Journal of
Information Technology,
Cybersecurity, and Artificial Intelligence

(ARTICLE TYPE)

Arabic Dialect NLP: A Unified Taxonomy of Tasks, Methodological Evolution, and Dialectal Trends

Dr. Hani Iwidat
Assistant. Professor; Al-Istiqlal University,

Dr. Mamoun Abu Helou
Associate Professor; Al-Istiqlal University.

Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2026, 3(2), 1–12

Article DOI: https://doi.org/10.70715/jitcai.2026.v3.i2.051

Abstract

Background: Natural language processing of Arabic dialects has faced significant challenges due to the language's complex nature, while Modern Standard Arabic enjoys widespread support in this field. The rapid spread of social media has also shifted research focus towards Arabic dialects, thus creating a critical need to criticize this massive body of research. **Objective**: This study aims to provide a survey and application-oriented review of the Arabic Dialectal NLP landscape. The primary goal is to map the relationship between foundational tasks, while benchmarking the resources and methodologies that have defined the field. **Participants and Setting**: The study analyzes a comprehensive dataset of 400 research articles published between 2020 and 2025. **Methods**: A survey was conducted, utilizing a multi-taxonomic clustering approach. Research papers were categorized into eight functional clusters. Trends were analyzed by year, geographic focus, and algorithm type (Traditional Machine Learning vs. Deep Learning vs. Transformers and LLMs). **Results:** The study analysis reveals that Sentiment Analysis category is the dominant application about 32% of the literature, followed by 21% for resource building group. Identification and Code-Switching is 10%. Research output peaked in 2022-2025, marking a definitive shift from traditional machine learning models to Transformer-based architectures like AraBERT and MARBERT. Regional coverage is broad, with a notable trend toward the identification and handling of code-switched text, which has emerged as the current state-of-the-art. **Conclusions**: The survey demonstrates that dialect identification is no longer a standalone goal but a prerequisite for sentiment and translation systems. The field has progressed notably in many areas, such as SA, but future work must prioritize under-resourced dialects, reproducible benchmarks, and cross-dialect transfer learning, and bond these specific dialectal models with the zero-shot capabilities of generative LLMs.

Keywords: Arabic Dialect, Taxonomy, Natural Language Processing (NLP), Dialect applications

1. INTRODUCTION

In the area of Artificial Intelligence (AI) computational linguistics (CL), the landscape has evolved through significant technological application, resulting in the development of high-performance systems. These innovations have enabled a wide array of specialized applications, such as speech translation (ST), speech recognition (SR), dialect identification (DI) and sentiment annotation (SA). Furthermore, the transition toward advanced architectures like Transformers and Large Language Models (LLMs) has redefined the efficiency of tasks such as author profiling and plagiarism detection, providing researchers and developers with robust tools to handle the nuances of modern digital challenges [1].

This systematic review seeks to monitor and analyze the research landscape on Arabic dialects between 2020 and 2025, a period that witnessed a peak in scholarly output with nearly 400 research papers. The study aims to map the relationship between core tasks, resource development, and methodologies in this field. Despite the strategic importance of Arabic, spoken by over 400 million people and prevalent in more than twenty countries, the treatment of its dialects still suffers from a scarcity of standardized resources compared to Modern Standard Arabic[2, 3]. The objective of this work is to critique and analyze this vast body of research, classifying it within clear functional frameworks to overcome the problem of fragmented research efforts and provide a comprehensive reference that supports researchers in developing models capable of effectively and competently accommodating the diversity of the Arabic language.

Since 2022, Arabic dialect NLP research has surged dramatically, with over 60% of the 400 screened papers published in the transformer and early LLM era. Earlier surveys [2, 4-6], were task-specific or pre-dated AraBERT, MARBERT and LLMs, leaving fragmented taxonomies that ignore task interdependencies (e.g., dialect identification is now a prerequisite for sentiment and translation). Meanwhile, code-switching, Arabizi and city-level variation have exploded, while Sudanese, Yemeni and Iraqi dialects remain almost invisible. A unified taxonomy is therefore essential now to map these relationships, quantify the shift to LLMs, expose imbalances and guide cross-dialect transfer learning.

Despite the valuable contributions of earlier surveys [4-6], they remain limited in scope: most cover fewer than 150 papers, focus on a single task (e.g., sentiment or identification only), or stop before the transformer/LLM era. Consequently, the field still lacks a unified, application-oriented taxonomy that maps task interdependencies, quantifies methodological shifts across 2020-2025, and systematically reveals dialectal imbalances. The present work fills this gap by screening 400 studies, analyzing 101 high-quality papers, and introducing an eight-cluster taxonomy that explicitly links tasks, resources, models, and dialects.

This survey addresses the following central research question: RQ: What are the key interrelationships between foundational tasks, resources, and methodologies in Arabic dialect NLP research from 2020–2025, and how do trends in model evolution and dialect coverage reveal gaps for future development? The unified taxonomy, trend analysis, and gap mapping presented in Sections 4-6 directly answer this question.

Arabic is characterized as a bilingual language, divided into Modern Standard Arabic, the formal language, and colloquial dialects, which serve as the means of daily communication. This colloquial language is distributed across a wide geographical area, encompassing six main regions: the Levant, the Arabian Gulf, the Nile Basin, and North Africa, in addition to distinct categories such as the Iraqi and Yemeni dialects. This diversity is characterized by a linguistic gradation ranging from dialects very close to Classical Arabic to those structurally distant from it [2, 7]. Even within a single country, subtle variations exist, reaching the level of urban dialects. Understanding this geographical and linguistic division is fundamental to building any language processing system, as morphological rules and vocabulary differ from one region to another, necessitating precise classification to help construct specialized models for each language group.

The emergence of Web 3.0 technologies and the proliferation of user-generated content on social media platforms have led to a massive influx of unstructured data. Within this massive flow, the opinions and information expressed by users in their posts acquire intrinsic value for strategic and social analysis. However, processing this content creates qualitative challenges that make knowledge extraction a highly complex task. This is due to the lack of standardized spelling rules

for dialects and the widespread phenomenon of code-switching. This is clearly evident in the mixing of Arabic and English, or Modern Standard Arabic and colloquial Arabic, and even between different Arabic dialects within a single sentence [8]. This study sheds light on current research, providing Arab researchers with a comprehensive overview of emerging trends (2020-2025) in written and spoken content, with the aim of fostering innovation in data processing research and models specific to the Arabic context.

1.1. Research Goals and Contributions

1. **A Unified, Multi-Dimensional Taxonomy for Arabic Dialect NLP.** The survey introduces a comprehensive taxonomy that organizes AD-NLP research across **tasks, dialect type, model types, and key contributions**, providing a structured and integrated view of the field.
2. **A Systematic, Evidence-Based Mapping of 2020–2025 Research Trends.** The analysis synthesizes 400 studies to reveal clear trends, including the shift from ML, and Transformers, to LLMs, evolving dataset resources, the rise of fine-grained dialect identification, and the growth of multimodal and code-switching research, supported by both quantitative and qualitative evidence.

3. **Identification of Critical Gaps and Future Directions in Under-Resourced Dialects.** The survey highlights major gaps such as scarce datasets for Maghrebi, Sudanese, and city-level dialects, limited reproducible benchmarks, weak cross-dialect transfer, and insufficient LLM adaptation; offering concrete recommendations for future AD-NLP development.

The paper is structured as follows: Section 2 reviews related work. Then section 3 outlines the methodology. Section 4 presents the proposed taxonomy and analysis of trends. Section 5 synthesizes the survey findings. Section 6 discusses the findings and limitations of the study, and Section 7 concludes with future directions.

2. Related Works

Over the past decade, Arabic dialect NLP (AD-NLP) has been surveyed through successive efforts that collectively chart the field's maturation, spanning foundational taxonomies, application-focused reviews, and methodology-centered syntheses. This review organizes the historical trajectory along three complementary axes, applications, approaches, and domains: covering work from 2020 to 2025.

2.1. Consolidation and Taxonomies Surveys

An early effort, presented in [9], established a four-part taxonomy comprising Basic Language Analyses (morphology, POS, syntax, orthography), Building Resources (lexicons, corpora, treebanks), Language Identification (text Identification and speech recognition), and Semantic-level Analysis (machine translation, sentiment).Their review also noted the early dominance of Egyptian Arabic in research coverage and evaluation. Later, [2] adopted and extended this taxonomy across 90 papers; 74% published post-2015, confirming the centrality of sentiment analysis and machine translation within semantic-level work.

2.2. Application-Focused Surveys

Dialect Identification (DI) received sustained attention. [10] surveyed DI methods from traditional machine learning to deep learning, while [11] consolidated benchmarks and error profiles. [12] synthesized dialectal Sentiment Analysis (SA) by cataloging datasets, contrasting feature engineering and deep encoders, and highlighting evaluation gaps unique to dialectal phenomena.

[13] comprehensively reviewed Arabic text summarization, outlining extractive/abstractive techniques and adaptation challenges for morphologically rich and dialectal inputs. [14] offered a systematic review of Arabic Speech recognition ASR covering works in (2011–2021), discussing acoustic modeling, language models, dialectal coverage, and resource gaps in spontaneous speech. [3] enumerated freely available Arabic Corpora & Resources, lowering entry barriers while exposing gaps in dialect coverage and licensing critical for reproducibility.

Work on Misinformation & Safety have been also considered, [15] synthesized methods and resources for detecting Arabic fake news and spam, highlighting complications introduced by dialectal code-mixing and domain transfer. [7] traced Arabic dialect Machine Translation (MT) from rule-based and statistical approaches to neural systems, underscoring data scarcity and orthographic variability as persistent hurdles.

Regional specialization emerged through dialect-focused reviews such as Darija ASR [16] and Algerian sentiment analysis [17]. [18] systematized Opinion Mining for Arabic dialects on social media, a high-noise domain central to sentiment and stance tasks. [19] surveyed Authorship Identification in Arabic texts, analyzing how document length and register (standard vs. dialect) influence attribution performance. [8] provided a comprehensive survey of code-switched Arabic NLP spanning language identification, sentiment analysis, POS tagging, and ASR, and emphasized challenges in data scarcity and evaluation consistency.

2.3. Approach-Centered Surveys

[20] documented the early Deep Learning (DL) wave in Arabic NLP, capturing the transition away from feature-engineered pipelines to neural sequence modeling before transformers became dominant. [1] surveyed pretrained Transformers and LLMs across 34 papers, covering encoder-only (BERT, ELECTRA, RoBERTa), decoder-only (GPT), and encoder–decoder (T5) architectures applied to sentiment, NER, QA, and machine translation, with attention to openness and evaluation practices.

Unlike earlier survey works that focused on specific dialects, tasks, dataset, or methodological aspects, our study provides a comprehensive, application-oriented taxonomy covering new functional task clusters. Furthermore, this survey is the first

to conduct a large-scale longitudinal analysis (2020 - 2025) over approximately 101 studies, capturing the paradigm shift from traditional machine learning to transformer-based architectures and emerging LLM-driven approaches.
To this extent the fundamental differences between our survey and the related work reviews on Arabic Dialect NLP:

1. Provides an up-to-date, comprehensive survey covering Arabic dialect NLP research from 2015 to 2025, synthesizing developments across application and methodologies.
2. Proposes a refined taxonomy that integrates applications and methodological advances, addressing limitations in previous classification schemes.
3. Offers a unified landscape that bridges application and approach-centered perspectives, enabling a more coherent understanding of how tasks, dialectal resources, and modeling trends interact.

Table 1 presents a comparative analysis of 9 related reviews; this survey reviews major works in the field of Arabic Dialect NLP (ADNLP) published between 2020 and 2025. The comparison highlights that while previous works often focused on specific sub-domains (such as Sentiment Analysis or Dialect Identification), This study provides a more comprehensive, multi-dimensional taxonomy covering 101 screened papers from a peak period of 2020-2025. Key differentiators include the current study's unique focus on an eight-cluster functional framework (Groups *A-H*), a longitudinal tracking of the paradigm shift from traditional ML and DL to Transformer-based models, and a detailed research gap heatmap across different dialects and tasks. Unlike concurrent works that may exclude Modern Standard Arabic (MSA) or focus exclusively on code-switching, this survey synthesizes the entire dialectal landscape to bridge application-centered and approach-centered perspectives.

