Global Journal of Sciences
Volume 3, Issue 1, 2026, Pages 13-25
DOI: https://doi.org/10.48165/gjs.2026.3102; ISSN: 3049-0456
https://acspublisher.com/journals/index.php/gjs

Research Article

Systematic Literature Review of Arabic NLP Datasets: A Meta-Study

Amani Jamal[1,2]

[1]*Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*

[2]*Center of Research Excellence in Artificial Intelligence and Data Science, King Abdulaziz university, Jeddah, Saudi Arabia*

ARTICLE INFO

Article History
Received 02 Mar 2026
Revised 06 Apr 2026
Accepted 18 Apr 2026
Published 12 May 2026

Keywords
Arabic Natural Language Processing;
PRISMA Systematic Review;
Sentiment Analysis;
Dialect Identification;
Dataset Catalogues.

ABSTRACT

This meta-analysis is a systematic review and synthesis of the Arabic Natural Language Processing (NLP) dataset landscape, in accordance with PRISMA guidelines. The review locates and categorizes publicly accessible datasets in NLP tasks that span a spectrum of sentiment analysis, text classification, question answering, summarization, dialect detection, and paraphrasing. Key datasets catalogues like Masader and Masader Plus are underscored as driven by enhancing discoverability and metadata standardisation, whereas task and domain-specific creation of resources are represented by ArabSis, SNAD, A-MASA, AGS and WiHArD. As can be seen, although the amount and variety of the datasets have grown in recent years, there is still a considerable number of gaps in the coverage of dialects, domain specificity and quality of annotations. The review ends with suggestions on how to extend underexploited areas, annotation behavior, and open-access to accelerate the Arabic NLP research and application.

1. INTRODUCTION

The Arabic Natural Language Processing (NLP) has reaped the benefits of increasing awareness regarding the global significance of the Arabic language; as one of the six official languages of the United Nations and the native language of more than 400 million people in the world. The high rate of digital revolution in the Arab world and the increasing need of smart systems that can be able to facilitate communication in the Arabic language has attracted a lot of interest in the creation of computational systems that are Arabic oriented. Social media sentiment analysis, automated question answering, smart chatbots, machine translation, and information retrieval have identified the need to have robust, well-labeled, and varied Arabic datasets [1].

Regardless of these developments, Arabic NLP still has structural and linguistic issues that make it different among most of the other languages [2]. The diglossic characteristic of the language, in which Modern Standard Arabic (MSA) and a variety of dialects co-exist, makes it hard to create resources as well as generalize models. Regional dialectal diversity including the Gulf, North Africa, the Levant, and Egypt add further variability to orthography, phonology, and lexicon, which current data sets frequently cannot fully reflect [3]. Moreover, Arabic morphology with a rich inflection, deriving, and cliticization is

*Corresponding author. Email: atjamal@kau.edu.sa

a significant obstacle to preprocessing, tokenization, and downstream NLP [4]. These peculiarities of the language highlight the necessity of developing datasets that would encompass MSA and at the same time represent dialects and specialized areas appropriately [5].

The absence, fragmentation and low accessibility of datasets is another urgent problem in Arabic NLP. Arabic corpora tend to be fragmented, unexpressed or encrypted by proprietary rights, unlike English and other languages that are more resource rich [6]. The resulting lack disrupts reproducibility, benchmarking of a model, and the creation of generalized solutions [7]. Such projects, including Masader and Masader Plus, have left a significant impact by indexing hundreds of Arabic collections with a standardized metadata, which enhances the process of discoverability and usability [8]. But there are still major gaps, notably in realms like domain-specific corpora (e.g., healthcare, legal, and religious domains), dialogue and conversational data, and materials in underrepresented dialects [9].

Regarding the research methodology, systematic reviews of the available data sets play a crucial role in accumulating the knowledge, revealing gaps, and informing the creation of the new datasets [10]. Although some of the previous scoping reviews and benchmarking projects have surveyed Arabic NLP aspects, there still is a necessity to conduct a thorough PRISMA-directed synthesis that categories the resources based on the tasks, domains, and language varieties and accessibility levels [11]. The given meta-study answers this need, as it examines publicly available Arabic NLP data sets in a systematic manner, emphasizing their features, domain, and shortcomings [12]. The ultimate aim is to give the researchers and practitioners a systematized point of reference when choosing the datasets, not mention give recommendations on how to fill critical gaps in the landscape.

This review provides a map of the current situation with the Arabic NLP resources, as well as points out the strategic ways of future research by synthesizing the contributions of dataset catalogues, benchmarking and task-specific studies. They are development of inclusive dialectal corpora, adoption of uniform annotation standards, and encouragement of open access programs which democratize access to resources. By so doing, the research is aimed at hastening the creation of strong Arabic NLP mechanisms capable of supporting the linguistic, cultural and technological interests of the Arabic speaking nations in the world.

