively small number of labeled samples.	Increases the risk of overfitting and limits the reliability of deep learning and transformer-based models.	Construct larger, manually verified, and publicly reusable Arabic datasets.
Class imbalance	Fake and real classes are often unevenly distributed.	Accuracy may become misleading, and models may become biased toward the majority class.	Report macro-F1 and class-wise metrics; use balanced sampling or cost-sensitive learning.
Domain specificity	Several datasets focus on one topic, such as COVID-19, politics, health, or rumors.	Models may learn topic-specific cues instead of general fake news patterns.	Use multi-topic datasets and evaluate models under cross-domain settings.
Limited dialectal coverage	Many datasets depend mainly on MSA or formal news sources.	Reduces robustness on dialectal Arabic and informal social-media text.	Include multiple Arabic dialects, code-switched text, and social media language varieties.
Lack of multimodal and contextual features	Most datasets are text-only and lack images, metadata, and temporal signals.	Important misinformation cues are ignored, especially in social media contexts.	Develop multimodal datasets combining text, images, metadata, timestamps, and interaction signals.
Poor reporting and reproducibility	Some studies do not clearly report data size, source, label distribution, preprocessing, or train/test splits.	Makes fair comparison between studies difficult and reduces reproducibility.	Use dataset cards, public splits, transparent annotation protocols, and clear availability statements.

E. Implications for Future Arabic Fake News Datasets

The surveyed dataset landscape shows that future research on Arabic fake news detection should move from isolated benchmark evaluation to more robust and transparent dataset design. An ideal dataset should be large enough to cover modern models, balanced enough for fair evaluation, diverse enough to cover multiple topics and platforms, and linguistically rich enough to represent MSA, dialectal Arabic, informal writing, and code-switched forms [13], [74], [87]. Such characteristics are necessary to build models that can generalize to real-world misinformation scenarios, not just perform well on a single curated dataset.

Another important direction is to incorporate multimodal and contextual signals. Many fake news items circulate with images, videos, source metadata, user reactions, publication time, and propagation patterns. Therefore, future Arabic datasets should not be restricted to text when the target environment is social media or online news ecosystems [4]–[6], [46]. More comprehensive datasets can support stronger transformer-based, hybrid, and multimodal approaches, while also enabling fairer comparisons across machine learning, deep learning, and transformer-based models.

In general, this section shows that the quality of the dataset is one of the main factors that influence the progress in Arabic fake news detection. The field has been bolstered by large and widely used datasets, such as AFND, AraNews, and ArCOV19-Rumors, but still suffers from imbalance, limited dialectal coverage, domain restriction, and reporting inconsistencies. Addressing these limitations is important for building

reliable, reproducible, and publication-worthy research for Arabic fake news detection.

VI. EVALUATION METRICS

Evaluation metrics are needed to measure the reliability, comparability, and practical effectiveness of Arabic fake news detection models. Since the reviewed studies use different modelling paradigms (e.g., machine learning, deep learning, transformer-based models, and hybrid architectures), there is a need for consistent evaluation criteria to compare studies fairly and to determine whether the improvements reported are due to real progress or dataset-specific effects [100], [101]. This is particularly true in Arabic fake news detection, where datasets usually vary in size, class distribution, domain, source, dialectal coverage, and annotation strategy. Thus, the use of a single metric such as accuracy can lead to an incomplete or even misleading interpretation of model performance, especially when dealing with imbalanced datasets. Therefore, the evaluation of Arabic fake news detection systems should include a set of complementary metrics such as accuracy, precision, recall, F1-score, Macro-F1, and AUC-ROC [65], [102], [103].

A. Confusion Matrix

The confusion matrix provides the foundation for most classification metrics. In binary Arabic fake news detection, predictions are usually organized into four outcomes:

True Positives (TP): fake news samples correctly classified as fake.

True Negatives (TN): real news samples correctly classified as real.

False Positives (FP): real news samples incorrectly classified as fake.

False Negatives (FN): fake news samples incorrectly classified as real.

The four outputs are used to calculate the main assessment metrics and to analyze the types of errors made by the model. This distinction is important as the ramifications of false positives and false negatives are different in the disinformation detection setting: the former can erode trust in legitimate sources, and the latter allows the continued presence of incorrect content.

B. Standard Classification Metrics

Accuracy is the ratio of the total number of correct predictions to the total number of predictions evaluated. It is widely reported, but it should be carefully interpreted when datasets are imbalanced, as a model might achieve high accuracy by favoring the majority class. It is defined in Eq. (2) [102].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision indicates the proportion of samples predicted as fake that are actually fake. In fake news detection, high precision means that the model produces fewer false alarms when labeling content as fake, as defined in Eq. (3) [102].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

Recall, also known as sensitivity or true positive rate, measures the ability of the model to identify all fake news samples, as defined in Eq. (4) [102]. High recall is especially important when the main objective is to reduce the risk of missing misleading content.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

F1-score is the harmonic mean of precision and recall. It provides a more balanced measure of performance when false positives and false negatives are both important, and it is defined in Eq. (5) [102]. It is more informative than accuracy in many fake news detection scenarios.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

C. Metrics for Imbalanced Datasets

Macro-F1 is particularly useful for evaluating a model's robustness across classes, since many Arabic fake news datasets are imbalanced; it is defined in Eq. (6) [102]. Macro-F1 calculates the F1-score for each class separately and gives equal importance to each class, unlike a standard or weighted average, which can be dominated by the majority class. This makes it more suitable for finding if a model performs well on both fake and real news classes.

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (6)$$

where, C is the number of classes, and $F1_c$ is the F1-score computed for class c .

D. Threshold-Independent Evaluation

AUC-ROC is a threshold-independent measure of a model's ability to distinguish fake news from real news. It is more robust than accuracy in measuring performance when class imbalance or threshold selection may impact performance, as it describes the trade-off between the true positive rate and the false positive rate.

E. Summary of Evaluation Metrics

Table IX summarizes The main evaluation metrics and their relevance to Arabic fake news detection. Using these metrics together provides a more complete evaluation than relying on any single measurement.