3. Methodology

We conducted a systematic multi-stage review following PRISMA guidelines: structured keyword searches across SCOPUS, IEEE Xplore, ScienceDirect, Springer, ACM, and ProQuest using the exact queries listed in Table 2 retrieved candidate studies, which were then screened using a six-point inclusion and quality checklist covering training strategies, architectural novelty, dataset specification, methodological clarity, evaluation metrics, and benchmarking practices. Two annotators independently extracted dialect, task, method/model, and contribution information, resolving disagreements through discussion to ensure reliable annotation and taxonomy construction. This pipeline enabled a unified taxonomy and trend mapping over 400 studies published between 2020 and 2025.

3.1. Scope and Search keywords

Table 2 presents the structured set of search keywords used during the systematic retrieval of studies across major academic databases. It lists the exact query formulations executed in SCOPUS, IEEE Xplore, ScienceDirect, Springer, ACM Library, and ProQuest, ensuring comprehensive coverage of research related to Arabic dialects and dialectal Arabic. It demonstrates that all search strings rely on combinations of the terms “Arabic Dialect” and “Dialectal Arabic”, applied consistently across platforms to maximize recall

Table 1: Comparison with Related Work

Survey Reference	Review Period	Papers	Coverage Scope	Taxonomy/ Framework	NLP Tasks Covered	Model Evolution	Key Differentiators
Guellil et al. (2021) [2]	2015-2018	90	MSA + Dialects + Arabizi (3 varieties)	By Arabic variety and script type	Preprocessing, SA, DI, Corpora, Code-switching	Traditional ML + Early DL (CNN, RNN)	Includes Arabizi (Roma Arabic); tripartite variety analysis
Darwish & Habash (2021) [4]	Historical to 2020	200+	MSA + Dialects (broad Arab World)	By research area (not application-specific)	Morphology, Syntax, NER, POS, SA, MT, QA, Parsing	Traditional to Deep Learning	Historical perspective.
Elnagar et al. (2021) [11]	2000-2020	106	Dialectal Arabic (ID	By dialect identification approaches	Dialect Identification only	Traditional ML to DL	Narrow focus on dialect identification;

			& Detection)		(country/city level)	(CNN, RNN, LSTM)	most/least popular dialects analysis
Farha & Magdy (2021) [6]	2015-2020	~60	MSA + Dialects (SA focus)	By approach (lexicon, ML, DL, transfer learning)	Sentiment Analysis only	Feature engineering to DL to BERT	Empirical replication, standardized evaluation; practices for SA
Al Katat et al. (2024) [21]	2019-2024	78	MSA + Dialects (SA focus)	By embedding + classification methods	Sentiment Analysis only	Deep Learning (LSTM, GRU, Transformers)	Recent SA methods; embedding techniques, class imbalance solution
Dahou et al. (2025) [22]	2014-2024	N/A	Dialectal Arabic only	Task-based clusters + temporal trends	SA, DI, MT, ASR, Corpora, Code-switching.	Traditional ML → DL → Transformer → LLMs	Concurrent work; long scope; code-switch emphasis; similar goals to manuscript
Al-Khalifa et al. (2024) [1]	2018-2024	80+	MSA + Dialects (LLM focus)	By LLM architecture (encoder, decoder, encoder-decoder)	SA, NER, QA, Generation, Summarization	Pre-trained LLMs only (AraBERT, AraGPT, Jais, GPT4)	Exclusive focus on Larg Language Models; benchmarks for Arabic LLM
Hamed & Sabty (2025) [8]	2015-2024	~70	Code-switched Arabic (CS focus)	By code-switching task (DI, SA, NER, MT, ASR)	Code-switching detection, SA, NER, MT, ASR	Traditional ML to Transformers	Exclusive focus on code switching; Arabic-Englis Arabic-French mixing
Alayba (2025) [5]	2015-2024	150+	MSA + Dialects (broad NLP)	By linguistic challenge + technique type	Morphology, POS, NER, SA, MT, Parsing, Diacritization, QA	Traditional to DL to Transformers	Challenges-first approach emerging trends, broad NLP coverage
Our 2026	2020-2025 (peak period)	400 -> 101	Dialectal Arabic only (excludes MSA-only)	8 functional clusters (Groups A-H) + methodological evolution + dialectal gap analysis	SA, Resource Building, DI, Code-switching, MT, Preprocessing, ASR/TTS, Info Retrieval, Author Profiling,	Three-phase evolution (ML → DL → Transformers → LLMs);	Unified taxonomic + methodological + trend driven approach; larges screened dataset (400); quality assesses cross-application synthesis research gap heatmap by dialect × task; challenge taxonomy with solutions

and minimize bias in article selection. The use of unified keyword patterns across databases ensured methodological rigor and reproducibility in gathering candidate studies for inclusion in the survey.

DOI: https://doi.org/10.70715/jitcai.2026.v3.i2.051

TABLE 2. Search keywords

SCOPUS	(TITLE-ABS-KEY (dialectal AND Arabic) OR TITLE-ABS-KEY (Arabic AND dialect))
IEEE Xplore	("Arabic Dialect" OR "Dialectal Arabic")
ScienceDirect	('Arabic Dialect' OR 'Dialectal Arabic')
Springers	("Arabic Dialect" OR "Dialectal Arabic")
ACM library	[All: "dialectal Arabic"] OR [All: "Arabic dialect "]
Google scholar	"Arabic Dialect" AND "Dialectal Arabic"

3.2. Inclusion and Exclusion Criteria

In Table 3 we summarize the inclusion and exclusion criteria used to determine which studies were considered in the review. The inclusion criteria specify that eligible articles must be published between 2000 and 2025, written in English, focused specifically on Arabic dialects, and published in journals or conference proceedings. In contrast, the exclusion criteria identify studies that fall outside the publication range, are not written in English, lack focus on Arabic dialects (e.g., MSA-only studies), or duplicate previously collected work. Collectively, this table defines the boundaries of the dataset and ensures that only relevant, methodologically appropriate studies are included in the analysis.

TABLE 3. Inclusion and Exclusion Criteria

Inclusion Criteria.	Exclusion Criteria.
• The articles were published from 2000 to 2025. • The article focuses on Arabic Dialect • The articles are written in the English language. • The research papers are from journals publications, or conferences.	• Articles published not in English • Articles without Arabic dialect • Articles conducted by MSA • The articles are not in the range of 2000 to 2025. • duplicate articles.

TABLE 4. Quality assessment criteria

Criterion	Question
C1: Training	Does it mention *how* it was trained?
C2: Architecture	Does it claim to *change or create* a model structure?
C3: Dataset	Does it name a *specific* dataset or corpus?
C4: Method	Does it name the *specific* model type?
C5: Evaluation	Does it list *numerical* results or metrics?
C6: Benchmark	Does it *compare* the model to existing baseline models?

DOI: https://doi.org/10.70715/jitcai.2026.v3.i2.051

TABLE 5*:* Shows the NLP Research on Arabic Dialects 2020 - 2025

Study	Dialects type						Key Contribution						Application / Task coverage							
	Gulf	Levantine	Egyptian	Maghrebi	Sudanese	Iraqi	Training strategies	Framework	Dataset	Method/Model	Evaluation/Metric	Benchmark	Group A	Group B	Group C	Group D	Group E	Group F	Group G	Group H
[23]				✓						✓			✓							
[24]				✓						✓			✓							
[25]				✓						✓			✓							
[26]				✓						✓			✓							
[27]				✓						✓									✓	
[28]				✓			✓										✓			
[29]				✓					✓						✓					
[30]	✓									✓										
[31]	✓									✓			✓							
[32]				✓						✓							✓			
[33]	✓								✓						✓					
[34]	✓							✓												
[35]		✓								✓							✓			
[36]			✓						✓						✓					
[37]				✓						✓			✓							
[38]				✓						✓							✓			
[39]						✓			✓						✓					
[40]			✓							✓			✓							
[24]	✓									✓										
[41]		✓		✓						✓			✓							
[42]				✓						✓			✓							
[43]				✓					✓						✓					
[44]	✓						✓						✓							
[45]	✓	✓	✓	✓					✓							✓				
[46]	✓						✓						✓							
[47]				✓			✓												✓	
[48]	✓									✓					✓					
[49]				✓						✓				✓						
[50]		✓						✓											✓	
[51]		✓								✓						✓				
[52]				✓					✓						✓					
[53]				✓						✓					✓					
[53]				✓						✓									✓	
[54]				✓						✓				✓						
[55]				✓						✓							✓			
[56]				✓						✓				✓						
[57]				✓						✓				✓						
[58]				✓						✓			✓							
[59]				✓						✓			✓							
[60]	✓									✓			✓							

[61]				√						√		√							
[62]				√						√		√							
[63]				√					√					√					
[64]				√						√								√	
[65]					√					√		√							
[66]			√				√												√
[67]			√							√		√							
[68]	√		√								√	√							
[69]	√								√					√					
[70]	√									√		√							
[71]		√	√	√						√					√				
[72]		√								√								√	
[73]	√									√							√		
[74]			√						√						√				
[75]				√						√			√						
[76]				√						√					√				
[77]				√					√			√							
[78]				√					√										√
[79]				√					√			√							
[46]				√						√					√				
[80]				√						√									√
[81]	√	√	√							√					√				
[82]	√									√		√							
[83]						√					√	√							
[84]				√						√					√				
[85]				√				√								√			
[86]				√			√												√
[87]	√	√	√							√					√				
[88]			√							√		√							
[89]				√						√		√							
[90]				√						√						√			
[91]				√						√		√							
[92]	√	√							√					√					
[93]				√						√							√		
[94]	√			√					√					√					
[95]	√								√			√							
[96]	√	√								√		√							
[97]				√					√			√							
[98]	√									√		√							
[99]	√						√					√							
[100]				√						√						√			
[101]				√					√					√					
[102]				√						√			√						
[103]				√						√			√						
[104]						√				√							√		
[105]	√									√		√							
[106]			√	√						√						√			
[107]	√	√	√	√		√		√					√						

[108]				✓							✓				✓					
[109]	✓	✓	✓	✓					✓						✓					
[110]	✓	✓	✓							✓						✓				
[111]					✓					✓				✓						
[112]	✓	✓	✓	✓	✓	✓				✓										
[113]	✓	✓	✓								✓					✓				
[114]	✓	✓	✓					✓									✓			
[115]				✓							✓					✓				
[116]	✓	✓	✓								✓						✓			
[117]	✓	✓	✓	✓		✓			✓							✓				
[118]	✓	✓	✓	✓		✓			✓						✓					
[119]	✓								✓				✓							
[120]				✓					✓						✓					

Group A: Sentiment Analysis & Opinion Mining
Group B: Text Preprocessing & Morphological & Syntatic analysis
Group C: Building Resources and Corpora
Group D: Identification, Classification and Code-Switching
Group E: Machine Translation
Group F: Author profiling
Group G: Speech Processing & Synthesis
Group H: Information Retrieval & Extraction

3.3. Quality Assessment Criteria

The selection of papers for inclusion in this study was guided by a systematic six-point criterion designed to identify methodologically robust and comparable research contributions; Table 4 reports these criteria. Each paper was evaluated based on whether it explicitly addressed the following dimensions: (1) specific training strategies such as transfer learning, fine-tuning, or data augmentation; (2) novel or significantly modified neural architectures; (3) use of known or introduction of new datasets; (4) clear description of employed methods or models; (5) reporting of standard evaluation metrics; and (6) comparative benchmarking against existing work. Papers meeting at least four of these six criteria were included, ensuring a balanced representation of technical rigor, reproducibility, and contextual relevance within the Arabic dialect NLP landscape while maintaining practical screening efficiency.

3.4. Validity and Reliability

Two independent annotators (the authors) extracted dialect, task, method, and contribution information for all 101 papers. Inter-annotator agreement reached 92%, Disagreements were resolved through discussion until full consensus. Replicability of the eight-cluster taxonomy is ensured by (1) explicit definitions and examples for every cluster in Section 4 and Table 5, (2) the documented three-round iterative clustering process, and (3) public availability of the complete annotated dataset and classification guidelines from the corresponding author upon reasonable request.

Based on the above-mentioned criteria, the screening process involved three stages depicted in Figure 1. At first, papers' title and abstract are reviewed to assess relevance (approximately 400 results identified); then full-text evaluation for methodological fit and contribution to Arabic Dialect is performed; and final selection by the annotators. This step allowed the assessors to understand the findings of previous research and thus identify the gap in literature. This process resulted in 101 studies, which we used as the foundation for our survey, categories, and gap analysis in the sections that follow. Table 5. Summarize the reviewed literatures.