NLP has developed to offer new possibilities in both the acquisition of language and in the analysis of data. A number of benchmark datasets are in English and some other languages, yet Arabic remains underrepresented. This is impressive due to the number of dialects and speakers. The Arabic NLP has witnessed the publication of research over the past few years, giving rise to the development of datasets in sentiment analysis, text classification and question answering, machine translation and dialect identification just to mention a few. Such datasets can be used in academic research as well as in the real world through media monitoring, healthcare, education, religious studies and digital governance. The Arabic NLP data sets are problematic in terms of diversity, access and standardization, which complicates the chances of reproducing and generalizing research findings.

In the natural language processing of Arabic, e.g., we have seen related work on the construction of evaluation benchmarks and datasets. Good examples include Masader and Masader Plus, which have comprehensive catalogs, which increase the discoverability but also the consistency of metadata. Moreover, the fact that ArabSis, SNAD, ArabicaQA, and WiHArD datasets were specialized and refined specifically to take part in sentiment analysis, news classification, question answering, and text classification, respectively, reflects that progress in these fields is noteworthy. Nevertheless, the new advances also have the unresolved issues of dialectal and domain-specific coverage. Issues of the quality of annotations and the accessibility of the resources are also a problem. To consider the advantages and disadvantages of the data sets utilized in the Arabic NLP, this study gathers and appraises the data sets through PRISMA approach. Unlike the research mentioned, this research helps in the compilation of the existing resources, which in turn helps in identifying the shortcomings, which consequently helps in the construction and utilization of the resources in Arabic NLP.

2. PURPOSE AND CONTRIBUTIONS

This meta-study intends to conduct a systematic review of the existing body of Arabic NLP datasets. It aims to explore the data sets' range, strengths, and weaknesses. The study intends to achieve the objectives stated above by categorizing the existing datasets based on domains and use-cases per the PRISMA framework, assessing their accessibility and diversity, and shining a light on gaps such as restricted dialectal and domain coverage, fluctuating annotation quality, and lack of coverage in special domains. The study intends to deliver to the researchers and practitioners a singular reference that encloses the available datasets and the datasets that need to be developed to enrich the coverage and improve the quality of resources. This meta-study is intended to guide researchers in the development of resources to push the boundaries of research and applications of Arabic NLP.

The primary purpose of this review is to systematically catalog, classify, and summarize the landscape of Arabic NLP datasets, highlighting their characteristics, coverage, and accessibility. The main contributions are:

- A PRISMA-guided systematic review and meta-analysis of Arabic NLP datasets.
- A classification framework covering task, domain, language variety, and accessibility.
- Concise summaries of key datasets, including their size, focus, and public availability.
- Identification of research gaps and recommendations for future dataset development.

3. SELECTION METHODOLOGY

A comprehensive search was conducted over 170 million research papers in Consensus, including Semantic Scholar, PubMed, Scopus, and other sources. The search strategy followed PRISMA guidelines and included 20 targeted queries across 8 search groups, focusing on dataset catalogues, task-specific resources, dialectal coverage, annotation standards, and benchmarking. In total, 1038 papers were identified, 625 were screened, 494 were deemed eligible, and the 50 most relevant papers were included in this review. The Figure 1 depicts PRISMA-based selection process for Arabic NLP dataset literature.