TABLE IX. SUMMARY OF EVALUATION METRICS USED IN ARABIC FAKE NEWS DETECTION STUDIES

Metric	Main Formula / Basis	Evaluation Role	Relevance to Arabic Fake News Detection
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$	Measures overall correctness.	Useful as a general indicator, but may be misleading under class imbalance.
Precision	$TP / (TP + FP)$	Measures the correctness of fake-news predictions.	Reduces false alarms when reliable news is wrongly classified as fake.
Recall	$TP / (TP + FN)$	Measures the ability to detect fake news.	Important for minimizing undetected misinformation.
F1-score	Harmonic mean of precision and recall	Balances precision and recall.	Useful when false positives and false negatives are both important.
Macro-F1	Average of class-wise F1-scores	Evaluates all classes equally.	Recommended for imbalanced Arabic fake news datasets.
AUC-ROC	Area under the ROC curve	Measures threshold-independent discrimination.	Useful for comparing models across decision thresholds.

In general, the choice of evaluation metrics should be driven by the characteristics of the dataset and the aim of the detection task. Accuracy provides a general summary of performance, but it is not sufficient on its own if there is class imbalance, domain bias, or uneven dialect representation. Precision and recall clarify the trade-off between false alarms and missed fake news, while F1-score and Macro-F1 provide stronger evidence of balanced performance across classes. AUC-ROC further supports model comparison by measuring discrimination ability independently of a fixed decision threshold. Conse-

quently, Arabic fake news detection studies should report multiple complementary metrics to enable fairer comparison among machine learning, deep learning, transformer-based, and hybrid models and to provide a more reliable assessment of their practical effectiveness [3], [103].

VII. SYNTHESIS OF FINDINGS, RESEARCH GAPS, AND FUTURE DIRECTIONS

Based on the findings synthesized from the reviewed literature, this section presents an integrated discussion of the cur-

rent state, unresolved challenges, and future directions of Arabic fake news detection. In the first part, the main findings of the survey are explicitly linked to the predefined research questions by means of a research-question-based synthesis table (Table X). This table sheds light on the way each question is answered across the previous sections. The section then goes on to discuss the key research gaps and methodological challenges that continue to hinder progress in the field, especially those concerning dataset availability and quality, dialectal variation, cross-domain generalization, evaluation standardization, metadata and temporal modelling, multimodal learning, explainability, and real-world applicability. The section ends with a discussion on future research directions to improve the reliability, robustness, interpretability, and deployment of Arabic fake news detection systems in real-world scenarios. The purpose of this synthesis is to guide future research to develop more comprehensive and effective detection frameworks for the Arabic language that may help mitigate the harms to society, politics, and information posed by the spread of fake news.

A. Synthesis of Findings in Relation to the Research Questions

Table X provides a research-question-based synthesis of the main findings of this survey. Overall, the reviewed studies show a clear shift from classical machine learning and conventional deep learning models toward transformer-based and hybrid approaches, which generally provide stronger contextual and semantic representations for Arabic fake news detection. However, direct comparisons of studies are hindered by variations in datasets, preprocessing methods, evaluation metrics, and validation strategies. The synthesis also highlights the importance of dataset characteristics like size, source, annotation quality, class balance, dialectal coverage, and accessibility in reported performance and generalization. These results indicate that future progress will require not only stronger models but also more reliable datasets, standardized benchmarks, and consistent evaluation protocols.

TABLE X. SYNTHESIS OF THE MAIN FINDINGS IN RELATION TO THE RESEARCH QUESTIONS

Research Question	Main Findings	Sections Addressing It
RQ1	Arabic fake news detection employs ML, DL, transformer-based, and hybrid models. Although transformers and hybrid approaches often report better results, differences in datasets and evaluation protocols limit direct comparisons. Dataset-stratified results are reported in Table VI and per-study results in Table S1.	Sections IV and VI
RQ2	Transformer-based models often report stronger performance in Arabic fake news detection due to their contextual understanding. However, high computational costs, limited interpretability, and insufficient cross-dataset validation remain key challenges. Table VI shows that reported performance varies strongly across datasets, even for the same model.	Section IV
RQ3	The reviewed studies commonly use Arabic text preprocessing, followed by traditional feature representations (TF-IDF, BoW, and n-grams) or embeddings, including AraBERT and MARBERT.	Section IV
RQ4	LLMs offer new opportunities for Arabic fake news detection through prompt-based classification, explanation, augmentation, and evidence verification, but their adoption remains limited: only 10 of the 75 studies (13%) used LLMs or generative models, mostly GPT-2/AraGPT2 for generation or augmentation, and only two evaluated GPT-4 directly. Adoption is further challenged by hallucination, cost, prompt sensitivity, and weak benchmarking.	Section IV
RQ5	Explainability remains limited in Arabic fake news detection: only 3 of the 75 studies (4%) applied XAI methods (LIME and/or SHAP). SHAP, LIME, and attention visualization help interpret predictions, but challenges persist in explanation reliability, Arabic subword interpretation, computational cost, and human validation.	Section IV
RQ6	Reported performance and model comparability depend on dataset quality, size, balance, and diversity. Small, imbalanced, or poorly annotated datasets limit generalization and reliable cross-study comparisons. For example, the same model reaches 95.09% accuracy on AFND but 77.39% on ANS (Table VI).	Sections IV and V
RQ7	Key dataset challenges include scarcity, class imbalance, limited dialectal diversity, inconsistent annotation, and lack of standardized benchmarks. Future research should prioritize larger, accessible, diverse, and well-annotated Arabic datasets.	Sections V and VII

Beyond text-only classification, few of the 75 included studies address the other dimensions outlined in the Introduction. Two studies combined text with images [16], [17]; one analyzed the temporal and geographic propagation of fake news [31], and one dataset provides propagation networks [25]; one study addressed claim verification using evidence from authorities [34]; and others exploited reader comments [29], [48] or user-level signals [104]. This scarcity is itself a key research gap and motivates the future directions below.