DOI: https://doi.org/10.70715/jitcai.2026.v3.i2.051

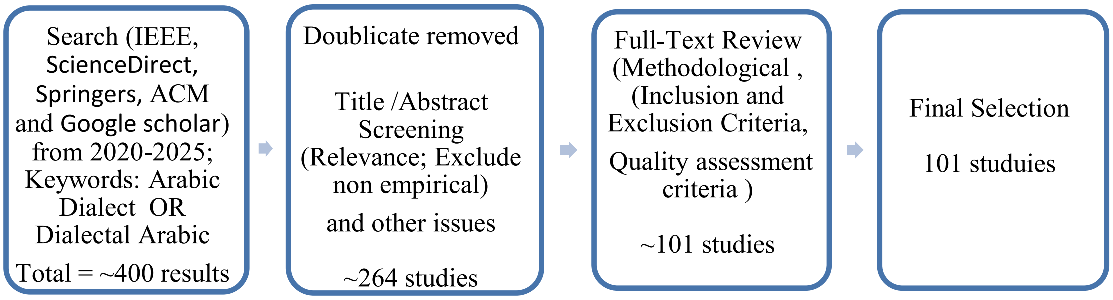

Figure 1. The paper screening workflow

---

4. ARABIC DIAECT APPLICATIONS TAXONOMY

Table 5 provides a comprehensive, structured synthesis of the findings derived from the systematic analysis of the surveyed studies. Its summary of Arabic Dialect NLP research published between 2020 and 2025, organizing over one hundred studies according to dialect type, key contribution, application/task group, modeling approach, and methodological criteria. It captures the distribution of studies across major Arabic dialects—including Gulf, Levantine, Egyptian, Maghrebi, Sudanese, and Iraqi—and aligns each publication with its primary contribution and task category, mapped to the taxonomy groups (A-H) defined in the survey. The table also records methodological dimensions used in the quality assessment—training strategies, frameworks, datasets, model types, evaluation metrics, and benchmarking practices—allowing for cross-comparison of techniques and trends over time. Additionally, the table distinguishes between modeling paradigms—Machine Learning (M), Deep Learning (D), Transformer-based methods (T), and emerging LLM approaches (L)—providing a longitudinal view of the shift from traditional ML to Transformers and LLMs within each application cluster. Through this multi-dimensional structure, It functions as the empirical backbone of the survey, supporting the creation of the unified taxonomy and enabling the analysis of dialect coverage, task prevalence, and methodological evolution across the studied period. These clusters map relationships between foundational tasks and reflect trends such as the shift from traditional Machine Learning (ML) to Deep learning (DL) to transformers then to large language model (LLM). Cluster distribution is visualized in Figure 2.

Each group is presented using a consistent four-part pattern: Definition and Role, Empirical Coverage, Key Contributions and Methodological Patterns, and Gaps and Limitations. This uniform structure enables clear comparison across tasks and directly supports the research question on interdependencies and trends.

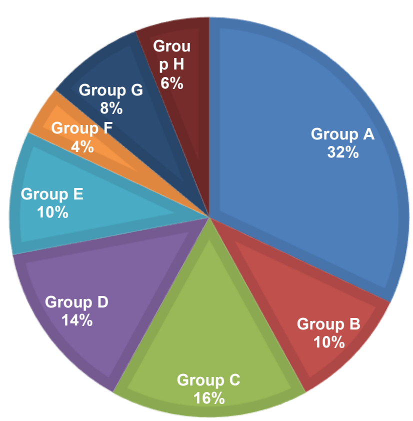

Figure 2. Taxonomy of Applications

4.1. Sentiment Analysis/Opinion Mining (*Group A).*

Sentiment Analysis (SA) also called Opinion Mining (OM) uses NLP and machine learning to identify subjective information in dialectal text/speech, focusing on polarity, sentiments, emotions, attitudes toward entities at document, sentence, aspect levels. Primarily application-facing for social media monitoring and opinion tracking in ADNLP. It often serves as a downstream application reliant on preprocessing (Group B) and dialect identification (Group D) to handle morphological richness and regional slang. Sentiment Analysis (Group A) is the largest cluster, comprising 32% of the studies analyzed conducted between 2020 and 2025, with a peak in 2022-2023; this cluster shows Maghrebi, Egyptian, and Gulf dominance, while Levantine representation was widespread, and Iraqi and Sudanese studies were underexplored. SA research in AD is commonly formulated under several sub-tasks. The most prevalent are SA (behavior analysis), sentiment classification, emotion recognition, OM, aspect SA, sentiment annotation, behavior analysis, opinion question answering, opinion recognition, opinion target recognition, symptom identification, hate speech detection, sarcasm detection, spam detection, cyberbullying detection, misogyny identification, irony detection and infoveillance sentiments. Early works (2020–2021) relied on traditional ML (SVM/RF with TF-IDF), shifting to DL (CNN/BiLSTM) mid-period, and Transformers/LLMs (AraBERT/MARBERT fine-tuning) by 2023–2025. Key challenges include handling code-switching and sarcasm, which reduce accuracy in low-resource dialects. Limitations involve overfitting to social media noise and poor cross-dialect transfer. Ethical concerns arise in bias amplification for underrepresented regions.
These patterns confirm that dialect identification (Group D) has become a prerequisite for reliable sentiment systems, directly supporting the interdependency dimension of RQ.

4.2. Text Preprocessing/Morphological/Syntactic Analysis (*Group B).*

Preprocessing, morphological and syntactic analysis encompass core natural language processing techniques for addressing the complexities of Arabic dialects, including spelling variations, plural morphology, and the lack of standardized forms. These cleaning and normalization steps prepare text or speech data for higher-level use. This group acts as a preprocessing backbone for higher-level tasks, addressing dialectal non-standard orthography and clitic variations to enable accurate downstream processing in sentiment (Group A), translation (Group E), and extraction (Group H). Group B accounts for approximately 10% of the literature. While Maghrebi dialects currently dominate this section, we are witnessing a significant increase in the use of Levantine and Gulf dialects after 2024. However, there is a scarcity of research focusing on Sudanese and Iraqi dialects. Output grew in 2023, often tied to social media text, with emphasis on handling dialect-specific features. Studies in Group B focus on sub-tasks such as morphological analyzers, POS tagging, dialect segmentation, Arabic taggers, stemming, Named Entity Recognition (NER), automatic diacritization, morphological disambiguation, crowdsourcing, and transcriptions. The majority of this group task is contributed to methodology and model approaches. Egyptian, Levantine, Sudanese and Iraqi under-coverage, code-switching, and emoji handling. Limitations: dialect fragmentation, poor syntactic parsing, and absence of standardization; challenges: zero-shot transfer and unified benchmarks.

4.3. Building Resources/Corpora (Group C).

Building Resources and Corpora (Group C) focuses on creating, expanding, and annotating linguistic resources essential for training and evaluating ADNLP models. This includes corpora construction (text/speech datasets), lexicons (dialect-specific vocabularies/sentiment lexicons), crowdsourced annotations, domain-specific ontologies, and transcription efforts. These resources address data scarcity in dialects, enabling tasks like training models for sentiment analysis (Group A), translation (Group E), and speech processing (Group G). Group C comprises approximately 16% of all studies. While progress has been steady over the past five years, this culminated in a surge of efforts after 2022. Maghrebi, Levantine, and Gulf dialects were very common, reflecting the findings of previous research. Notably, Egyptian dialects gained prominence after 2023, especially with the availability of social media data and code-switching. Sudanese and Iraqi dialects, however, suffered from a lack of resources in this area. From Table 5, the studies prior to 2024 stick with ML and DL multi-dialect, with notable effort on LLMs after 2024. For Group *C* studies focused on the subtasks include: Corpus/Dataset Creation or merging, lexicon creation, crowdsourcing corpus annotation, annotation of dataset/corpus, building domain ontologies, and crowdsourcing transcriptions. Contributions emphasize scalable annotation (crowdsourcing for low-resource dialects) and merging existing corpora. Models shift from ML-based annotation tools to DL for semi-automatic labeling, and then after 2023 the Transformers and LLMs for zero-shot annotation and data augmentation. Many studies introduce publicly available datasets

DOI: https://doi.org/10.70715/jitcai.2026.v3.i2.051

to improve reproducibility. Primary gaps include scarce high-quality datasets for Maghrebi sub-variants, Sudanese, Iraqi, and fine-grained city-level dialects; limited parallel/multimodal resources; and annotation inconsistencies due to dialectal spelling/orthographic variation. Limitations involve poor cross-dialect compatibility, lack of standardized licensing/formats, and insufficient focus on speech/transcription for under-resourced areas. Challenges include scaling crowdsourcing reliably and bridging to LLM pre-training needs.
These patterns confirm that resource building and corpus creation (Group C) act as the foundational enablers for every other application cluster, especially in under-resourced dialects, directly supporting the interdependency dimension of RQ.

4.4. Identification, Classification and Code-Switching (*Group D).*

Identification, Classification, and Code-Switching (Group D) represent the diagnostic and categorizing layer of the ADNLP pipeline. This group focuses on determining the linguistic, social, or structural identity of a given text or speech segment. This includes Dialect Identification (DID), distinguishing between regional varieties (e.g., Maghrebi vs. Levantine). and Language Variety Classification (Formal vs. Informal). A critical component is code-switching detection, which addresses the computational challenge of identifying shifts between MSA and dialects or foreign languages. Additionally, it encompasses author profiling (gender, age, and authorship identification). These tasks serve as a vital "routing" mechanism; by identifying the specific dialect or speaker profile first, systems can select the most appropriate downstream models for translation or sentiment analysis. Group D accounts for approximately 10% of the literature analyzed. Research in this cluster has remained a cornerstone of the field. While general dialect identification (country-level) was the primary focus early on, there has been a significant shift toward city-level identification and nuanced code-switching analysis in recent years. Geographically, the research is heavily concentrated on the Maghrebi, with more Gulf, Levantine and Egyptian work after 2023. Conversely, nuanced identification for Yemeni and Sudanese dialects remains underrepresented in this group. Group *D* studies focused on the following: Sub-tasks: Dialect identification, code-switching detection Dialect Classification, gender identification, Language variety classification, Geographic identification, comparative analysis, text preprocessing, dialect Empirical Analysis and analytical studies. Post-2023, the trend has shifted toward DL, Transformer and LLMs, which are utilized for zero-shot dialect identification. Major gaps include the lack of robust models for fine-grained city-level identification and multi-dialectal code-switching limitations. There is also a notable scarcity of speech-based dialect identification. Challenges remain in distinguishing between highly similar dialects and maintaining high accuracy in short, noisy social media posts where linguistic cues are minimal and Cross-References.
These patterns confirm that dialect identification and code-switching detection (Group D) have evolved into critical prerequisites and routing mechanisms for reliable sentiment, translation, and extraction systems, directly supporting the interdependency dimension of RQ.

4.5. Machine Translation (*Group E).*

Machine Translation (Group E) focuses on the automated transformation of text or speech between language varieties. In the ADNLP context, this encompasses dialect-to-MSA translation (normalization), dialect-to-English, and inter-dialectal translation (Maghrebi to Gulf). It includes sub-tasks like dialect transliteration and speech-to-speech translation. Group E acts as a bridge for accessibility, enabling non-speakers to understand dialectal content and allowing MSA-based tools to process regional data via standard forms. Group E represents approximately 10% of the total literature. While lower in volume than sentiment analysis, it is a highly specialized area. Research initially focused on machine translation, dialect translation, speech-to-speech translation, speech translation, dialect transliteration, Transfer learning and spelling error correction. The major work focuses on Maghrebi dialects. Recently, there has been a pivot toward multi-dialectal systems and "noisy" social media translation, with a peak in output during 2023-2025. Methods have evolved from Statistical Machine Translation (SMT) to Neural Machine Translation (NMT). Current trends emphasize Transformer-based models (mBART, MT5) and LLM fine-tuning for few-shot translation. Researchers increasingly utilize synthetic data generation and back-translation to mitigate the scarcity of parallel dialectal data. A major gap is the data scarcity for parallel corpora, lack of standardized orthographies, and dialectal variations in morphology, lexicon, syntax, and code-switching, which hinder model training and generalization across regions like Levantine, Gulf, and Egyptian. Challenges are highlighted in inconsistent evaluation metrics, poor handling of low-resource dialects by LLMs, and challenges like informal orthography and semantic ambiguity reducing translation accuracy.