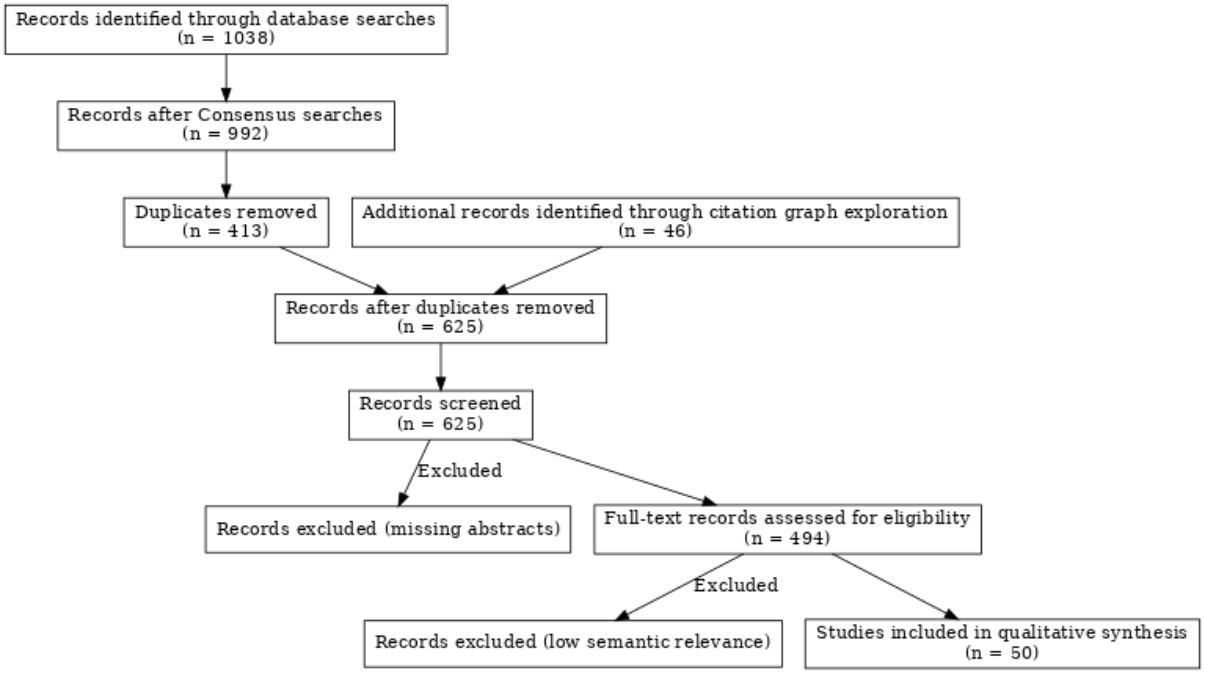


Fig. 1. PRISMA-based selection process for Arabic NLP dataset literature

4. SCOPE

- Publicly available Arabic NLP datasets for text and speech.
- Datasets supporting core NLP tasks: sentiment analysis, classification, question answering, summarization, dialect identification, named entity recognition, paraphrasing, and more.
- Both Modern Standard Arabic (MSA) and dialectal resources.
- Excludes proprietary or non-public datasets, focusing on resources accessible to the research community.

5. RELATED WORK

Several prior surveys and catalogues have mapped the Arabic NLP dataset landscape. Masader [6]and Masader Plus [9] are the largest public catalogues, indexing 200+ datasets with rich metadata. Scoping reviews have identified dozens of free and accessible corpora, but also highlight persistent gaps in dialectal, domain-specific, and complex-task resources [2,5]. Benchmarking initiatives such as ORCA [11], ARLUE [12], and GPTAraEval [13] aggregate and evaluate models on a wide range of datasets, revealing both progress and ongoing limitations in coverage and annotation quality. Recent works also emphasize the need for more comprehensive, well-annotated, and publicly available datasets, especially for underrepresented dialects and domains [1,3,7,8,10].

The largest Arabic NLP dataset catalogues were proposed by Masader [6], which indexes more than 200 datasets, including rich metadata. The work focuses on enhancing discoverability and standardization of datasets among researchers and practitioners. Indexing 500 + datasets with improved discovery interface MASader Plus [9], extension of Masader. It has more comprehensive metadata, and enhances the usability of dataset catalogues, handling the problems of accessibility and navigation.

In MSA SNAD [2] created a news classification dataset consisting of 45,000+ records, which experiments on deep learning confirmed. It shows how big, domain-specific corpus can be used to enhance the performance of classification. ArabSis[1] sentiment analysis dataset is based on five different emotions in universal Arabic. It is the attempt to overcome binary sentiment labels movement to multidimensional emotion classification.

ArabicQA [4] a large Arabic Question Answering dataset of more than 89,000 Q&A pairs in MSA. It aids open-domain QA tasks and is crowd-annotated, and this is essential to evaluate question answering systems. Twt15DA [5] a dialect identification dataset that includes more than 311, 000 tweets and 15 country-level dialects. It shows the dialectal diversity difficulties and gives a standard of categorizing dialect.

The data used in this study is Open Access Dialectal Dataset [8], which gathered more than 50,000 tweets in 5 dialects, and annotated by topic and sentiment. It is created as a multi-task dataset in dialect classification, sentiment analysis, and topic identification. MARSA [7] an emotion collection of 61,000+ tweets in Gulf dialects in various areas. It includes lexicons and points out underserved Gulf Arabic materials.

The SANAD [14] intermediate Arabic news dataset containing approximately 200,000 articles that were labeled under seven classes. It is among the biggest free datasets that are used in text classification in MSA. WiHArD [15] is a hierarchical dataset is a Wikipedia article in MSA to perform text classification tasks. It aims at testing multi-level classification models.

QADI [16] a dataset of 540,000 annotated-to-18-dialect tweets. It supports country-level dialect identification and provides an essential resource for studying dialect diversity. ArEntail [17] a natural language inference dataset of 6,000 manually curated sentence pairs from Arabic news headlines. It introduces entailment and contradiction tasks in Arabic. CAraNER [18] a COVID-19 news dataset with over 55,000 tokens annotated for named entities. It targets domain-specific NER in Arabic. Arabic Paraphrased Parallel Dataset [3] a synthetic dataset generated using back-translation methods for paraphrasing, summarization, and machine translation tasks.