B. Limitations of this Review

This review has several limitations. First, the review protocol was not registered in advance (e.g., in PROSPERO or the Open Science Framework). Second, screening, full-text assessment, and quality scoring were performed by the first author and verified by the co-authors, with disagreements re-

solved by consensus; formal inter-rater agreement statistics (e.g., Cohen's kappa) were therefore not computed. Third, the original search strings did not include terms such as rumor, misinformation, disinformation, and fact-checking, nor Arabic-script terms, and the records retrieved from each database were not counted separately before de-duplication; the missing terms were partly mitigated by backward and forward snowballing and by an updated search during revision, but some relevant studies may still have been missed. Fourth, the review was restricted to peer-reviewed journal and conference papers, which may exclude some shared-task system descriptions and grey literature. Fifth, the full texts of a small number of potentially eligible studies could not be retrieved. Finally, because the included studies differ in datasets, splits, and metrics, a formal meta-analysis was not feasible, and the performance

comparison is restricted to a dataset-stratified subset of studies (Table VI).

C. Challenges and Future Directions

Based on previous research, this section examines the challenges facing Arabic-language fake news detection models, their causes, and potential solutions to overcome them, as well as future research directions that can be explored. This is intended to help researchers develop highly efficient and effective Arabic-language models comparable to those in English, allowing them to be integrated into practical applications and mitigate the negative impact of fake news. These challenges include the following:

The task of Arabic fake news detection faces major challenges related to data quality and diversity. There are fewer large, high-quality Arabic datasets than English datasets, and most existing datasets suffer from class imbalance between true and false news [7]–[9], [91], [92]. Additionally, most datasets cover only a limited number of dialects and focus mainly on Modern Standard Arabic, which reduces their usefulness in diverse settings. Crowd-sourced datasets often contain labeling biases, and there is a lack of standardized benchmark datasets for equitable comparisons between studies. To tackle these challenges, subsequent efforts must focus on developing extensive, multi-domain, and multi-dialect Arabic corpora that encompass social media content and multimodal data, including text, images, and videos. To facilitate the comparison and replication of studies, researchers should create benchmark datasets and standardized evaluation frameworks that are openly accessible to all. These efforts will enhance Arabic fake news detection systems, making them stronger, fairer, and more effective for real-life applications.

Arabic fake news detection faces significant linguistic challenges because of the language's complex morphology. Arabic has a highly productive morphology and a root-and-pattern structure, which increases lexical variation and complicates tokenization and stemming [9], [24], [50]. Text normalization and model training are more challenging because there are no diacritics, words are often ambiguous, and there are multiple ways to inflect the words. Additionally, the presence of different dialects with slang on social media complicates the generalization of models, and linguistic tools such as lemmatizers and sentiment lexicons are still not very effective. Therefore, creating Arabic-specific resources for building datasets and developing effective morphological analyzers and normalization tools that encompass Modern Standard Arabic, its various dialects, and synonyms are among the most important tasks that researchers should focus on in the future. Improving the models' ability to handle sentiment analysis and understand contextual content will enhance the accuracy of fake news detection. If researchers can address these linguistic and semantic challenges and fully comprehend Arabic, they will be able to develop effective algorithms for detecting fake news in that language.

The feature extraction process is one of the most challenging aspects of detecting fake Arabic news. This is because it requires an understanding of the context and selection of relevant features of the Arabic language to detect both fake and real news. This knowledge is important for understanding the

meaning of sentences and words, which helps determine whether a text is genuine [87]. In this respect, traditional models have not produced satisfactory outcomes, and even hybrid models encounter difficulties in extracting relevant features in Arabic. Future researchers should enhance the existing models employed in the feature extraction process of previous studies, such as hybrid and transformer models, or develop new models that can improve the feature selection process in Arabic. They may also explore methods used in other languages and apply them to Arabic. Such cross-lingual transfer would allow Arabic research to benefit from experience gained in other languages and to compare results when developing effective detection models.

Another challenge lies in the modeling and classification stages, where very high reported results may reflect limited data or bias towards a particular class and may therefore indicate overfitting [5], [65]. This issue diminishes the use and reliability of models across various regions and dialects. Additionally, deep learning and transformer-based models can act like black boxes, meaning they lack transparency in their inner workings, making them difficult to understand. Therefore, future research should test whether hybrid models that combine transformer and deep learning components genuinely improve robustness, using ablation and cross-dataset evaluation. Additionally, utilizing domain adaptation and continuous learning will enable the models to remain current with evolving trends in Arabic fake news detection.

Finally, detection models could be integrated with social media platforms and practical applications in health, sports, and other domains to help limit the spread of false content. Furthermore, combining these systems with user behavior analysis and interaction patterns will enhance detection accuracy and adaptability, contributing to the development of effective models for detecting Arabic fake news.

VIII. CONCLUSION

This survey presented a systematic and taxonomy-based synthesis of Arabic fake news detection research published between January 2019 and April 2026. It examined the field along several dimensions, such as preprocessing strategies, feature extraction methods, model families, datasets, evaluation metrics, and open challenges by reviewing 75 studies. The reviewed literature shows a clear methodological transition from traditional machine learning models using handcrafted features to deep learning, transformer-based models, and hybrid architectures. Deep learning models enable automatic feature learning while classical machine learning models serve as interpretable and computationally efficient baselines. The transformer-based models provide stronger contextual and semantic representations for Arabic text. Hybrid models have become a promising direction, combining the strengths of different modelling paradigms. However, the survey also reveals that reported performance cannot be ascribed to model architecture alone, as it is strongly influenced by dataset size, source, annotation quality, class balance, dialectal coverage, preprocessing choices, feature representations, and evaluation protocols.

Overall, the findings indicate that Arabic fake news detection has made substantial progress, but several limitations continue to restrict its robustness, comparability, and real-world

applicability. The field still suffers from the lack of large, balanced, publicly available, and dialectally diverse datasets, limited use of standardized benchmarks, cross-dataset validation, ablation studies, temporal evaluation, explainability, and multimodal modelling. Future work should thus focus on creating high-quality Arabic datasets, transparent annotation procedures, fair evaluation protocols, and models that generalize across domains, dialects, platforms, and time periods. More attention is needed to the integration of textual content with metadata, temporal signals, source credibility, user behavior, propagation patterns, and visual information. Addressing these directions would help produce Arabic fake news detection systems that are accurate, explainable, reproducible, robust, and ready for deployment in real-world fact-checking, news verification, and social media monitoring scenarios.