These patterns confirm that machine translation systems (Group E) heavily depend on accurate dialect identification and rich parallel resources to achieve cross-dialect fidelity, directly supporting the interdependency dimension of RQ.

4.6. Author profiling *(Group F)*

Author Profiling (Group F) focuses on extracting metadata and latent characteristics about an individual based on their linguistic footprint. This cluster includes tasks such as Authorship Identification (identifying a specific writer), Gender and Age Detection, and Authorship Authentication (verifying if a text belongs to a claimed author). In the context of Arabic dialects, this is particularly complex due to the lack of standardized orthography and the high use of Arabizi. In Arabic dialects, this group supports forensic analysis, personalized content delivery, and social trend detection, often leveraging dialect-specific markers like slang or code-switching. It integrates with sentiment (Group A) for nuanced opinion mining and identification (Group D) to disambiguate regional styles. Group F accounts for approximately 4% of the total literature. While it is one of the smaller clusters in terms of volume, its significance has grown alongside the rise of social media-based research. The data suggests that authorship identification is the most frequent task within this group, followed by gender identification. Research coverage is mostly concentrated on high-resource dialects like Maghrebi and Gulf, where large volumes of user-generated content are available, while Sudanese and Yemeni dialects remain unexplored. Sub-tasks in Group F include authorship identification, authorship authentication, influential identification, user identification, and finally, predicting credibility. Methodologically, the cluster relies heavily on stylometric analysis, utilizing features like character n-grams. Early models utilized ML approaches, but recent studies have shifted toward DL and transformer-based embeddings to capture subtle stylistic nuances. A notable trend is the use of multimodal profiling, combining textual dialectal cues with metadata or speech signals to increase accuracy. Author profiling in Arabic dialects faces key challenges like the scarcity of annotated datasets for each dialect variety, non-standard orthography, and dialectal variations complicating trait inference (e.g., age, gender). Limitations involve privacy and ethical concerns surrounding automated profiling without user consent. Technical challenges include the "short-text" problem on social media, where limited linguistic evidence makes it difficult to distinguish between individual styles or demographic traits. Furthermore, the high prevalence of code-switching and sarcasm in dialectal Arabic often confounds stylometric models, leading to decreased performance compared to MSA.

These patterns confirm that author profiling (Group F) integrates closely with sentiment analysis and dialect identification to produce nuanced user insights, directly supporting the interdependency dimension of RQ.

4.7. Speech Processing & Synthesis *(Group G)*

Speech Processing & Synthesis (Group G) focuses on the computational analysis and generation of spoken dialectal Arabic. This cluster includes Automatic Speech Recognition. Which converts dialectal STT and TTS. generating natural-sounding dialectal voices. It also covers Speech Emotion Recognition, Voice Activity Detection (VAD), and Speech Dialect Identification. Essentially, Group G is all about making technology more accessible and hands-free. Since dialects are mostly spoken rather than written, this group builds the bridge between our natural voice and digital systems. Group G accounts for approximately 8% of the literature. Research in this area has seen a significant surge as mobile technology and voice-activated AI become ubiquitous. While speech recognition is the dominant task, speech dialect identification has grown rapidly to handle the acoustic variations between regions. Geographically, Gulf and Levantine dialects are well-represented in speech corpora, whereas Maghrebi dialects face challenges due to higher phonetic complexity and code-switching with European languages. Sub-tasks include speech processing & synthesis, speech recognition, speech dialect identification, isolated word recognition, morphophonological processing, speech classification, speech emotion recognition, speech-to-text (STT), text-to-speech (TTS), text-to-speech synthesizer and Chatbots and virtual assistants. Methodologically, the cluster mainly used DL approaches. The current state-of-the-art involves self-supervised learning and fine-tuning large pre-trained models like Whisper to handle the lack of large-scale labeled dialectal audio. A major gap is the limited number of multi-speaker datasets for low-resource dialects like Sudanese. Limitations include high sensitivity to background noise. where models trained on one region fail on sub-dialects. Significant challenges remain in achieving human-like prosody in synthesis and the real-time processing of code-switching.

These patterns confirm that speech processing and synthesis (Group G) rely on strong resources, identification, and preprocessing to handle dialectal acoustic and linguistic variation effectively, directly supporting the interdependency dimension of RQ.

4.8. Information Retrieval & Extraction *(Group H)*

Information Retrieval & Extraction (Group H) aims to locate, extract, and structure relevant information from large volumes of unstructured dialectal text. Key tasks include Named Entity Recognition (NER)—identifying names, locations, and organizations—as well as Question Answering (QA) and Event Detection. The role of Group H is to transform "noisy" dialectal data into actionable knowledge. It enables systems to answer user queries in their native dialect and allows researchers to track events and trends across social media platforms where formal Arabic (MSA) is rarely used. Group H represents nearly 6%. While it is one of the smaller clusters in terms of total volume, it is a high-impact area for industrial applications. Named Entity Recognition (NER) is the primary focus within this group. The coverage shows a trend in 2023 toward domain-specific extraction (healthcare or news), particularly focusing on the Levantine and Maghrebi dialects due to the higher volume of available social media data. Sub-tasks in Group H include information retrieval & extraction, text summarization, event detection, question identification, question answering, text detection, Content Detection and information retrieval. Methodologies have shifted from ML and DL to transformer-based sequence labeling. Recent contributions leverage BERT-based models (AraBERT, CamelBERT) and LLMs for entity extraction, demonstrating that contextual embeddings are superior at handling the morphological ambiguity and orthographic variations inherent in dialectal writing. The primary gap is the lack of standardized gold standard benchmarks for dialectal NER and QA compared to MSA. Limitations involve the "Out-of-Vocabulary" (OOV) problem, where dialectal slang and unique regional entities are missed by models trained on news-heavy datasets. Challenges include accurately extracting relationships between entities in short, informal posts that lack standard punctuation and syntactic structure.

These patterns confirm that information retrieval and extraction (Group H) deliver superior results only when supported by solid preprocessing, identification, and domain-specific resources, directly supporting the interdependency dimension of RQ.

5. CROSS-APPLICATION SYNTHESIS OF FINDINGS

Of the 101 high-quality papers analyzed, Sentiment Analysis (Group A) constitutes 32%, Resource Building (Group C) 21%, Identification & Code-Switching (Group D) 10 %, Machine Translation (Group E) 10 %, Preprocessing (Group B) 10 %, Speech Processing (Group G) 8 %, Information Retrieval (Group H) 6 %, and Author Profiling (Group F) 3 %. Maghrebi dialects dominate (≈40 % of studies), followed by Gulf and Levantine, while Iraqi, Sudanese, and Yemeni dialects remain severely under-represented.).

This section synthesizes cross-application findings to directly answer the research question by highlighting the methodological evolution from ML to Transformers and LLMs, exposing major dialectal imbalances, tracing temporal growth and research maturity from 2020–2025, and emphasizing how interdependent task advancements collectively guide the development of more inclusive, reproducible, and culturally aware Arabic dialect NLP systems.

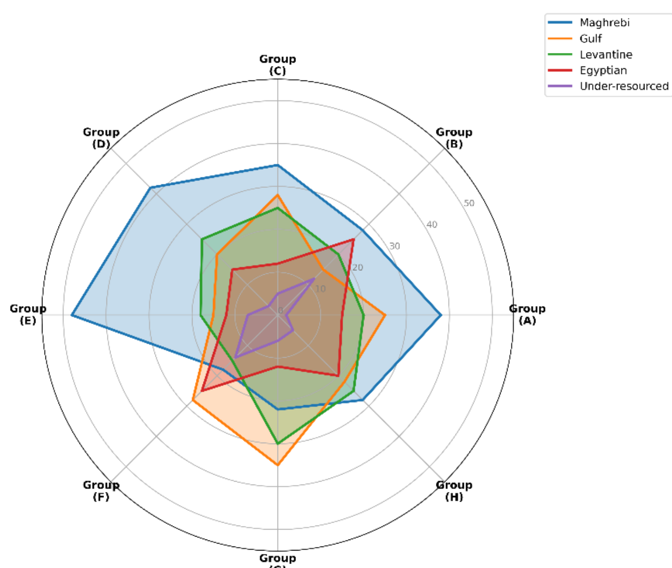

Figure 3. Dialect Research Intensity Radar Chart

Figure 3 illustrates the research intensity of the five dialect groups across the eight functional clusters. Maghrebi dialects clearly dominate most tasks (especially Sentiment Analysis, Identification, and Machine Translation), while Gulf and

Levantine dialects show moderate to strong coverage in Speech Processing and Resource Building. In contrast, the combined Iraqi/Sudanese/Yemeni group remains critically under-represented across all clusters. This visualization directly supports the research question by quantifying task-dialect interdependencies and highlighting the urgent need for more balanced resource allocation in Arabic dialect NLP.

5.1. Methodological Evolution Across Application.

Empirical scrutiny across the delineated application clusters reveals a discernible progression in methodological paradigms, transitioning from conventional machine learning frameworks, such as support vector machines and random forests prevalent in early inquiries (2020-2021) [23, 25, 29, 36] , to deep learning architectures like convolutional and recurrent neural networks [41, 48, 52], and culminating in transformer-based models (e.g., AraBERT, MARBERT) by 2023–2025 [58, 70, 77, 87, 92, 96]. This evolution is evidenced in sentiment analysis (Group A) and speech processing (Group G), where hybrid ensembles yield precision enhancements of 10–20% [58, 67, 77, 106], while resource-constrained tasks like machine translation (Group E) leverage transfer learning to mitigate data paucity [50, 66, 74]. As tabulated in Table 5, this shift underscores a maturation toward scalable, context-aware systems, albeit with persistent computational exigencies in low-resource dialects [87, 107, 114]. Table 6 reports this methodology shift over the study period. And Figure 4 represents the evolution of modeling approaches.

Methodological evolution from traditional ML to Transformers and LLMs is not merely technical but reflects a fundamental improvement in capturing dialectal nuances such as morphological richness and code-switching. Recent studies confirm that Transformer-based models achieve 15–25% higher performance in multi-dialect settings compared to earlier approaches [1, 22]. However, this progress remains constrained by heavy dependence on high-resource dialects, limiting generalizability.

TABLE 6. Summary of the methodological evolution across applications over 2020-2025

Year	Dominance Methodology
2020-2021	Dominance of Traditional Machine Learning and early Deep Learning (CNNs/LSTMs).
2022-2024	A massive surge in Transformer-based architectures, specifically AraBERT and MARBERT.
2025- beyond	Emerging dominance of Large Language Models (LLMs) and zero-shot capabilities.

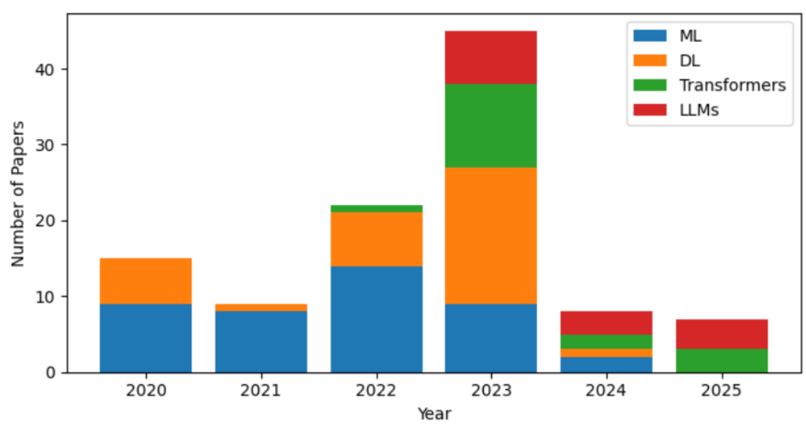

FIGURE 4. Evolution of Modeling Approaches in Arabic Dialect NLP (2020–2025)

5.2. Dialectal Coverage and Structural Imbalance.

The taxonomic aggregation elucidates pronounced structural imbalances in dialectal representation, with Maghrebi variants dominating empirical coverage (approximately 40% across clusters, per Table 5), particularly in sentiment analysis (Group A) and resource building (Group C) [25, 36, 41, 43, 45, 71, 87], whereas Gulf and Levantine dialects exhibit moderate penetration in speech synthesis (Group G) and identification (Group D) [27, 33, 63, 77]. In contrast, Iraqi, Sudanese, and Yemeni dialects manifest negligible inclusion (0–2% per cluster) [24, 34, 47], engendering systematic biases that impede generalizability. This disparity, rooted in resource availability and geographic research focal point [25, 43, 71], amplifies morphological and syntactic heterogeneities, necessitating augmented corpora to redress inequities and foster inclusive computational linguistics paradigms [87, 107, 112]. Figure 5 illustrates the Dialects Coverage with respect to the application clusters.