APD [19] a large punctuation prediction dataset with over 312 million words. It focuses on sentence boundary detection and punctuation restoration in Arabic. Fatwaset [20] a dataset of Islamic fatwas designed for NLP applications in religious domains. It is metadata-rich and highlights underexplored domains in Arabic NLP. A-MASA [21] an aspect-based sentiment analysis dataset covering 6,500 records across five domains. It bridges MSA and dialectal resources. AltibbiVec [22] a large-scale medical dataset of 1.5 million consultations used for creating healthcare embeddings. It represents domain-specific applications in medicine.

AraFinNLP / ArBanking77 [23] a financial intent detection dataset with 39,000 queries in MSA and dialects. It supports multi-intent classification for banking-related tasks. GPTAraEval [13] a benchmark for evaluating large language models (LLMs) on Arabic NLP tasks, covering 44 tasks and over 60 datasets. It provides a unified evaluation framework. ORCA [11] a benchmark suite of 60 datasets spanning seven task clusters. It enables comparative evaluation and standardized leaderboards for Arabic NLP systems.

6. CLASSIFICATION OF THE DATASETS

In this review, the datasets are classified based on their task focus, domain focus, and language variations to better understand their range and usage. Masader and Masader Plus act as catalogues that index hundreds of datasets spanning different NLP tasks and dialects to provide holistic overviews. On the other hand, datasets that serve specific tasks are collated under sentiment analysis, news classification, question answering, dialect identification, paraphrasing, and text classification. Moreover, the datasets are distinguished based on their relevance to specific sectors, such as healthcare, municipal services, and religion, with examples including CAraNER, the Municipal Arabic Dataset, and Fatwaset. Further, dialectal datasets such as Twt15DA, QADI, and Open Access Dialectal are noted separately due to their focus on regional and country-level dialects. This organization brings to attention not only the scope of datasets but also the misbalance in coverage with some tasks and domains being abundantly covered while others are lacking critical research.

Datasets are classified by:

- Task: Sentiment analysis, classification, question answering, summarization, discretization, paraphrasing, NER, etc.
- Language Variety: MSA, dialectal, or mixed.
- Domain: News, social media, reviews, Wikipedia, healthcare, religious, etc.
- Size: Ranging from thousands to millions of records.
- Accessibility: Publicly available or restricted.

7. SUMMARY OF EACH DATASET

Diverse and imbalanced among themselves, these reviewed datasets represent the landscape of Arabic NLP resources ranging from catalogues and task-specific corpora to domain-specific collections. Masader and Masader Plus, the comprehensive catalogues aggregating more than 700 datasets, have rich and standardized metadata and discovery tools. Arabic resources for sentiment analysis like ArabSis and

MARSA manifest multidimensional and domain-specific affective annotations, while SNAD and SANAD provide large-scale validated categories of news classification corpora. ArabicaQA denotes open domain question-answering data under a large dataset with crowd-annotated Q&A pairs. On the other hand, public service requests are covered in dialogue-oriented resources, such as the Municipal Arabic Dataset. Twt15DA, Open Access Dialectal, and QADI provide dialectal variation: they combined include data from up to 18 dialects through social media. WiHArD focuses on hierarchical classification, ArEntail focuses on natural language inference, CAraNER focuses on named entity recognition concerning COVID-19 news, Arabic Paraphrased Parallel includes paraphrase and summary, and APD deals with punctuation prediction on large text corpora. Fatwaset indicates the religious angle with an Islamic fatwa corpus example of domain-specific initiatives. These datasets represent significant advances in developing Arabic NLP resources while revealing persistent gaps in dialectal balance, domain-specific depth, and annotation quality.

The summary of each dataset is given in Table 1, 2 and 3.

Table 1 Classification of Major Arabic NLP Datasets

Dataset Name / Paper	Task / Domain	Content	Variety	Availability	Key Notes
Masader/[6]	Catalogue (all)	200 datasets, 25 attributes	MSA & Dialects	Yes	Largest public catalogue; rich metadata
Masader Plus/[9]	Catalogue (all)	500+ datasets indexed	MSA & Dialects	Yes	Enhanced interface for dataset discovery
ArabSis\[1]	Sentiment (5 emotions)	Not specified	Universal Arabic	Yes	Multidimensional sentiment, 5 emotions
SNAD/[2]	News classification	45,000+ records	MSA	Yes	6 news classes, deep learning validated
ArabicaQA/[4]	Question answering	89,095 Q&A pairs	MSA	Yes	Large-scale, open-domain QA, crowd-annotated
Municipal Arabic Dataset/[39]	Municipal services	3,714 requests	MSA	Yes	8 classes, municipal sector focus
Twt15DA/[10]	Dialect identification	311,785 tweets	15 dialects	Yes	Country-level dialects, Twitter-based
Open Access Dialectal/[8]	Dialect, topic, sentiment	50,000+ tweets	5 dialects	Yes	Social media, multi-task labels
MARSA/[7]	Sentiment (multi-domain)	61,353 tweets	Gulf dialects	Yes	4 domains, affect lexicons
SANAD \ [14]	News classification	~200,000 articles	MSA	Yes	7 categories, 3 news portals
WiHArD \[15]	Text classification	Not specified	MSA	Yes	Hierarchical, Wikipedia-based
QADI /[16]	Dialect identification	540,000 tweets	18 dialects	Yes	Country-level, Twitter-based