DECLARATIONS

Data availability: The supplementary materials supporting this review are openly available in the Zenodo repository at <https://doi.org/10.5281/zenodo.23474152>. They comprise Supplementary Table S1 (characteristics and reported results of the 75 included studies), Table S2 (quality assessment scores), Table S3 (quality assessment checklist and scoring rules), Table S4 (reports excluded at full-text stage with reasons), and Table S5 (PRISMA 2020 checklist).

Funding: This research received no external funding.

Conflicts of interest: The authors declare no conflicts of interest.

REFERENCES

- [1] E. C. Tandoc Jr., Z. W. Lim, and R. Ling, "Defining 'Fake News': A typology of scholarly definitions," *Digital Journalism*, vol. 6, no. 2, pp. 137–153, 2018, doi: 10.1080/21670811.2017.1360143.
- [2] S. Mahmoud, A. E. Aboutabl, and E. H. Mohamed, "Arafakedetect: enhancing fake health news detection with ensemble learning on AraCovidVac," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, Art. no. 237, Dec. 2024, doi: 10.1007/s13278-024-01390-y.
- [3] E. Mohamed, W. N. Ismail, and E. O. Eldawy, "Enhancing Arabic healthcare fake news detection with data augmentation and multi-metric analysis using large language models," *Sci. Rep.*, vol. 16, Art. no. 5364, 2026, doi: 10.1038/s41598-025-21733-9.
- [4] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, "Fake news on Social Media: the Impact on Society," *Information Systems Frontiers*, vol. 26, no. 2, pp. 443–458, Apr. 2024, doi: 10.1007/s10796-022-10242-z.
- [5] A. Awajan, "Enhancing Arabic fake news detection for Twitter's social media platform using shallow learning techniques," *J. Theor. Appl. Inf. Technol.*, vol. 101, no. 5, 2023. [Online]. Available: <http://www.jatit.org/volumes/Vol101No5/12Vol101No5.pdf>
- [6] R. Vicari and N. Komendantova, "Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters," *Humanit. Soc. Sci. Commun.*, vol. 10, Art. no. 332, 2023, doi: 10.1057/s41599-023-01838-0.
- [7] I. Boulesnam and R. Boucetti, "Arabic Language Characteristics that Make its Automatic Processing Challenging," *International Arab Journal of Information Technology*, vol. 22, no. 4, pp. 814–831, Jul. 2025, doi: 10.34028/iajit/22/4/14.
- [8] A. Y. Muaad, S. Raza, U. Naseem, and H. J. J. Davanagere, "Arabic text detection: a survey of recent progress, challenges and opportunities," *Applied Intelligence*, vol. 53, no. 24, pp. 29845–29862, Dec. 2023, doi: 10.1007/s10489-023-04992-9.
- [9] M. Sellami, N. Kamel, and A. Lakhfif, "Fake news detection in Arabic: a survey of linguistic challenges, resources, and classification approaches," *Int. J. Mach. Learn. Cybern.*, vol. 17, no. 7, Art. no. 321, Jul. 2026, doi: 10.1007/s13042-026-03144-y.
- [10] T. Duridi, D. Eleyan, A. Eleyan, and T. Bejaoui, "Arabic Fake News Detection Using Machine Learning Approach," in *Proc. 2024 Int. Symp. Networks, Computers and Communications (ISNCC)*, IEEE, 2024, pp. 1–7, doi: 10.1109/ISNCC62547.2024.10758936.
- [11] E. S. Alboutoush, K. H. Gan, and S. A. A. Alrababah, "Evaluation of Machine Learning and Deep Learning Models for Fake News Detection in Arabic Headlines," *IEEE Access*, vol. 13, pp. 162009–162026, 2025, doi: 10.1109/ACCESS.2025.3606114.
- [12] E. Aljohani, "Enhancing Arabic Fake News Detection: Evaluating Data Balancing Techniques Across Multiple Machine Learning Models," *Engineering, Technology and Applied Science Research*, vol. 14, no. 4, pp. 15947–15956, Aug. 2024, doi: 10.48084/etasr.8019.
- [13] A. Jaber, T. Duridi, E. Daraghmi, and P. Martínez, "Fine-Tuning Arabic BERT Models with Hybrid CNN–BiLSTM Architecture for Disinformative Tweet Detection," *Arab. J. Sci. Eng.*, vol. 51, no. 9, pp. 12285–12297, May 2026, doi: 10.1007/s13369-025-11031-x.
- [14] O. Ibrahim Aboulola and M. Umer, "Novel approach for Arabic fake news classification using embedding from large language features with CNN-LSTM ensemble model and explainable AI," *Sci. Rep.*, vol. 14, Art. no. 30463, 2024, doi: 10.1038/s41598-024-82111-5.
- [15] H. M. Suliman, "Arabic Fake News Detection Using a Hybrid Arabert and 2-D CNN Architecture," *Journal of Academic Research (Majallat al-Buhuth al-Akadiimiyya)*, vol. 30, no. 1, pp. 101–111, 2026, doi: 10.65540/jar.v30i1.1439.
- [16] R. M. Albalawi, A. T. Jamal, A. O. Khadidos, and A. M. Alhothali, "Multimodal Arabic rumors detection," *IEEE Access*, vol. 11, pp. 9716–9730, 2023, doi: 10.1109/ACCESS.2023.3240373.
- [17] K. Tabet, K. E. Daouadi, A. Bennour, A. Brek, and A. Taleb-Ahmed, "MAFIA-Net: Multimodal Arabic fake-news identification via hybrid attention networks," *Array*, vol. 30, Art. no. 100924, 2026, doi: 10.1016/j.array.2026.100924.
- [18] S. Althabiti, M. A. Alsalka, and E. Atwell, "A Survey: Datasets and Methods for Arabic Fake News Detection," *Int. J. Islamic Appl. Comput. Sci. Technol. (IJASAT)*, vol. 11, no. 1, pp. 19–28, 2023.
- [19] E. S. Alboutoush, K. H. Gan, and S. A. A. Alrababah, "Fake news detection: state-of-the-art review and advances with attention to Arabic language aspects," *PeerJ Comput. Sci.*, vol. 11, Art. no. e2693, 2025, doi: 10.7717/peerj-cs.2693.
- [20] I. Touahri and A. Mazroui, "Survey of machine learning techniques for Arabic fake news detection," *Artif. Intell. Rev.*, vol. 57, no. 6, Art. no. 157, Jun. 2024, doi: 10.1007/s10462-024-10778-3.
- [21] M. Abbas Yousef, A. ElKorany, and H. Bayomi, "Fake-news detection: a survey of evaluation Arabic datasets," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, Art. no. 225, 2024, doi: 10.1007/s13278-024-01370-2.
- [22] M. Amjad, S. Butt, A. Zhila, G. Sidorov, L. Chanona-Hernandez, and A. Gelbukh, "Survey of Fake News Datasets and Detection Methods in European and Asian Languages," *Acta Polytechnica Hungarica*, vol. 19, no. 10, 2022.
- [23] B. M. Nema and M. H. Al-Tai, "Detecting Arabic Misinformation Using an Attention Mechanism-Based Model," *Iraqi Journal for Computer Science and Mathematics*, vol. 5, no. 1, pp. 285–298, 2024, doi: 10.52866/ijcs.2024.05.01.020.
- [24] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, no. 4, pp. 1–22, Dec. 2009, doi: 10.1145/1644879.1644881.
- [25] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOVID19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection," in *Proc. Sixth Arabic Natural Language Processing Workshop (WANLP)*, Kyiv, Ukraine (Virtual), Apr. 2021, pp. 72–81.
- [26] A. R. Mahlous and A. Al-Laith, "Fake News Detection in Arabic Tweets during the COVID-19 Pandemic," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 776–785, 2021, doi: 10.14569/IJACSA.2021.0120691.
- [27] M. S. Hadj Ameur and H. Aliane, "AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset," *Procedia Comput. Sci.*, vol. 189, pp. 232–241, 2021, doi: 10.1016/j.procs.2021.05.086.