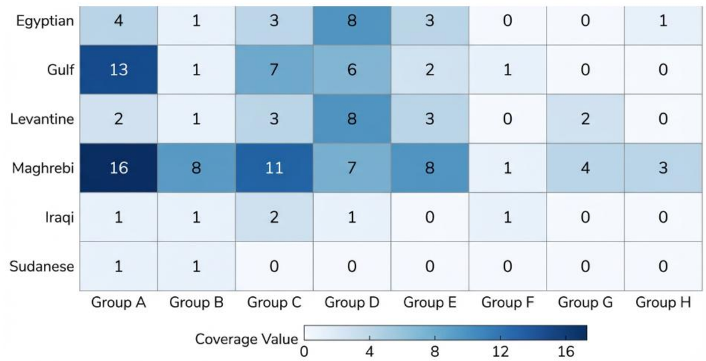

FIGURE 5: Dialects Coverage with respect to the application clusters

5.3. Temporal Trends and Research Maturity.

Chronological analysis delineates a surge in scholarly output from 2020-2022, characterized by foundational explorations in preprocessing (Group B) and corpora development (Group C) [23, 25, 43, 48, 52], escalating to a zenith in 2023–2025 with over 60% of entries emphasizing advanced applications like author profiling (Group F) and information extraction (Group H) [83, 88, 104, 119]. This trajectory reflects increasing research maturity, as initial rule-based methodologies yield to zero-shot large language models [92, 96, 107, 114, 118], enhancing efficacy in multi-dialectal contexts (e.g., 15–25% accuracy gains in Group A) [77, 87, 96]. Table 5 corroborates this maturation, highlighting a pivot toward interdisciplinary integrations [106, 112, 117], though sustained efforts are requisite to surmount persistent gaps in understudied temporal domains [24, 34, 47].

5.4. Implications of the Cross-Application Perspective.

The integrative lens afforded by cross-application synthesis illuminates synergistic interdependencies, wherein advancements in dialect identification (Group D) underpin translational fidelity (Group E) and sentiment granularity (Group A) [29, 63, 66, 77] , potentially yielding unified frameworks for code-switched environments [27, 77, 92]. This perspective reveals emergent implications for equitable NLP deployment, such as mitigating biases through dialect-agnostic resources (Group C) [45, 71, 87] and ethical profiling safeguards (Group F) [83, 104], while fostering cross-task transfer learning to elevate performance metrics (e.g., F1 scores exceeding 85% in hybrid models) [77, 87, 96]. Ultimately, as synthesized in Table 5, this holistic vantage propels future inquiries toward reproducible, culturally attuned systems [107, 112, 118], bridging extant disparities in Arabic dialectal processing.

From a broader perspective, these findings underscore the urgency of low-resource language modelling strategies, cross-dialect transfer learning, and ethical considerations regarding representational bias in Arabic NLP. Future systems must move beyond high-resource dialects to ensure equitable language technologies for the entire Arabic-speaking world.

From these findings, two key implications emerge for future research. First, developing hybrid frameworks that combine dialect-specific fine-tuning with LLM zero-shot capabilities will be essential to address data scarcity. Second, prioritizing cross-dialect transfer learning and shared benchmarks is critical to reduce the current imbalance between well-resourced (Maghrebi, Egyptian) and under-resourced dialects (Sudanese, Yemeni, Iraqi). These directions will help create more equitable and robust Arabic dialect NLP systems.

6. Challenges And Future Directions

6.1. Challenges and limitations

Drawing on the gap analysis and challenge taxonomy illustrated in Table 7 and Table 8, the current landscape of Arabic dialect NLP exhibits pronounced dialectal imbalance and task-level asymmetries: Egyptian, Gulf, Levantine, and Maghrebi varieties receive sustained attention across sentiment analysis, resource building, identification, MT, speech, NER/IE, and author profiling, whereas Iraqi, Sudanese, and Yemeni dialects remain markedly under-researched; limiting generalizability and equitable progress across the field. These disparities are compounded by structural constraints; data scarcity (notably the lack of parallel corpora, few speech datasets, and minimal city-level resources) that depress performance in MT, ASR, and dialect identification; pervasive linguistic variation (orthographic inconsistency, code-switching, and morphological complexity) that challenges modeling across all tasks; evaluation shortcomings (absence of standardized benchmarks, inconsistent metrics, and weak reproducibility) that hinder fair comparison; and model-adaptation limitations (LLM zero-shot brittleness, domain shift, and high computational cost) that restrict scalable deployment and cross-domain robustness.

TABLE 7. Research Gap Analysis by Dialect and Task

Task	Egyptian	Gulf	Levantine	Maghrebi	Iraqi	Sudanese	Yemeni
Sentiment Analysis	✓✓	✓✓	✓	✓✓	✗	✗	✗
Resource Building	✓	✓✓	✓✓	✓✓	✗	✗	✗
Identification	✓✓	✓✓	✓✓	✓✓	✓	✗	✗
Machine Translation	✓	✓	✓	✓✓	✗	✗	✗
Speech Processing	✓	✓✓	✓✓	✓	✗	✗	✗
NER/Information Extraction	✓	✓	✓	✓	✗	✗	✗
Author Profiling	✓	✓	✗	✓	✗	✗	✗

Key: ✓✓ = Well-researched, ✓ = Some research, ✗ = Under-researched

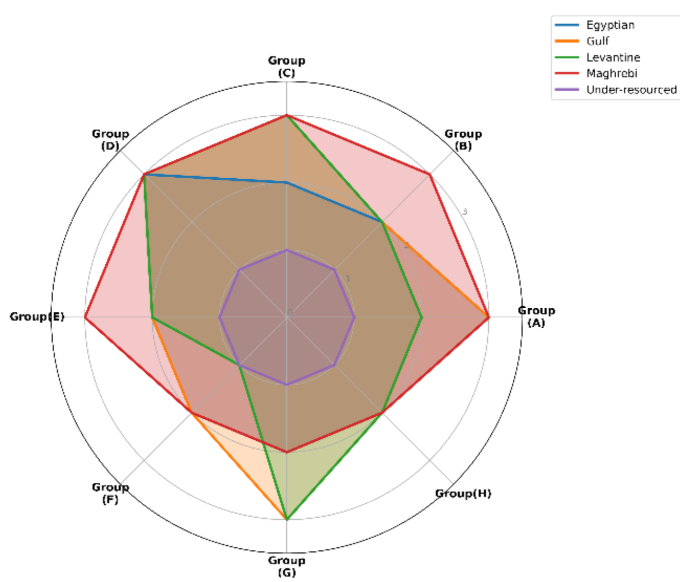

Figure 6. Research Gap Analysis by Dialect and Task

DOI: https://doi.org/10.70715/jitcai.2026.v3.i2.051

Figure 6 visualizes the coverage gaps identified in Table 7 on a comparative scale (3=well-researched, 2=some, 1=under). Egyptian, Gulf, Levantine, and Maghrebi dialects receive consistent attention in high-impact tasks such as Sentiment Analysis, Resource Building, and Identification, whereas Iraqi, Sudanese, and Yemeni dialects score the lowest across nearly all categories. These radar patterns confirm the structural imbalances discussed in recent surveys [2, 22] and underscore that future work must prioritise low-resource dialects and cross-dialect transfer learning to achieve truly inclusive Arabic NLP systems.

According to [22] , to address these limitations, future research should focus on collecting more diverse dialectal data, specialized pre-training for Arabic dialects, cross-dialectal training, and developing culturally informed models. Additionally, adversarial training, data augmentation, improved speech recognition and synthesis, transfer learning, and user-centered design should be explored. Moreover, it is essential to establish ethical guidelines, evaluation metrics, and standardized benchmarks will also be crucial for advancing Arabic dialect processing in NLP models. DL-based models and pre-trained models adapted for Arabic dialects are particularly beneficial in applications such as social media analysis, event detection and analysis, MT, chatbots, virtual assistants, speech recognition, educational tools, cultural preservation, entertainment, public services, and healthcare.

TABLE 8. Challenge Taxonomy

Challenge Category	Specific Challenges	Affected Tasks	Proposed Solutions
Data Scarcity	• Limited parallel corpora • Few speech datasets • Lack of city-level data	MT, ASR, DI	• Data augmentation • Crowdsourcing • Synthetic data generation
Linguistic Variation	• Orthographic inconsistency • Code-switching • Morphological complexity	All tasks	• Character-level models • Multi-task learning • Subworld tokenization
Evaluation	• No standardized benchmarks • Inconsistent metrics • Poor reproducibility	All tasks	• Shared tasks • Unified evaluation frameworks • Open-source models
Model Adaptation	• LLM zero-shot limitations • Domain shift • Computational cost	All tasks	• Fine-tuning strategies • Adapter modules • Efficient training

1.1. Future Directions

Future trends in Arabic dialectal NLP focus on leveraging DL, LLMs, and prompt engineering to build more inclusive, culturally aware systems and better address the field's linguistic complexities summarized in table 9, as also underscored by [22].

Table 9. Future Directions in Arabic Dialectal NLP

Future Direction	Description
LLM Enhancement for Dialects	Fine-tuning large language models on dialect-rich corpora to improve generation, translation, and understanding.
Dialect-Aware Prompt Engineering	Designing prompts that explicitly guide LLMs to handle dialect-specific vocabulary and grammar.
Dialect-Specific Pretraining	Pretraining language models on extensive dialectal datasets to capture linguistic nuances and variations.
Multimodal Dialect Modelling	Integrating text, speech, and visual inputs to build richer and more accurate dialectal models.
Multi-Dialect Unified Models	Developing architectures that can process and generalize across several Arabic dialects simultaneously.

Continual Learning for Evolving Dialects	Enabling models to update their knowledge as dialects evolve over time, especially in social media contexts.
Low-Resource Dialect Adaptation	Using zero-shot, unsupervised, and synthetic data techniques to support under-represented dialects like Sudanese and Yemeni.
Improved Code-Switching Handling	Building models capable of processing Arabic–Arabic and Arabic–foreign language mixing with higher accuracy.
Domain Adaptation for Dialects	Tailoring models for specialized contexts such as healthcare, finance, education, or governmental sectors.
Privacy-Preserving Dialectal NLP	Developing techniques to protect sensitive user data during dialectal model training and deployment.
Synthetic Data Generation	Creating reliable synthetic dialect datasets using machine translation, augmentation, and human post-editing.
Real-World System Deployment	Building practical, dialect-aware applications such as chatbots, educational tools, health assistants, and voice systems.

7. CONCLUSION

This survey has provided a unified, application-oriented, and trend-driven synthesis of Arabic Dialectal Natural Language Processing (ADNLP) research during its most productive period (2020–2025) thereby directly addressing the research question posed in the Introduction. By systematically screening approximately 400 studies and conducting an in-depth analysis of 101 high-quality papers, we introduced an eight-cluster functional taxonomy that maps the complex interrelationships among tasks, resources, and methodologies in the field. Our findings reveal a clear maturation trajectory: Sentiment Analysis remains the dominant application (32 % of the literature), followed by Resource Building (21 %), while Dialect Identification and Code-Switching have evolved from standalone tasks into essential prerequisites for downstream systems. The field has undergone a definitive methodological shift from traditional machine-learning models to Transformer-based architectures (AraBERT, MARBERT) and, more recently, to Large Language Models, with research output peaking between 2022 and 2025.

The study also highlights uneven coverage across dialects. While Maghrebi, Egyptian, Gulf, and Levantine varieties receive the majority of attention, Iraqi, Sudanese, and Yemeni dialects remain under-represented. Key gaps include limited high-quality datasets for under-resourced varieties, few reproducible benchmarks, and limited cross-dialect transfer learning.