ArEntail/ [17]	Natural language inference	6,000 pairs	MSA	Yes	Manually curated, news headlines
CAraNER /[18]	Named entity recognition	1,278 sentences, 55,389 tokens	MSA	Yes	COVID-19 news, 5 entity types
Arabic Paraphrased Parallel /[3]	Paraphrasing, MT, summarization	Not specified	MSA	Yes	Synthetic, back-translation
APD / [19]	Punctuation prediction	312M words, 12M sentences	MSA	Yes	Sentence boundary, punctuation
Fatwaset/[20]	Religious / Islamic	Not specified	MSA	Yes	Islamic fatwas, metadata-rich

Table 2 Summary and Classification of Major Arabic NLP Datasets and Catalogues

Dataset Name / Paper	Task / Domain	Content Language	Variety	Public Availability	Key Notes
A-MASA [21]	Aspect-based sentiment	6,500 records, 5 domains	MSA & Dialects	Yes	Multidomain, aspect extraction
AltibbiVec [22]	Healthcare embeddings	1.5M consultations	MSA & Dialects	Yes	Medical, word embeddings
AraFinNLP/ArBanking77 [23]	Financial intent detection	39,000 queries	MSA & 4 dialects	Yes	Banking, multi-intent
GPTAraEval [13]	Benchmarking LLMs	44 tasks, 60+ datasets	MSA & Dialects	Yes	LLM evaluation, NLU/NLG
ORCA [11]	Benchmark suite	60 datasets, 7 task clusters	MSA & Dialects	Yes	Unified evaluation, leaderboard

8. RESULTS

Their results highlight the increasing amount of data in Arabic NLP, particularly in sentiment analysis, news classification and question answering. The data sets are dialect recognition and niche markets like healthcare and religion. And there are still a lot of holes on the dialect, subject matter, and coverage of annotation standards. The results section conducts a thorough overview of datasets and offers a review of the resource landscape. The section also brings out the key points, bottlenecks and research gaps that limit the future vision of Arabic NLP. Findings of this meta-study allow one to get a picture of the current Arabic NLP data. In this summary datasets are classify based on the field of activities, data properties and availability. The overview presents a narrative synthesis of materials based on catalogues and corpora collections based on the PRISMA methodology and denotes the increasing tendency in the volume of information, language diversity, and application scenario.

Table 3 Comparison of Key Arabic NLP Dataset Catalogues and Individual Datasets.

Paper (Year)	Focus / Dataset	Size / Content	Main Application(s)	Availability
[6]	Masader (Catalogue)	200 datasets	Metadata, discovery	Yes
[9]	Masader Plus (Catalogue)	500+ datasets	Metadata, discovery	Yes
[1]	ArabSis (Sentiment)	5 emotions, size not specified	Multi-dimensional sentiment	Yes
[2]	SNAD (News)	45,000+ records	News classification	Yes
[4]	ArabicaQA (QA)	89,095 Q&A pairs	Question answering	Yes

8.1 Dataset Catalogues and Accessibility

- Masader and similar catalogues provide comprehensive listings of Arabic NLP datasets, with Masader indexing 200+ datasets and classifying them by task, domain, and accessibility [5].
- A recent scoping review identified 48 freely accessible Arabic corpora, categorized by primary purpose (e.g., sentiment, QA, dialogue, religious, news) [5].
- Most available datasets focus on Modern Standard Arabic (MSA), with fewer resources for dialects, domainspecific applications, or low-resource tasks [5, 24, 25, 26, 27, 28].

8.2 Task-Specific and Domain-Specific Datasets

- Sentiment Analysis: Numerous datasets exist for sentiment analysis, especially for social media and product reviews, but aspect-based and dialectal sentiment datasets remain limited [27, 28, 29, 30, 31, 32].
- Dialect Identification: Datasets for dialect identification and processing are growing, with Egyptian, Saudi, Algerian, and Jordanian dialects most represented; others (e.g., Mauritanian, Yemeni) are underrepresented [25, 26].
- Hate Speech and Offensive Language: Several datasets target hate speech and offensive language detection on Arabic Twitter, with a focus on MSA and Gulf dialects [33].
- Religious and Qur'anic NLP: Specialized datasets exist for Qur'anic research, including morphological analysis, recitation correction, and question answering, but many are not publicly available [24,34].
- Mental Health and Chatbots: Recent work has begun to address mental health applications and chatbot development, but these areas are still in early stages and face data scarcity [5, 35, 36].