- [28] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.
- [29] H. F. Al-Saif and H. Z. Al-Dossari, "Exploring the role of emotions in Arabic rumor detection in social media," *Applied Sciences*, vol. 13, no. 15, Art. no. 8815, 2023, doi: 10.3390/app13158815.
- [30] G. Amoudi, R. Albalawi, F. Baothman, A. Jamal, H. Alghamdi, and A. Alhothali, "Arabic rumor detection: A comparative study," *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 12511–12523, 2022, doi: 10.1016/j.aej.2022.05.029.
- [31] F. L. Alotaibi and M. M. Alhammad, "Using a rule-based model to detect Arabic fake news propagation during Covid-19," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130114.
- [32] N. O. Bahurmuz, G. A. Amoudi, F. A. Baothman, A. T. Jamal, H. S. Alghamdi, and A. M. Alhothali, "Arabic rumor detection using contextual deep bidirectional language modeling," *IEEE Access*, vol. 10, pp. 114907–114918, 2022, doi: 10.1109/ACCESS.2022.3217522.
- [33] A. Bahaaluddin et al., "AraDetector at AraIEval shared task: An ensemble of Arabic-specific pre-trained BERT and GPT-4 for Arabic disinformation detection," in *Proc. First Arabic Natural Language Processing Conf. (ArabicNLP 2023)*, Singapore, 2023, pp. 530–535, doi: 10.18653/v1/2023.arabicnlp-1.51.
- [34] F. Haouari, T. Elsayed, and R. Suwaileh, "AuRED: Enabling Arabic rumor verification using evidence from authorities over Twitter," in *Proc. Second Arabic Natural Language Processing Conf. (ArabicNLP 2024)*, Bangkok, Thailand, 2024, pp. 27–41, doi: 10.18653/v1/2024.arabicnlp-1.3.
- [35] R. S. Aziz, A. T. Sadiq, M. Kherallah, and A. Douik, "Arabic fake news detection for Covid-19 using deep learning and machine learning," *Periodicals of Engineering and Natural Sciences*, vol. 11, no. 6, pp. 56–72, 2023, doi: 10.21533/pen.v11i6.194.
- [36] A. Abdedaïem, A. H. Dahou, and M. A. Cheragui, "Fake news detection in low resource languages using SetFit framework," *Inteligencia Artificial*, vol. 26, no. 72, pp. 178–201, 2023, doi: 10.4114/intartif.vol26iss72pp178-201.
- [37] A. O. Alharbi et al., "Evaluating large language models on health-related claims across Arabic dialects," in *Proc. 1st Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2025)*, 2025, pp. 95–103.
- [38] S. Alyoubi, M. Kalkatawi, and F. Abukhodair, "The Detection of Fake News in Arabic Tweets Using Deep Learning," *Applied Sciences*, vol. 13, no. 14, Art. no. 8209, Jul. 2023, doi: 10.3390/app13148209.
- [39] H. Saadany, C. Orasan, and E. Mohamed, "Fake or Real? A Study of Arabic Satirical Fake News," in *Proc. 3rd Int. Workshop on Rumours and Deception in Social Media (RDSM)*, Barcelona, Spain (Online), Dec. 2020, pp. 70–80.
- [40] N. M. Alkudah, N. B. Idris, A. Q. M. Sabri, and M. A. M. Abushariah, "An empirical comparison of ensemble and deep learning models for multi-level Arabic fake news detection using the JoNewsFake dataset," *PeerJ Comput. Sci.*, vol. 12, Art. no. e3510, 2026, doi: 10.7717/peerj-cs.3510.
- [41] A. Altamimi, "Novel approach for predicting fake news stance detection using large word embedding blending and customized CNN model," *PLOS ONE*, vol. 19, no. 12, Art. no. e0314174, 2024, doi: 10.1371/journal.pone.0314174.
- [42] H. Najadat, M. Tawalbeh, and R. Awawdeh, "Fake news detection for Arabic headlines-articles news data using deep learning," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 4, pp. 3951–3959, 2022, doi: 10.11591/ijece.v12i4.pp3951-3959.
- [43] F. J. Fernández-Bravo Peña, "Deception Detection in Arabic Tweets and News," in *Working Notes of FIRE 2019 – Forum for Information Retrieval Evaluation*, CEUR Workshop Proceedings, vol. 2517, Kolkata, India, 2019, pp. 122–126.
- [44] M. Abd Elaziz, A. Dahou, D. A. Orabi, S. Alshathri, E. M. Soliman, and A. A. Ewees, "A Hybrid Multitask Learning Framework with a Fire Hawk Optimizer for Arabic Fake News Detection," *Mathematics*, vol. 11, no. 2, Art. no. 258, Jan. 2023, doi: 10.3390/math11020258.
- [45] M. Alkhair, K. Meftouh, K. Smaïli, and N. Othman, "An Arabic Corpus of Fake News: Collection, Analysis and Classification," in *Arabic Language Processing: From Theory to Practice (ICALP)*, Communications in Computer and Information Science, Springer, 2019, pp. 292–302, doi: 10.1007/978-3-030-32959-4_21.
- [46] N. M. Alkudah, N. B. Idris, M. A. M. Abushariah, A. Q. M. Sabri, and M. Alkhaili, "FakeNews-TransAug: an enhanced AraBERT-Based deep learning model with data augmentation for addressing class imbalance in Arabic fake news detection," *PeerJ Comput. Sci.*, vol. 12, Art. no. e3654, 2026, doi: 10.7717/peerj-cs.3654.
- [47] A. Y. Muad et al., "Arabic Document Classification: Performance Investigation of Preprocessing and Representation Techniques," *Math. Probl. Eng.*, vol. 2022, Art. no. 3720358, 2022, doi: 10.1155/2022/3720358.
- [48] S. O. Althabiti, M. A. Alsalka, and E. Atwell, "Detecting Arabic Fake News on Social Media using Sarcasm and Hate Speech in Comments," *International Journal on Islamic Applications in Computer Science and Technology-IJASAT*, vol. 10, no. 4, pp. 28–36, 2022, [Online]. Available: <https://eprints.whiterose.ac.uk/>
- [49] H. Ala'a Alkaabi, A. Kadhim Jasim, and A. Darroudi, "Arabic NLP: A Survey of Pre-Processing and Representation Techniques," *Journal of Computer Science, Information Technology and Telecommunication Engineering (JCoSITE)*, vol. 6, no. 2, pp. 876–890, 2025, doi: 10.30596/jcositte.v6i2.25562.
- [50] S. Larabi Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali, and I. Abunadi, "Arabic natural language processing and machine learning-based systems," *IEEE Access*, vol. 7, pp. 7011–7020, 2019, doi: 10.1109/ACCESS.2018.2890076.
- [51] N. M. Alkudah, N. B. Idris, and M. A. M. Abushariah, "Fake news detection in Arabic media: comparative analysis of machine learning and deep learning algorithms using the Arabic Fake News Dataset," *PeerJ Comput. Sci.*, vol. 11, Art. no. e3272, 2025, doi: 10.7717/peerj-cs.3272.
- [52] T. Thaher, M. Saheb, H. Turabieh, and H. Chantar, "Intelligent detection of false information in Arabic tweets utilizing hybrid Harris Hawks-based feature selection and machine learning models," *Symmetry*, vol. 13, no. 4, Art. no. 556, Apr. 2021, doi: 10.3390/sym13040556.
- [53] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.
- [54] R. Al Mamlook et al., "Optimizing Machine Learning Models based on Hyperparameter Tuning for Arabic Fake News Detection," in *ISAS 2025 - 9th International Symposium on Innovative Approaches in Smart Technologies, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2025, doi: 10.1109/ISAS66241.2025.11101917.
- [55] E. M. B. Nagoudi, A. Elmadany, M. Abdul-Mageed, T. Alhindi, and H. Cavusoglu, "Machine Generation and Detection of Arabic Manipulated and Fake News," in *Proc. Fifth Arabic Natural Language Processing Workshop (WANLP)*, Barcelona, Spain (Online), Dec. 2020, pp. 69–84.
- [56] A. Chatta, M. Bouakkaz, and B. Latreche, "Sentiment Analysis-Driven Approach for Enhancing Arabic Fake News Detection Using Machine and Deep Learning," *International Journal of Informatics and Applied Mathematics*, vol. 8, no. 2, pp. 19–36, Jan. 2026, doi: 10.53508/ijiam.1838771.
- [57] K. M. Fouad, S. F. Sabbeh, and W. Medhat, "Arabic fake news detection using deep learning," *Computers, Materials and Continua*, vol. 71, no. 2, pp. 3647–3665, 2022, doi: 10.32604/cmc.2022.021449.
- [58] E. T. Khudair, O. Lazzez, M. Zaied, T. M. Hamdani, A. T. Sadiq, H. Chabchoub, and A. M. Alimi, "Deep Learning Techniques To Detect Arabic Fake News," *Int. J. Adv. Signal Image Sci.*, vol. 12, no. 1s, pp. 760–772, 2026, doi: 10.29284/c5dxbw85.
- [59] A. A. Elrowayati and M. M. Shumani, "Arabic Fake News Detection Using Genetically Optimized Deep Learning Models," *Journal of Technology Research (JTR)*, vol. 4, special issue, pp. 865–879, 2026, doi: 10.26629/jtr.2025.82.
- [60] I. kh. Alnabrisi and M. Saad, "Detect Arabic fake news through deep learning models and Transformers," *Expert Syst. Appl.*, vol. 251, Art. no. 123997, Oct. 2024, doi: 10.1016/j.eswa.2024.123997.
- [61] T. A. Wotaifi and B. N. Dhannoon, "An Effective Hybrid Deep Neural Network for Arabic Fake News Detection," *Baghdad Science Journal*, vol. 20, no. 4, pp. 1392–1401, 2023, doi: 10.21123/bsj.2023.7427.