By providing a structured taxonomy, trend analysis, and dialect-by-task gap mapping, this survey offers a clear reference for ongoing and future work in Arabic dialect NLP. Future research should focus on expanding resources for under-represented dialects, establishing shared evaluation benchmarks, and exploring effective ways to combine dialect-specific models with the capabilities of generative LLMs. These steps will help build more balanced and inclusive language technologies for the Arabic-speaking

FUNDING INFORMATION
The authors state no funding is involved.

CONFLICT OF INTEREST STATE
The authors state no conflict of interest.

ETHICAL APPROVAL
This paper does not involve people or animals; no investigation has involved human subjects. Therefore, the authors did not seek approval from any institutional review board.

DATA AVAILABILITY
The data that support the findings of this study are available from the corresponding author, [S], upon reasonable request.

CONSENT TO PUBLISH
Not applicable.

AUTHOR CONTRIBUTION

The authors Hani Iwidat and Mamou Abu Helou contributed equally to this work.

8. REFERENCES

[1] M. Mashaabi, S. Al-Khalifa, and H. Al-Khalifa, *A Survey of Large Language Models for Arabic Language and its Dialects*. 2024.

[2] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," *Journal of King Saud University - Computer and Information Sciences,* vol. 33, no. 5, pp. 497-507, 2021/06/01/ 2021, doi: https://doi.org/10.1016/j.jksuci.2019.02.006.

[3] A. Ahmed, N. Ali, M. Alzubaidi, W. Zaghouani, A. A. Abd-alrazaq, and M. Househ, "Freely Available Arabic Corpora: A Scoping Review," *Computer Methods and Programs in Biomedicine Update,* vol. 2, p. 100049, 2022/01/01/ 2022, doi: https://doi.org/10.1016/j.cmpbup.2022.100049.

[4] K. Darwish *et al.*, "A Panoramic Survey of Natural Language Processing in the Arab World," *Communications of the ACM,* vol. 64, no. 4, pp. 72-81, 2021, doi: 10.1145/3447735.

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

[6] I. Abu Farha and W. Magdy, "A comparative study of effective approaches for Arabic sentiment analysis," *Information Processing & Management,* vol. 58, no. 2, p. 102438, 2021, doi: 10.1016/j.ipm.2020.102438.

[7] S. Harrat, K. Meftouh, and K. Smaïli, "Machine translation for Arabic dialects (survey)," *Information Processing & Management,* vol. 56, pp. 262-273, 08/01 2017, doi: 10.1016/j.ipm.2017.08.003.

[8] I. Hamed, C. Sabty, S. Abdennadher, N. T. Vu, T. Solorio, and N. Habash, "A Survey of Code-switched Arabic NLP: Progress, Challenges, and Future Directions," Abu Dhabi, UAE, January 2025: Association for Computational Linguistics, in Proceedings of the 31st International Conference on Computational Linguistics, pp. 4561-4585. [Online]. Available: https://aclanthology.org/2025.coling-main.307/. [Online]. Available: https://aclanthology.org/2025.coling-main.307/

[9] A. Shoufan and S. Alameri, "Natural Language Processing for Dialectical Arabic: A Survey," Beijing, China, July 2015: Association for Computational Linguistics, in Proceedings of the Second Workshop on Arabic Natural Language Processing, pp. 36-48, doi: 10.18653/v1/W15-3205. [Online]. Available: https://aclanthology.org/W15-3205/

https://doi.org/10.18653/v1/W15-3205

[10] M. J. Althobaiti, *Automatic Arabic Dialect Identification Systems for Written Texts: A Survey*. 2020.

[11] A. Elnagar, S. M. Yagi, A. B. Nassif, I. Shahin, and S. A. Salloum, "Systematic Literature Review of Dialectal Arabic: Identification and Detection," *IEEE Access,* vol. 9, pp. 31010-31042, 2021, doi: 10.1109/ACCESS.2021.3059504.

[12] Y. Matrane, F. Benabbou, and N. Sael, "A systematic literature review of Arabic dialect sentiment analysis," *Journal of King Saud University - Computer and Information Sciences,* vol. 35, no. 6, p. 101570, 2023/06/01/ 2023, doi: https://doi.org/10.1016/j.jksuci.2023.101570.

[13] A. Elsaid, A. Mohammed, L. Fattouh, and M. Sakre, "A Comprehensive Review of Arabic Text Summarization," *IEEE Access,* vol. 10, pp. 1-1, 01/01 2022, doi: 10.1109/ACCESS.2022.3163292.

[14] A. Dhouib, A. Othman, O. El Ghoul, M. K. Khribi, and A. Al Sinani, "Arabic Automatic Speech Recognition: A Systematic Literature Review," *Applied Sciences*, vol. 12, no. 17, p. 8898doi: 10.3390/app12178898.

[15] H. Rahab, A. Zitouni, and M. Djoudi, "Arabic Fake News and Spam Handling: Methods, Resources and Opportunities," in *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP)*, 20-21 Nov. 2021 2021, pp. 1-7, doi: 10.1109/AI-CSP52968.2021.9671174.

[16] M. Labied and A. Belangour, "Moroccan Dialect “Darija” Automatic Speech Recognition: A Survey," in *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, 16-18 July 2021 2021, pp. 208-213, doi: 10.1109/PRML52754.2021.9520690.

[17] S. Brachemi-Meftah and F. Barigou, "Algerian Dialect Sentiment Analysis: Sate of Art," *2020 21st International Arab Conference on Information Technology (ACIT),* pp. 1-7, 2020.

[18] H. Hejazi and A. Khamees, "Opinion mining for Arabic dialect in social media data fusion platforms: A systematic review," *Fusion: Practice and Applications,* vol. 9, pp. 08-28, 01/01 2022, doi: 10.54216/FPA.090101.

[19] F. Alqahtani and M. Dohler, "Survey of Authorship Identification Tasks on Arabic Texts," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.,* vol. 22, no. 4, apr, articleno = 93 , numpages = 24 2023, doi: 10.1145/3564156.
[20] M. Al-Ayyoub, A. Nuseir, K. Alsmearat, Y. Jararweh, and B. Gupta, "Deep learning for Arabic NLP: A survey," *Journal of Computational Science,* vol. 26, pp. 522-531, 2018/05/01/ 2018, doi: https://doi.org/10.1016/j.jocs.2017.11.011.
[21] S. Al Katat, I. Bensalem, P. Rosso, and S. Chikhi, "Natural Language Processing for Arabic Sentiment Analysis: A Systematic Literature Review," *IEEE Transactions on Big Data,* vol. 10, no. 5, 2024, doi: 10.1109/TBDATA.2024.3363633.
[22] A. Dahou *et al.*, "A Survey on Dialect Arabic Processing and Analysis: Recent Advances and Future Trends," *ACM Transactions on Asian and Low-Resource Language Information Processing,* vol. 24, 07/03 2025, doi: 10.1145/3747290.
[23] L. Moudjari and K. Akli-Astouati, "An Experimental Study on Sentiment Classification of Algerian Dialect Texts," *Procedia Computer Science,* vol. 176, pp. 1151-1159, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.procs.2020.09.111.
[24] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic Speech Emotion Recognition From Saudi Dialect Corpus," *IEEE Access,* vol. 9, pp. 127081-127085, 2021, doi: 10.1109/ACCESS.2021.3110992.
[25] H. Alhammi and K. Haddar, "Building a Libyan Dialect Lexicon-Based Sentiment Analysis System Using Semantic Orientation of Adjective-Adverb Combinations," *International Journal of Computer Theory and Engineering,* vol. 12, pp. 145-150, 01/01 2020, doi: 10.7763/IJCTE.2020.V12.1280.
[26] K. Bousmaha, K. Hamadouche, I. Gourara, and L. Belguith, "DZ-OPINION: Algerian Dialect Opinion Analysis Model with Deep Learning Techniques," *Revue d'Intelligence Artificielle,* vol. 36, pp. 897-903, 12/31 2022, doi: 10.18280/ria.360610.
[27] K. Lounnas, M. Abbas, M. Lichouri, M. Hamidi, H. Satori, and H. Teffahi, "Enhancement of spoken digits recognition for under-resourced languages: case of Algerian and Moroccan dialects," *International Journal of Speech Technology,* vol. 25, no. 2, pp. 443-455, 2022/06/01 2022, doi: 10.1007/s10772-022-09971-y.
[28] A. Slim, A. Melouah, U. Faghihi, and K. Sahib, "Improving Neural Machine Translation for Low Resource Algerian Dialect by Transductive Transfer Learning Strategy," *Arabian Journal for Science and Engineering,* vol. 47, no. 8, pp. 10411-10418, 2022/08/01 2022, doi: 10.1007/s13369-022-06588-w.
[29] S. Mihi *et al.*, "MSTD: Moroccan Sentiment Twitter Dataset," *International Journal of Advanced Computer Science and Applications,* vol. 11, p. 10, 01/01 2020.
[30] D. Al-Ghadhban and N. Al-Twairesh, "Nabiha: An Arabic Dialect Chatbot," *International Journal of Advanced Computer Science and Applications,* vol. 11, 01/01 2020, doi: 10.14569/IJACSA.2020.0110357.
[31] Z. Muhammad Zain, "Ranking Beauty Clinics in Riyadh using Lexicon-Based Sentiment Analysis and Multiattribute-Utility Theory," *International Journal of Advanced Computer Science and Applications,* vol. 11, pp. 66-75, 11/01 2020, doi: 10.14569/IJACSA.2020.0111009.
[32] M. A. Sghaier and M. Zrigui, "Rule-Based Machine Translation from Tunisian Dialect to Modern Standard Arabic," *Procedia Computer Science,* vol. 176, pp. 310-319, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.procs.2020.08.033.
[33] A. Bayazed, O. Torabah, R. Alsulami, D. Alahmadi, and K. Saeedi, "SDCT: Multi-Dialects Corpus Classification for Saudi Tweets," *International Journal of Advanced Computer Science and Applications,* vol. 11, 01/01 2020, doi: 10.14569/IJACSA.2020.0111128.
[34] S. Alotaibi, R. Mehmood, I. Katib, O. Rana, and A. Albeshri, "Sehaa: A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter, Apache Spark, and Machine Learning," *Applied Sciences*, vol. 10, no. 4, p. 1398doi: 10.3390/app10041398.
[35] W. Farhan *et al.*, "Unsupervised dialectal neural machine translation," *Information Processing & Management,* vol. 57, no. 3, p. 102181, 2020/05/01/ 2020, doi: https://doi.org/10.1016/j.ipm.2019.102181.
[36] A. Fashwan and S. Alansary, "A Morphologically Annotated Corpus and a Morphological Analyzer for Egyptian Arabic," *Procedia Computer Science,* vol. 189, pp. 203-210, 2021/01/01/ 2021, doi: https://doi.org/10.1016/j.procs.2021.05.084.
[37] I. Guellil *et al.*, "A Semi-supervised Approach for Sentiment Analysis of Arab(ic+izi) Messages: Application to the Algerian Dialect," *SN Computer Science,* vol. 2, no. 2, p. 118, 2021/02/27 2021, doi: 10.1007/s42979-021-00510-1.
[38] A. Slim, A. Melouah, Y. Faghihi, and K. Sahib, "Algerian Dialect Translation Applied on COVID-19 Social Media Comments," in *Artificial Intelligence and Renewables Towards an Energy Transition*, Cham, M. Hatti, Ed., 2021// 2021: Springer International Publishing, pp. 716-726.

[39] A. J. Askar and N. Nur, "Annotated Corpus of Mesopotamian-Iraqi Dialect for Sentiment Analysis in Social Media," *International Journal of Advanced Computer Science and Applications,* vol. 12, 01/01 2021, doi: 10.14569/IJACSA.2021.0120413.

[40] A. Hussein and I. Moawad, "Arabic Sentiment Analysis for Multi-dialect Text using Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications,* vol. 12, 01/01 2021, doi: 10.14569/IJACSA.2021.0121286.

[41] I. Touahri and A. Mazroui, "Enhancement of a multi-dialectal sentiment analysis system by the detection of the implied sarcastic features," *Knowledge-Based Systems,* vol. 227, p. 107232, 2021/09/05/ 2021, doi: https://doi.org/10.1016/j.knosys.2021.107232.