8.3 Feature Extraction and Deep Learning

- Word embeddings (Word2Vec, FastText, AraVec, GloVe, MARBERT, AraBERT) are widely used for semantic feature extraction, with transformer-based models (AraBERT, MARBERT) showing superior performance, especially for dialectal and contextual tasks [26, 27, 28, 30, 31, 35, 37].

- Deep learning models (CNN, RNN, LSTM, Bi-LSTM) are increasingly applied, but their effectiveness is often limited by dataset size and annotation quality [27, 28, 30, 31, 38].

8.4 Dataset Characteristics and Gaps

- Size and Coverage: Dataset sizes range from a few thousand to millions of records, but many are small or narrowly focused [5, 24, 25, 26, 27, 28, 29, 30, 32].
- Annotation Quality: Annotation practices and quality vary widely, with some datasets lacking clear guidelines or inter-annotator agreement reporting [5, 27, 28, 30, 35].
- Public Availability: As the quantity of the free datasets increases, many of them are not available because of the licensing or the absence of documentation [5, 24, 34].

9. DISCUSSION

The landscape of Arabic NLP datasets has become much more advanced, with an increasing number of free and available corpora, in particular, sentiment analysis, dialect recognition, and hate speech detection [5, 26, 27, 28, 29, 30, 33] the lack of representation of many dialects and areas, (2) inconsistent annotation, (3) the inadequacy of datasets and variety, and (4) a deficit of standardized benchmarks on many tasks [5, 24, 25, 26, 27, 28, 34, 35]. Transformer-based models (AraBERT, MARBERT) have set new performance standards, but their effectiveness is constrained by the availability and quality of training data [26, 27, 28, 30, 31, 35, 37]. There is a clear need for more comprehensive, well-annotated, and publicly available datasets, especially for underrepresented dialects, domains (e.g., healthcare, legal, religious), and complex tasks (e.g., aspect-based sentiment, emotion recognition, dialogue systems) [32]. It is given in Table 4.

Table 4 Key Claims and Support Evidence Identified in These Papers

Claim	Evidence Strength	Reasoning	Papers
The number and diversity of Arabic NLP datasets have increased in recent years.	Strong	Multiple systematic reviews and catalogues document dozens of new datasets across tasks and domains.	[5, 26, 27, 28, 29, 30, 32]
Most available datasets focus on MSA and a few major dialects; many dialects and domains are underrepresented.	Strong	Reviews consistently report gaps in dialectal and domainspecific resources.	[5, 24, 25, 26, 27, 28, 32]
Annotation quality and documentation vary widely, impacting dataset usability and model performance.	Moderate	Several reviews highlight inconsistent annotation practices and lack of guidelines.	[5, 27, 28, 30, 35]
Transformer-based models outperform traditional methods but require large, high-quality datasets.	Strong	results for AraBERT/MARBERT, but note data limitations.	[26, 27, 28, 30, 31, 35, 37]
Public availability of datasets is improving, but many remain inaccessible or poorly documented.	Moderate	Scoping reviews and catalogues report growth in free datasets, but also access barriers.	[5, 24, 34]

Some tasks (e.g., aspect- recognition, dialogue) lack large, high-quality datasets.	Moderate	NLP based sentiment, emotion literature.	[27, 28, 32, 36]

10. Conclusion

The Arabic NLP data ecosystem has grown in leaps and bounds, with much of this effort being done by catalogues such as Masader and Masader Plus, and an increasing amount of task- and domain-specific datasets. Although strides have been made in the sentimental analysis, classification, and question answering as well as dialect identification, there are still difficulties in dialectal and domain coverage, quality of annotation and accessibility of data. These gaps still require efforts to fill them and make sure that the Arabic language technologies are advanced. Although it is improving, dialectal coverage, domain diversity (e.g., healthcare, legal, religious) and specific tasks of NLP (e.g., aspect-based sentiment, emotion recognition, dialogue) all have significant gaps. The table below emphasizes the allocation of the datasets on the basis of task and domain. The research gap matrix is given in Table 5.

Table 5 Distribution of Arabic NLP datasets by task and domain, highlighting research gaps.

Task / Domain	General	Media	Healthcare	Religious	Dialogue / Chatbot
Sentiment Analysis	7	10	1	GAP	1
Dialect Identification	2	5	GAP	GAP	GAP
Hate Speech / Offensive	1	4	GAP	GAP	GAP
Question Answering	2	1	GAP	2	GAP
Aspect / Emotion / Dialogue	1	1	GAP	GAP	1

Declarations

Conflicts of Interest

The author declares no conflicts of interest.

Funding

This research received no external funding.