- [62] W. Shishah, "JointBert for Detecting Arabic Fake News," *IEEE Access*, vol. 10, pp. 71951–71960, 2022, doi: 10.1109/ACCESS.2022.3185083.
- [63] N. A. Othman, D. S. Elzanfaly, and M. M. M. Elhawary, "Arabic Fake News Detection Using Deep Learning," *IEEE Access*, vol. 12, pp. 122363–122376, 2024, doi: 10.1109/ACCESS.2024.3451128.
- [64] N. M. Alkudah, S. Abufakher, R. Masadeh, E. Masadeh, and N. Bataina, "A Context Aware Deep Learning Framework for Multi-Level Arabic Fake News Detection Using Hybrid Linguistic Representations," *Journal of Computer Science*, vol. 22, no. 3, pp. 1127–1138, Mar. 2026, doi: 10.3844/jcsp.2026.1127.1138.
- [65] A. B. Nassif, A. Elnagar, O. Elgendy, and Y. Afadar, "Arabic fake news detection based on deep contextualized embedding models," *Neural Comput. Appl.*, vol. 34, no. 18, pp. 16019–16032, 2022, doi: 10.1007/s00521-022-07206-4.
- [66] B. M. Merzah, J. Razmara, and Z. Salmanian, "Hybrid deep learning models for fake news detection: case study on Arabic and English languages," *Front. Big Data*, vol. 8, Art. no. 1683786, Jan. 2026, doi: 10.3389/fdata.2025.1683786.
- [67] H. J. Alshahrani *et al.*, "Hunter Prey Optimization with Hybrid Deep Learning for Fake News Detection on Arabic Corpus," *Computers, Materials and Continua*, vol. 75, no. 2, pp. 4255–4272, 2023, doi: 10.32604/cmc.2023.034821.
- [68] R. S. Matti and S. A. Yousif, "Leveraging Arabic BERT for High-Accuracy Fake News Detection," *Iraqi Journal of Science*, vol. 66, no. 2, pp. 751–764, 2025, doi: 10.24996/ijis.2025.66.2.18.
- [69] F. Harrag, M. Dabbah, K. Darwish, and A. Abdelali, "BERT Transformer model for Detecting Arabic GPT -2 Auto-Generated Tweets," in *Proc. Fifth Arabic Natural Language Processing Workshop (WANLP)*, Barcelona, Spain (Online), Dec. 2020, pp. 207–214.
- [70] G. Jardaneh, H. Abdelhaq, M. Buzz, and D. Johnson, "Classifying Arabic Tweets Based on Credibility Using Content and User Features," in *Proc. 2019 IEEE Jordan Int. Joint Conf. Electrical Engineering and Information Technology (JEEIT)*, 2019, pp. 596–601, doi: 10.1109/JEEIT.2019.8717386.
- [71] D. H. Abd, M. F. Mahdi, M. A. Jassim, and A. Hussain, "Arabic fake news detection using ensemble technique," in *Proc. 16th Int. Conf. Developments in eSystems Engineering (DeSE)*, Istanbul, Turkey, 2023, pp. 292–297, doi: 10.1109/DeSE60595.2023.10469046.
- [72] M. Al-Yahya, H. Al-Khalifa, H. Al-Baity, D. Alsaeed, and A. Essam, "Arabic Fake News Detection: Comparative Study of Neural Networks and Transformer-Based Approaches," *Complexity*, vol. 2021, Art. no. 5516945, 2021, doi: 10.1155/2021/5516945.
- [73] M. E. Almandouh, M. F. Alrahmawy, M. Eisa, M. Elhoseny, and A. S. Tolba, "Ensemble-based high-performance deep learning models for fake news detection," *Sci. Rep.*, vol. 14, Art. no. 26591, 2024, doi: 10.1038/s41598-024-76286-0.
- [74] B. M. Merzah, J. Razmara, and J. Karimpour, "Self-Attention Enhanced Dual BiGRU for Arabic Fake News Detection," *Mesopotamian Journal of Computer Science*, vol. 2025, pp. 247–257, Sep. 2025, doi: 10.58496/MJCSC/2025/016.
- [75] N. M. Alkudah, "Arabic Fake News Detection Across Generational Text Representations: From Traditional Models to Transformer-Based Methodologies," *Journal of Computer Science*, vol. 22, no. 4, pp. 1313–1329, Apr. 2026, doi: 10.3844/jcsp.2026.1313.1329.
- [76] A. Alrayzah, S. M. Alqhtani, and B. Alrashidi, "AraBLA: cross-layer enhanced AraBERT with BiLSTM and attention for Arabic fake news detection," *Sci. Rep.*, vol. 16, 2026, doi: 10.1038/s41598-026-49960-8.
- [77] M. A. Mostafa and A. Almogren, "VERA-ARAB: unveiling the Arabic tweets credibility by constructing balanced news dataset for veracity analysis," *PeerJ Comput. Sci.*, vol. 10, Art. no. e2432, 2024, doi: 10.7717/peerj-cs.2432.
- [78] B. Taboubi, M. A. Ben Nessim, and H. Haddad, "iCompass at CheckThat! 2022: ARBERT and AraBERT for Arabic Checkworthy Tweet Identification," in *Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Bologna, Italy, 2022.