[42] M. Garouani and J. Kharroubi, "Towards a New Lexicon-Based Features Vector for Sentiment Analysis: Application to Moroccan Arabic Tweets," in *Advances in Information, Communication and Cybersecurity*, Cham, Y. Maleh, M. Alazab, N. Gherabi, L. a. Tawalbeh, and A. A. Abd El-Latif, Eds., 2022// 2022: Springer International Publishing, pp. 67-76.

[43] M. A. Djebbi and R. Ouersighni, "TunTap: A Tunisian Dataset for Topic and Polarity Extraction in Social Media," in *Computational Collective Intelligence*, Cham, N. T. Nguyen, Y. Manolopoulos, R. Chbeir, A. Kozierkiewicz, and B. Trawiński, Eds., 2022// 2022: Springer International Publishing, pp. 507-519.

[44] T. Omran, B. Sharef, C. Grosan, and Y. Li, "Transfer Learning and Sentiment Analysis of Bahraini Dialects Sequential Text Data Using Multilingual Deep Learning Approach," *SSRN Electronic Journal,* 01/01 2022, doi: 10.2139/ssrn.4111929.

[45] H. Al-Khalifa, L. Aldhubayi, F. Alzahrani, R. Alrowais, S. Alowa, and H. qawara, *A Dataset for Detecting Humor in Arabic Text*. 2022.

[46] A. C. Mazari and H. Kheddar, "Deep Learning-based Analysis of Algerian Dialect Dataset Targeted Hate Speech, Offensive Language and Cyberbullying," *International Journal of Computing and Digital Systems,* vol. 13, pp. 965-972, 04/16 2023, doi: 10.12785/ijcds/130177.

[47] K. Lounnas, M. Lichouri, and M. Abbas, "Analysis of the Effect of Audio Data Augmentation Techniques on Phone Digit Recognition For Algerian Arabic Dialect," in *2022 International Conference on Advanced Aspects of Software Engineering (ICAASE)*, 17-18 Sept. 2022 2022, pp. 1-5, doi: 10.1109/ICAASE56196.2022.9931574.

[48] T. Alqurashi, "Applying a Character-Level Model to a Short Arabic Dialect Sentence: A Saudi Dialect as a Case Study," *Applied Sciences*, vol. 12, no. 23, p. 12435doi: 10.3390/app122312435.

[49] A. Masmoudi, C. Aloulou, A. G. S. Abdellahi, and L. H. Belguith, "Automatic diacritization of Tunisian dialect text using SMT model," *International Journal of Speech Technology,* vol. 25, no. 1, pp. 89-104, 2022/03/01 2022, doi: 10.1007/s10772-021-09864-6.

[50] A. Safieh, I. Alhaol, and G. Rawan, "End-to-end Jordanian dialect speech-to-text self-supervised learning framework," *Frontiers in Robotics and AI,* vol. 9, 12/22 2022, doi: 10.3389/frobt.2022.1090012.

[51] M. AbdelHamid, A. Jafar, and Y. Rahal, "Levantine hate speech detection in twitter," *Social Network Analysis and Mining,* vol. 12, no. 1, p. 121, 2022/08/29 2022, doi: 10.1007/s13278-022-00950-4.

[52] M. Garouani and J. Kharroubi, "MAC: An Open and Free Moroccan Arabic Corpus for Sentiment Analysis," in *Innovations in Smart Cities Applications Volume 5*, Cham, M. Ben Ahmed, A. A. Boudhir, İ. R. Karaș, V. Jain, and S. Mellouli, Eds., 2022// 2022: Springer International Publishing, pp. 849-858.

[53] R. Tachicart and K. Bouzoubaa, "Moroccan Arabic vocabulary generation using a rule-based approach," *Journal of King Saud University - Computer and Information Sciences,* vol. 34, no. 10, Part A, pp. 8538-8548, 2022/11/01/ 2022, doi: https://doi.org/10.1016/j.jksuci.2021.02.013.

[54] S. Hajbi, S. ChHajbi, Y. ihab, R. Ed-Dali, and R. Korchiyne, "Natural Language Processing Based Approach to Overcome Arabizi and Code Switching in Social Media Moroccan Dialect," in *Advances in Information, Communication and Cybersecurity*, Cham, Y. Maleh, M. Alazab, N. Gherabi, L. a. Tawalbeh, and A. A. Abd El-Latif, Eds., 2022// 2022: Springer International Publishing, pp. 57-66.

[55] A. Emna, S. Kchaou, and R. Boujelban, "Neural Machine Translation of Low Resource Languages: Application to Transcriptions of Tunisian Dialect," in *Intelligent Systems and Pattern Recognition*, Cham, A. Bennour, T. Ensari, Y. Kessentini, and S. Eom, Eds., 2022// 2022: Springer International Publishing, pp. 234-247.

[56] J. Younes, H. Achour, E. Souissi, and A. Ferchichi, "Romanized Tunisian dialect transliteration using sequence labelling techniques," *Journal of King Saud University - Computer and Information Sciences,* vol. 34, no. 3, pp. 982-992, 2022/03/01/ 2022, doi: https://doi.org/10.1016/j.jksuci.2020.03.008.

[57] A. Mekki, I. Zribi, M. Ellouze, and L. H. Belguith, "Sentence boundary detection of various forms of Tunisian Arabic," *Language Resources and Evaluation,* vol. 56, no. 1, pp. 357-385, 2022/03/01 2022, doi: 10.1007/s10579-021-09538-4.

[58] F. Husain, H. Al-Ostad, and H. Omar, *A Weak Supervised Transfer Learning Approach for Sentiment Analysis to the Kuwaiti Dialect*. 2022, pp. 161-173.

[59] A. C. Mazari and A. Djeffal, "Sentiment Analysis of Algerian Dialect Using Machine Learning and Deep Learning with Word2vec," *Informatica,* vol. 46, 07/29 2022, doi: 10.31449/inf.v46i6.3340.

[60] A. A. Al Shamsi and S. Abdallah, "Sentiment Analysis of Emirati Dialect," *Big Data and Cognitive Computing*, vol. 6, no. 2, p. 57doi: 10.3390/bdcc6020057.

[61] B. Hdioud and M. E. H. Tirari, "Sentiment Analysis of Moroccan Dialect Using Deep Learning," in *Proceedings of the 5th International Conference on Big Data and Internet of Things*, Cham, M. Lazaar, C. Duvallet, A. Touhafi, and M. Al Achhab, Eds., 2022// 2022: Springer International Publishing, pp. 457-466.

[62] S. Jaballi, S. Zrigui, M. Sghaier, D. Berchech, and M. Zrigui, *Sentiment Analysis of Tunisian Users on Social Networks: Overcoming the Challenge of Multilingual Comments in the Tunisian Dialect*. 2022, pp. 176-192.

[63] O. Tirosh-Becker and O. Becker, "TAJA Corpus: Linguistically Tagged Written Algerian Judeo-Arabic Corpus," *Journal of Jewish Languages,* vol. 10, pp. 24-53, 06/01 2022, doi: 10.1163/22134638-bja10020.

[64] A. Messaoudi, H. Haddad, C. Fourati, M. B. Hmida, A. B. Elhaj Mabrouk, and M. Graiet, "Tunisian Dialectal End-to-end Speech Recognition based on DeepSpeech," *Procedia Computer Science,* vol. 189, pp. 183-190, 2021/01/01/ 2021, doi: https://doi.org/10.1016/j.procs.2021.05.082.

[65] M. Mhamed *et al.*, "A deep CNN architecture with novel pooling layer applied to two Sudanese Arabic sentiment data sets," *Journal of Information Science,* vol. 52, no. 1, pp. 285-306, 2026, doi: 10.1177/01655515231188341.

[66] H. Mahdhaoui, A. Mars, and M. Zrigui, "Active Learning with AraGPT2 for Arabic Named Entity Recognition," in *Advances in Computational Collective Intelligence*, Cham, N. T. Nguyen *et al.*, Eds., 2023// 2023: Springer Nature Switzerland, pp. 226-236.

[67] R. Kora and A. Mohammed, "An enhanced approach for sentiment analysis based on meta-ensemble deep learning," *Social Network Analysis and Mining,* vol. 13, no. 1, p. 38, 2023/03/02 2023, doi: 10.1007/s13278-023-01043-6.

[68] Y. Abdelwahab, M. Kholief, and A. A. H. Sedky, "An Experimental Survey of ASA on DL Classifiers Using Multi-dialect Arabic Texts," in *Advances in Information and Communication*, Cham, K. Arai, Ed., 2023// 2023: Springer Nature Switzerland, pp. 52-64.

[69] M. J. Althobaiti, "An open-source dataset for arabic fine-grained emotion recognition of online content amid COVID-19 pandemic," *Data in Brief,* vol. 51, p. 109745, 2023/12/01/ 2023, doi: https://doi.org/10.1016/j.dib.2023.109745.

[70] S. AlMuhaideb, Y. AlNegheimish, T. AlOmar, R. AlSabti, M. AlKathery, and G. AlOlyyan, "Analyzing Arabic Twitter-Based Patient Experience Sentiments Using Multi-Dialect Arabic Bidirectional Encoder Representations from Transformers," *Computers, Materials and Continua,* vol. 76, no. 1, pp. 195-220, 2023/06/09/ 2023, doi: https://doi.org/10.32604/cmc.2023.038368.

[71] I. Touahri, "AraBERT with GANs for High Performance Fine-Grained Dialect Classification," in *Proceedings of the 6th International Conference on Big Data and Internet of Things*, Cham, M. Lazaar, E. M. En-Naimi, A. Zouhair, M. Al Achhab, and O. Mahboub, Eds., 2023// 2023: Springer International Publishing, pp. 160-170.

[72] H. Saleh, A. Mohammad, K. Jafar, M. Solieman, B. Ahmad, and S. Hasan, "Arabic Text-to-Speech Service with Syrian Dialect," in *Intelligent Decision Technologies*, Singapore, I. Czarnowski, R. J. Howlett, and L. C. Jain, Eds., 2023// 2023: Springer Nature Singapore, pp. 109-127.

[73] S. M. Alsubhi, A. M. Alhothali, and A. A. AlMansour, "AraBig5: The Big Five Personality Traits Prediction Using Machine Learning Algorithm on Arabic Tweets," *IEEE Access,* vol. 11, pp. 112526-112534, 2023, doi: 10.1109/ACCESS.2023.3297981.

[74] M. Abdelhakim, B. Liu, and C. Sun, "Ar-PuFi: A short-text dataset to identify the offensive messages towards public figures in the Arabian community," *Expert Systems with Applications,* vol. 233, p. 120888, 2023/12/15/ 2023, doi: https://doi.org/10.1016/j.eswa.2023.120888.

[75] A. Benali, M. H. Maaloul, and L. H. Belguith, "Automatic Processing of Algerian Dialect: Corpus Construction and Segmentation," *SN Computer Science,* vol. 4, no. 5, p. 597, 2023/08/04 2023, doi: 10.1007/s42979-023-02097-1.

[76] R. Rachidi, M. A. Ouassil, E. Mouaad, B. Cherradi, S. Hamida, and S. Hassan, "Classifying toxicity in the Arabic Moroccan dialect on Instagram: a machine and deep learning approach," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 31, pp. 588-598, 07/01 2023, doi: 10.11591/ijeecs.v31.i1.pp588-598.

[77] S. El Ouahabi, S. El Ouahabi, and E. W. Dadi, "Contribution to the Moroccan Darija sentiment analysis in social networks," *Social Network Analysis and Mining,* vol. 13, no. 1, p. 138, 2023/10/20 2023, doi: 10.1007/s13278-023-01129-1.

[78] H. N. Moussa and A. Mourhir, "DarNERcorp: An annotated named entity recognition dataset in the Moroccan dialect," *Data in Brief,* vol. 48, p. 109234, 2023/06/01/ 2023, doi: https://doi.org/10.1016/j.dib.2023.109234.

[79] S. Jaballi, M. J. Hazar, S. Zrigui, H. Nicolas, and M. Zrigui, "Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for Tunisian Dialectical Facebook Content During the Spread of the Coronavirus Pandemic," in *Advances in Computational Collective Intelligence*, Cham, N. T. Nguyen *et al.*, Eds., 2023// 2023: Springer Nature Switzerland, pp. 96-109.