Data Availability Statements

The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

REFERENCES

[1]. Doughan, Z., Itani, S., & Itani, S. (2025). ArabSis: Arabic Corpus Sentiment Analysis. *IEEE Access*, 13, 81083–81095. https://doi.org/10.1109/ACCESS.2025.3567755

[2]. Alsaleh, D., Alamir, M., & Marie-Sainte, S. (2020). SNAD Arabic Dataset for Deep Learning. 630–640. https://doi.org/10.1007/978-3-030-55180-3_47

[3]. Al-Shameri, N., & Al-Khalifa, H. (2024). Arabic paraphrased parallel synthetic dataset. *Data in Brief*, 57.

[4]. Abdallah, A., Kasem, M., Abdalla, M., Mahmoud, M., Elkasaby, M., Elbendary, Y., & Jatowt, A. (2024). ArabicaQA: A Comprehensive Dataset for Arabic Question Answering. *Proceedings of the 47th International ACM SIGIR Conference*.

[5]. Ahmed, A., Ali, N., Alzubaidi, M., Zaghouani, W., Abd-Alrazaq, A., & Househ, M. (2022). Arabic chatbot technologies: A scoping review. *Computer Methods and Programs in Biomedicine Update*. https://doi.org/10.1016/j.cmpbup.2022.100057

[6]. Alyafeai, Z., Al-Shaibani, M., Ghaleb, M., & Ahmad, I. (2021). Evaluating various tokenizers for Arabic text classification. *Neural Processing Letters*, 55, 2911–2933. https://doi.org/10.1007/s11063-022-10990-8

[7]. Alowisheq, A., Al-Twairesh, N., Altuwaijri, M., AlMoammar, A., Alsuwailem, A., Albuhairi, T., Alahaideb, W., & Alhumoud, S. (2021). MARSA: Multi-domain Arabic resources for sentiment analysis. *IEEE Access*. https://doi.org/10.1109/ACCESS.2021.3120746

[8]. Boujou, E., Chataoui, H., Mekki, A., Benjelloun, S., Chairi, I., & Berrada, I. (2021). An open access NLP dataset for Arabic dialects: Data collection, labeling, and model construction. *arXiv*.

[9]. Altaher, Y., Fadel, A., Alotaibi, M., Alyazidi, M., Al-Mutairi, M., Aldhbuiub, M., Mosaibah, A., Rezk, A., Alhendi, A., Shal, M., Alghamdi, E., Alshaibani, M., Zakraoui, J., Mohammed, W., Gaanoun, K., Elmadani, K., Ghaleb, M., Tazi, N., Alharbi, R., Masoud, M., & Alyafeai, Z. (2022). Masader Plus: A new interface for exploring 500+ Arabic NLP datasets.

[10]. Althobaiti, M. (2021). Creation of annotated country-level dialectal Arabic resources: An unsupervised approach. *Natural Language Engineering*, 28, 607–648. https://doi.org/10.1017/s135132492100019x

[11]. Elmadany, A., Nagoudi, E., & Abdul-Mageed, M. (2022). ORCA: A challenging benchmark for Arabic language understanding. *arXiv*. https://doi.org/10.48550/arXiv.2212.10758

[12]. Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. (2020). ARBERT & MARBERT: Deep bidirectional transformers for Arabic.

[13]. Khondaker, M., Waheed, A., Nagoudi, E., & Abdul-Mageed, M. (2023). GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. *arXiv*. https://doi.org/10.48550/arXiv.2305.14976

[14]. Einea, O., Elnagar, A., & Debsi, R. (2019). SANAD: Single-label Arabic news articles dataset for automatic text categorization. *Data in Brief*, 25. https://doi.org/10.1016/j.dib.2019.104076

[15]. Bouchiha, D., Bouziane, A., Doumi, N., Berbouchi, F., Kebir, A., Mebarki, N., & Benameur, B. (2024). WiHArD: Wikipedia based hierarchical Arabic dataset for text classification. *Proceedings of the 4th International Conference on Embedded & Distributed Systems (EDiS)*, 115–118. https://doi.org/10.1109/EDiS63605.2024.10783418

[16]. Abdelali, A., Mubarak, H., Samih, Y., Hassan, S., & Darwish, K. (2020). QADI: Arabic dialect identification in the wild.