- [79] K. Hamadouche, K. Z. Bousmaha, M. Y. B. Amar, and L. Hadrich-Belguith, "Detection of Arabic and Algerian Fake News," *Applied Computer Systems*, vol. 29, no. 2, pp. 14–21, Dec. 2024, doi: 10.2478/acss-2024-0017.
- [80] E. Papageorgiou, C. Chronis, I. Varlamis, and Y. Himeur, "A Survey on the Use of Large Language Models (LLMs) in Fake News," *Future Internet*, vol. 16, no. 8, Art. no. 298, Aug. 2024, doi: 10.3390/fi16080298.
- [81] H. Himdi, N. Zamzami, F. Najjar, M. Alrehaili, and N. Bouguila, "Arabic Fake News Dataset Development: Humans and AI-Generated Contributions," *IEEE Access*, vol. 13, pp. 62234–62253, 2025, doi: 10.1109/ACCESS.2025.3556376.
- [82] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023, doi: 10.48550/arXiv.2302.13971.
- [83] E. Almazrouei *et al.*, "The Falcon series of open language models," *arXiv preprint arXiv:2311.16867*, 2023, doi: 10.48550/arXiv.2311.16867.
- [84] N. Sengupta *et al.*, "Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models," *arXiv preprint arXiv:2308.16149*, 2023, doi: 10.48550/arXiv.2308.16149.
- [85] L. Al-Zahrani and M. Al-Yahya, "Pre-Trained Language Model Ensemble for Arabic Fake News Detection," *Mathematics*, vol. 12, no. 18, Art. no. 2941, Sep. 2024, doi: 10.3390/math12182941.
- [86] H. M. Turki *et al.*, "Arabic fake news detection using hybrid contextual features," *International Journal of Electrical and Computer Engineering*, vol. 15, no. 1, pp. 836–845, Feb. 2025, doi: 10.11591/ijece.v15i1.pp836-845.
- [87] N. Zamzami, H. Himdi, and R. K. Qarout, "An Automated Unsupervised Model Using Probabilistic Mixture Models and Textual Analysis for Arabic Fake News Detection," *Mathematics*, vol. 14, no. 8, Art. no. 1250, Apr. 2026, doi: 10.3390/math14081250.
- [88] V. U. Gongane, M. V. Munot, and A. D. Anuse, "A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms," *J. Comput. Soc. Sci.*, vol. 7, no. 1, pp. 587–623, Apr. 2024, doi: 10.1007/s42001-024-00248-9.
- [89] G. George and T. Abdullahi, "A SHAP-Based XAI Approach to Evaluating Machine Learning Classification of Fake News," *Nile Journal of Engineering and Applied Science*, vol. 2, no. 2, p. 1, 2025, doi: 10.5455/njeas.214016.
- [90] T. Aljrees, "Improving Prediction of Arabic Fake News Using ELMO's Features-Based Tri-Ensemble Model and LIME XAI," *IEEE Access*, vol. 12, pp. 63066–63076, 2024, doi: 10.1109/ACCESS.2024.3392297.
- [91] M. Saad, S. Abdelrazek, and I. R. Abdelmaksoud, "An algorithmic system for arabic fake news detection using neural networks and transformer embeddings with class weighting," *Sci. Rep.*, vol. 16, 2026, doi: 10.1038/s41598-026-45653-4.
- [92] F. K. Abu Salem, R. Al Feel, S. Elbassuoni, M. Jaber, and M. Farah, "FA-KES: A Fake News Dataset around the Syrian War," in *Proc. Int. AAAI Conf. Web and Social Media (ICWSM)*, vol. 13, no. 1, 2019, pp. 573–582, doi: 10.1609/icwsml.v13i01.3254.
- [93] B. Hawashin, A. Althunibat, T. Kanan, S. Alzu'Bi, and Y. Sharrah, "Improving Arabic Fake News Detection Using Optimized Feature Selection," in *2023 International Conference on Information Technology: Cybersecurity Challenges for Sustainable Cities, ICIT 2023 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 690–694, doi: 10.1109/ICIT58056.2023.10225974.
- [94] H. Himdi, G. Weir, F. Assiri, and H. Al-Barhamtoshy, "Arabic Fake News Detection Based on Textual Analysis," *Arab. J. Sci. Eng.*, vol. 47, no. 8, pp. 10453–10469, Aug. 2022, doi: 10.1007/s13369-021-06449-y.
- [95] A. Hocini and K. Smaili, "Detecting Fake News: Exploring Key Features in Multilingual Arabic Dialect Corpus," in *Arabic Language Processing: From Theory to Practice (ICALP)*, Communications in Computer and Information Science, Springer, 2025, pp. 236–248, doi: 10.1007/978-3-031-80438-0_18.
- [96] A. Khalil, M. Jarrah, M. Aldwairi, and M. Jaradat, "AFND: Arabic fake news dataset for the detection and classification of articles credibility," *Data in Brief*, vol. 42, Art. no. 108141, 2022, doi: 10.1016/j.dib.2022.108141.
- [97] J. Khouja, "Stance prediction and claim verification: An Arabic perspective," in *Proc. Third Workshop on Fact Extraction and*