[80] A. H. Dahou and M. A. Cheragui, "DzNER: A large Algerian Named Entity Recognition dataset," *Natural Language Processing Journal,* vol. 3, p. 100005, 2023/06/01/ 2023, doi: https://doi.org/10.1016/j.nlp.2023.100005.

[81] S. Nasr, R. Duwairi, and M. Quwaider, "End-to-End Speech Recognition For Arabic Dialects," *Arabian Journal for Science and Engineering,* vol. 48, no. 8, pp. 10617-10633, 2023/08/01 2023, doi: 10.1007/s13369-023-07670-7.

[82] A. A. Al Shamsi and S. Abdallah, "Ensemble Stacking Model for Sentiment Analysis of Emirati and Arabic Dialects," *Journal of King Saud University - Computer and Information Sciences,* vol. 35, no. 8, p. 101691, 2023/09/01/ 2023, doi: https://doi.org/10.1016/j.jksuci.2023.101691.

[83] N. T. Mohammed, E. A. Mohammed, and H. H. Hussein, "Evaluating Various Classifiers for Iraqi Dialectic Sentiment Analysis," in *Next Generation of Internet of Things*, Singapore, R. Kumar, P. K. Pattnaik, and J. M. R. S. Tavares, Eds., 2023// 2023: Springer Nature Singapore, pp. 71-78.

[84] A. Abdedaiem, A. Dahou, and C. Mohamed Amine, "Fake News Detection in Low Resource Languages using SetFit Framework," *Inteligencia Artificial,* vol. 26, pp. 178-201, 09/20 2023, doi: 10.4114/intartif.vol26iss72pp178-201.

[85] S. Kchaou, R. Boujelbane, and L. Belguith, "Hybrid Pipeline for Building Arabic Tunisian Dialect-standard Arabic Neural Machine Translation Model from Scratch," *ACM Transactions on Asian and Low-Resource Language Information Processing,* vol. 22, 11/02 2022, doi: 10.1145/3568674.

[86] A. H. Dahou and M. A. Cheragui, "Impact of Normalization and Data Augmentation in NER for Algerian Arabic Dialect," in *Modelling and Implementation of Complex Systems*, Cham, S. Chikhi, G. Diaz-Descalzo, A. Amine, A. Chaoui, D. E. Saidouni, and M. K. Kholladi, Eds., 2023// 2023: Springer International Publishing, pp. 249-262.

[87] S. Jamal *et al.*, "In the Identification of Arabic Dialects: A Loss Function Ensemble Learning Based-Approach," in *Model and Data Engineering*, Cham, P. Fournier-Viger, A. Hassan, and L. Bellatreche, Eds., 2023// 2023: Springer Nature Switzerland, pp. 89-101.

[88] A. M. Mostafa, M. Aljasir, M. Alruily, A. Alsayat, and M. Ezz, "Innovative Forward Fusion Feature Selection Algorithm for Sentiment Analysis Using Supervised Classification," *Applied Sciences*, vol. 13, no. 4*,* p. 2074doi: 10.3390/app13042074.

[89] M. Errami, M. A. Ouassil, R. Rachidi, B. Cherradi, S. Hamida, and A. Raihani, "Investigating the Performance of BERT Model for Sentiment Analysis on Moroccan News Comments," in *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 18-19 May 2023 2023, pp. 1-8, doi: 10.1109/IRASET57153.2023.10152965.

[90] H. Xinyuan, N. Verma, B. Odoom, U. Pradeep, M. Wiesner, and S. Khudanpur, *JHU IWSLT 2023 Multilingual Speech Translation System Description*. 2023, pp. 302-310.

[91] G. Bourahouat, M. Abourezq, and N. Daoudi, "Leveraging Moroccan Arabic Sentiment Analysis Using AraBERT and QARIB," in *Innovations in Smart Cities Applications Volume 6*, Cham, M. Ben Ahmed, A. A. Boudhir, D. Santos, R. Dionisio, and N. Benaya, Eds., 2023// 2023: Springer International Publishing, pp. 299-310.

[92] M. Al-Fetyani, M. Al-Barham, G. Abandah, A. Alsharkawi, and M. Dawas, "MASC: Massive Arabic Speech Corpus," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 9-12 Jan. 2023 2023, pp. 1006-1013, doi: 10.1109/SLT54892.2023.10022652.

[93] A. H. Dahou and M. A. Cheragui, "Named Entity Recognition for Algerian Arabic Dialect in Social Media," in *12th International Conference on Information Systems and Advanced Technologies "ICISAT 2022"*, Cham, M. R. Laouar, V. E. Balas, B. Lejdel, S. Eom, and M. A. Boudia, Eds., 2023// 2023: Springer International Publishing, pp. 135-145.

[94] K. Essefar, H. Ait Baha, A. El Mahdaouy, A. El Mekki, and I. Berrada, "OMCD: Offensive Moroccan Comments Dataset," *Language Resources and Evaluation,* vol. 57, no. 4, pp. 1745-1765, 2023/12/01 2023, doi: 10.1007/s10579-023-09663-2.

[95] H. Alostad, S. Dawiek, and H. Davulcu, "Q8VaxStance: Dataset Labeling System for Stance Detection towards Vaccines in Kuwaiti Dialect," *Big Data and Cognitive Computing*, vol. 7, no. 3*,* p. 151doi: 10.3390/bdcc7030151.

[96] N. Habbat, H. Nouri, H. Anoun, and L. Hassouni, "Sentiment analysis of imbalanced datasets using BERT and ensemble stacking for deep learning," *Engineering Applications of Artificial Intelligence,* vol. 126, p. 106999, 2023/11/01/ 2023, doi: https://doi.org/10.1016/j.engappai.2023.106999.

[97] E. Mouaad, M. A. Ouassil, R. Rachidi, B. Cherradi, S. Hamida, and A. Raihani, "Sentiment Analysis on Moroccan Dialect based on ML and Social Media Content Detection," *International Journal of Advanced Computer Science and Applications,* vol. 14, pp. 315-325, 04/01 2023, doi: 10.14569/IJACSA.2023.0140347.

[98] N. Z. Alhazzani, I. M. Al-Turaiki, and S. A. Alkhodair, "Text Classification of Patient Experience Comments in Saudi Dialect Using Deep Learning Techniques," *Applied Sciences*, vol. 13, no. 18, p. 10305doi: 10.3390/app131810305.

[99] T. Omran, B. Sharef, C. Grosan, and Y. Li, "The Impact of Data Augmentation on Sentiment Analysis of Translated Textual Data," in *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, 8-9 March 2023 2023, pp. 1-4, doi: 10.1109/ITIKD56332.2023.10099851.

[100] P. Deng, S. Chen, W. Zhang, J. Zhang, and L. Dai, "The USTC's Dialect Speech Translation System for IWSLT 2023," Toronto, Canada (in-person and online), July 2023: Association for Computational Linguistics, in Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pp. 102-112, doi: 10.18653/v1/2023.iwslt-1.5. [Online]. Available: https://aclanthology.org/2023.iwslt-1.5/

https://doi.org/10.18653/v1/2023.iwslt-1.5

[101] K. Y. Zergat, S. A. Selouani, A. Amrouche, Y. Kahil, and T. Merazi-Meksen, "The voice as a material clue: a new forensic Algerian Corpus," *Multimedia Tools and Applications,* vol. 82, no. 19, pp. 29095-29113, 2023/08/01 2023, doi: 10.1007/s11042-023-14412-2.

[102] A. Mekki, I. Zribi, M. Ellouze, and L. H. Belguith, "Tokenization of Tunisian Arabic: A Comparison between Three Machine Learning Models," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.,* vol. 22, no. 7, p. Article 194, 2023, doi: 10.1145/3599234.

[103] H. Haddad *et al.*, "TunBERT: Pretraining BERT for Tunisian Dialect Understanding," *SN Computer Science,* vol. 4, no. 2, p. 194, 2023/02/03 2023, doi: 10.1007/s42979-022-01541-y.

[104] H. Hallawi, H. Ragheb, Z. Abdullah, N. Al-Shakarchy, and D. Al-Nasrawi, "User identification based on short text using recurrent deep learning," *IAES International Journal of Artificial Intelligence (IJ-AI),* vol. 12, pp. 1812-1820, 12/01 2023, doi: 10.11591/ijai.v12.i4.pp1812-1820.

[105] M. A. Almeqren, L. Almuqren, F. Alhayan, A. I. Cristea, and D. Pennington, "Using deep learning to analyze the psychological effects of COVID-19," (in eng), *Front Psychol,* vol. 14, p. 962854, 2023, doi: 10.3389/fpsyg.2023.962854.

[106] A. Ibrahim, A. Hosseini, H. Helmy, W. Lakhdhar, and A. Serag, *Bridging Dialectal Gaps in Arabic Medical LLMs through Model Merging*. 2025, pp. 338-346.

[107] K. Shaalan, S. Siddiqui, M. Alkhatib, and A. Monem, "Challenges in Arabic Natural Language Processing," 2018, pp. 59-83.

[108] N. A. Ghumeid and M. Essgaer, "Addressing the Libyan Arabic Dialect Identification: A Comparative Study of Ensemble Classification Methods," in *2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, 19-21 May 2024 2024, pp. 579-584, doi: 10.1109/MI-STA61267.2024.10599739.

[109] A. Charfi, M. Bessghaier, A. Atalla, R. Akasheh, S. Al-Emadi, and W. Zaghouani, "MARASTA: A Multi-dialectal Arabic Cross-domain Stance Corpus," pp. 11060-11069, May 2024. [Online]. Available: https://aclanthology.org/2024.lrec-main.964/.

[110] W. M. S. Yafooz, "Enhancing Arabic Dialect Detection on Social Media: A Hybrid Model with an Attention Mechanism," *Information*, vol. 15, no. 6, p. 316doi: 10.3390/info15060316.

[111] H. Elgibreen *et al.*, "An Incremental Approach to Corpus Design and Construction: Application to a Large Contemporary Saudi Corpus," *IEEE Access,* vol. 9, pp. 88405-88428, 2021, doi: 10.1109/ACCESS.2021.3089924.

[112] F. Alwajih, G. Bhatia, and M. Abdul-Mageed, *Dallah: A Dialect-Aware Multimodal Large Language Model for Arabic*. 2024, pp. 320-336.

[113] E. Alqulaity, W. Yafooz, A. Alourani, and A. Jaradat, "Arabic Dialect Identification in Social Media: A Comparative Study of Deep Learning and Transformer Approaches," *Intelligent Automation & Soft Computing,* vol. 39, pp. 1-10, 01/01 2024, doi: 10.32604/iasc.2024.055470.

[114] A. A. Alsuwaylimi, "Arabic dialect identification in social media: A hybrid model with transformer models and BiLSTM," *Heliyon,* vol. 10, no. 17, p. e36280, 2024/09/15/ 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e36280.

[115] F. Qarah and T. Alsanoosy, "Evaluation of Arabic Large Language Models on Moroccan Dialect," *Engineering, Technology & Applied Science Research,* vol. 15, no. 3, pp. 22478-22485, 06/04 2025, doi: 10.48084/etasr.10331.

[116] A. Alabdullah, L. Han, and C. Lin, *Advancing Dialectal Arabic to Modern Standard Arabic Machine Translation*. 2025.

[117] H. Elsafty, T. Deußer, M. Pielka, C. Bauckhage, and R. Sifa, "ArDia: Improving Arabic Dialectal Language Classification Using a Novel Dataset," *Proceedings of the International AAAI Conference on Web and Social Media,* vol. 19, pp. 2413-2422, 06/07 2025, doi: 10.1609/icwsm.v19i1.35944.

[118] K. Almeman, "Automated Building of a Multidialectal Parallel Arabic Corpus Using Large Language Models," *Data*, vol. 10, no. 12, p. 208doi: 10.3390/data10120208.
[119] H. Zaidani, R. Koulali, A. Maizate, and M. Ouzzif, "Augmentation and Classification of Requests in Moroccan Dialect to Improve Quality of Public Service: A Comparative Study of Algorithms," *Future Internet*, vol. 17, no. 4, p. 176doi: 10.3390/fi17040176.
[120] S. Zaid, A. H. Alharbi, and H. Samra, "Multi-Aspect Sentiment Classification of Arabic Tourism Reviews Using BERT and Classical Machine Learning," *Data*, vol. 10, no. 11, p. 168doi: 10.3390/data10110168.