[17]. Obeidat, R., Al-Harahsheh, Y., Al-Ayyoub, M., & Gharaibeh, M. (2024). ArEntail: Manually-curated Arabic natural language inference dataset from news headlines. *Language Resources and Evaluation*, 59, 509–535. https://doi.org/10.1007/s10579-024-09731-1

[18]. Al-Thubaity, A., Alkhereyf, S., Al-Zahrani, W., & Bahanshal, A. (2022). CAraNER: The COVID-19 Arabic named entity corpus. *Proceedings of WANLP*. https://doi.org/10.18653/v1/2022.wanlp-1.1

[19]. Yagi, S., Elnagar, A., & Yaghi, E. (2024). Arabic punctuation dataset. *Data in Brief*, 53. https://doi.org/10.1016/j.dib.2024.110118
[20]. Alyemny, O., Al-Khalifa, H., & Mirza, A. (2023). A data-driven exploration of a new Islamic Fatwas dataset for Arabic NLP tasks. *Data*, 8, 155. https://doi.org/10.3390/data8100155
[21]. Eid, Y., Zayed, H., & Medhat, W. (2024). A-MASA: Arabic multi-domain aspect-based sentiment analysis datasets. *Procedia Computer Science*, 202–211. https://doi.org/10.1016/j.procs.2024.10.193
[22]. Habib, M., Faris, M., Alomari, A., & Faris, H. (2021). AltibbiVec: A word embedding model for medical and health applications in the Arabic language. *IEEE Access*, 9, 133875–133888. https://doi.org/10.1109/ACCESS.2021.3115617
[23]. Malaysha, S., El-Haj, M., Ezzini, S., Khalilia, M., Jarrar, M., Almujaiwel, S., Berrada, I., & Bouamor, H. (2024). AraFinNLP 2024: The first Arabic financial NLP shared task.
[24]. Bashir, M., Azmi, A., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A., & Qadir, J. (2021). Arabic natural language processing for Qur'anic research: A systematic review. *Artificial Intelligence Review*, 56, 6801–6854. https://doi.org/10.1007/s10462-022-10313-2
[25]. Hejazi, H., & Khamees, A. (2022). Opinion mining for Arabic dialect in social media data fusion platforms: A systematic review. *Fusion: Practice and Applications*. https://doi.org/10.54216/fpa.090101
[26]. Elnagar, A., Yagi, S., Nassif, A., Shahin, I., & Salloum, S. (2021). Systematic literature review of dialectal Arabic: Identification and detection. *IEEE Access*, 9, 31010–31042. https://doi.org/10.1109/ACCESS.2021.3059504
[27]. Obiedat, R., Al-Darras, D., Alzaghoul, E., & Harfoushi, O. (2021). Arabic aspect-based sentiment analysis: A systematic literature review. *IEEE Access*. https://doi.org/10.1109/ACCESS.2021.3127140
[28]. Alyami, S., Alhothali, A., & Jamal, A. (2022). Systematic literature review of Arabic aspect-based sentiment analysis.
[29]. Nelci, J., Salloum, S., & Shaalan, K. (2021). An overview of sentiment analysis with dialectical processing. *Proceedings of the International Conference on Emerging Technologies and Intelligent Systems*. https://doi.org/10.1007/978-3-030-82616-1_1
[30]. Nassif, A., Elnagar, A., Shahin, I., & Henno, S. (2020). Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities. *Applied Soft Computing*, 98, 106836. https://doi.org/10.1016/j.asoc.2020.106836
[31]. Alhumoud, S., & Wazrah, A. (2021). Arabic sentiment analysis using recurrent neural networks: A review. *Artificial Intelligence Review*, 55, 707–748. https://doi.org/10.1007/s10462-021-09989-9
[32]. Matrane, Y., Benabbou, F., & Sael, N. (2023). A systematic literature review of Arabic dialect sentiment analysis. *Journal of King Saud University – Computer and Information Sciences*, 35, 101570. https://doi.org/10.1016/j.jksuci.2023.101570
[33]. Alhazmi, A., Mahmud, R., Idris, N., Abo, M., & Eke, C. (2024). A systematic literature review of hate speech identification on Arabic Twitter data: Research challenges and future directions. *PeerJ Computer Science*, 10. https://doi.org/10.7717/peerj-cs.1966
[34]. Alrayzah, A., Alsolami, F., & Saleh, M. (2023). Challenges and opportunities for Arabic question-answering systems.
[35]. Alasmari, A. (2025). A scoping review of Arabic natural language processing for mental health. *Healthcare*, 13. https://doi.org/10.3390/healthcare13090963

[36]. Ouali, S., & Said, E. (2024). Arabic chatbots challenges and solutions: A systematic literature review. *Iraqi Journal for Computer Science and Mathematics*. https://doi.org/10.52866/ijcsm.2024.05.03.007
[37]. Bourahouat, G., Abourezq, M., & Daoudi, N. (2024). Word embedding as a semantic feature extraction technique in Arabic natural language processing: An overview. *International Arab Journal of Information Technology*, 21, 313–325. https://doi.org/10.34028/21/2/13
[38]. Bouzahir, M., Abdelouahad, A., & Nabil, M. (2022). How far can deep learning improve Arabic part-of-speech tagging? 206–215. https://doi.org/10.1007/978-3-031-06458-6_17
[39]. Alaloye, H., Alkhodre, A., & Nabil, E. (2025). Utilizing NLP to optimize municipal services delivery using a novel municipal Arabic dataset. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2025.0160278