- VERification (FEVER), Online, Jul. 2020, pp. 8–17, doi: 10.18653/v1/2020.fever-1.2.
- [98] M. K. Elhadad, K. F. Li, and F. Gebali, “COVID-19-FAKES: A Twitter (Arabic/English) dataset for detecting misleading information on COVID-19,” in *Advances in Intelligent Networking and Collaborative Systems (INCoS 2020)*, Advances in Intelligent Systems and Computing, vol. 1263. Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-57796-4_25.
- [99] A. M. K. Shehata, M. N. Al-Suqri, N. E. M. E. Osman, F. Hamad, Y. N. Alhusaini, and A. Mahfouz, “ArabFake: A Multitask Deep Learning Framework for Arabic Fake News Detection, Categorization, and Risk Prediction,” *IEEE Access*, vol. 12, pp. 191345–191360, 2024, doi: 10.1109/ACCESS.2024.3518204.
- [100] M. Umer, A. A. Jamjoom, S. Alsubai, A. A. AlArfaj, E. A. Alabdulqader, and I. Ashraf, “Predictive Modeling for Arabic Fake News Detection: Leveraging Language Model Embeddings and Stacked Ensemble,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, early access, Jul. 2024, doi: 10.1145/3677016.
- [101] A. Qusef, H. Mosa, and A. El-Hassan, “Improving Fake News Detection in Arabic Using ML and DL Algorithms,” in *2025 14th International Conference on Software and Computer Applications, ICSCA 2025*, Association for Computing Machinery, Inc, Sep. 2025, pp. 279–283, doi: 10.1145/3731806.3731828.
- [102] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci. Rep.*, vol. 14, Art. no. 6086, 2024, doi: 10.1038/s41598-024-56706-x.
- [103] M. Huisman, “When AUC-ROC and accuracy are not accurate: what everyone needs to know about evaluating artificial intelligence in radiology,” *Eur. Radiol.*, vol. 34, no. 12, pp. 7892–7894, Dec. 2024, doi: 10.1007/s00330-024-10859-5.
- [104] Z. Sheikh Ali, A. Al-Ali, and T. Elsayed, “Detecting Users Prone to Spread Fake News on Arabic Twitter,” in *Proc. 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, Marseille, France, Jun. 2022, pp. 12–22.