*mathematics*

MDPI

*Article*

# Advancing AI-Driven Linguistic Analysis: Developing and Annotating Comprehensive Arabic Dialect Corpora for Gulf Countries and Saudi Arabia

**Nouf Al-Shenaifi, Aqil M. Azmi * and Manar Hosny**

Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; noalshenaifi@ksu.edu.sa (N.A.-S.); mifawzi@ksu.edu.sa (M.H.)
* Correspondence: aqil@ksu.edu.sa

**Abstract:** This study harnesses the linguistic diversity of Arabic dialects to create two expansive corpora from X (formerly Twitter). The Gulf Arabic Corpus (GAC-6) includes around 1.7 million tweets from six Gulf countries—Saudi Arabia, UAE, Qatar, Oman, Kuwait, and Bahrain—capturing a wide range of linguistic variations. The Saudi Dialect Corpus (SDC-5) comprises 790,000 tweets, offering in-depth insights into five major regional dialects of Saudi Arabia: Hijazi, Najdi, Southern, Northern, and Eastern, reflecting the complex linguistic landscape of the region. Both corpora are thoroughly annotated with dialect-specific seed words and geolocation data, achieving high levels of accuracy, as indicated by Cohen's Kappa scores of 0.78 for GAC-6 and 0.90 for SDC-5. The annotation process leverages AI-driven techniques, including machine learning algorithms for automated dialect recognition and feature extraction, to enhance the granularity and precision of the data. These resources significantly contribute to the field of Arabic dialectology and facilitate the development of AI algorithms for linguistic data analysis, enhancing AI system design and efficiency. The data provided by this research are crucial for advancing AI methodologies, supporting diverse applications in the realm of next-generation AI technologies.

**Keywords:** Arabic dialects; Arabic corpora; Twitter; dialect identification; lexicon

**MSC:** 68T50; 68T37; 62H30; 91B74



**Citation:** Al-Shenaifi, N.; Azmi, A.M.; Hosny, M. Advancing AI-Driven Linguistic Analysis: Developing and Annotating Comprehensive Arabic Dialect Corpora for Gulf Countries and Saudi Arabia. *Mathematics* **2024**, *12*, 3120. https://doi.org/10.3390/math12193120

Academic Editors: Teng Huang, Qiong Wang and Yan Pang

Received: 24 August 2024
Revised: 28 September 2024
Accepted: 30 September 2024
Published: 5 October 2024



## 1. Introduction

Arabic is not only a major world language, spoken natively by approximately three hundred million people primarily in the Middle East and North Africa (MENA), but also the liturgical language of two billion Muslims globally. It features one of the most widely used writing systems in the world. This script transcends its native speakers, extending throughout the Islamic world, as it is employed to write the Qur'an, the holy book of Muslims [1].

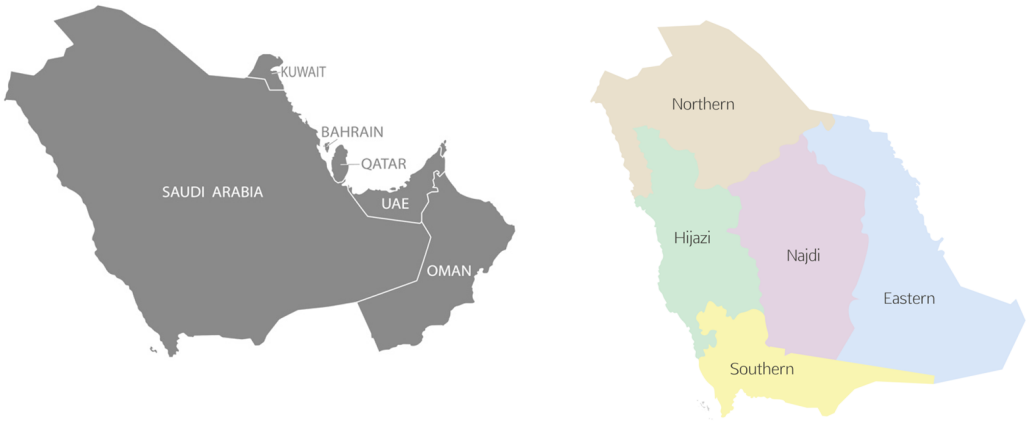
In the contemporary landscape of the Arab region, there has been a marked escalation in the deployment of Arabic dialects for informal written communication and interactions on social media platforms [2–6]. This trend has led to an exponential growth in the quantity of Dialectal Arabic content across these digital venues, particularly on social media, as evidenced by recent studies [7–10]. Consequently, this surge has catalyzed considerable scholarly interest among researchers in Arabic Natural Language Processing (NLP). There is a robust initiative underway to cultivate annotated linguistic resources specifically designed for these dialects. The primary objectives of this initiative are to enrich our understanding of the linguistic intricacies of Dialectal Arabic (DA) [11] and to accelerate the advancement of specialized tools and applications for its processing [12,13].

Despite the broad spectrum of tools and resources dedicated to Arabic NLP, the primary emphasis remains on Modern Standard Arabic (MSA), the lingua franca language

universally utilized throughout the Arab world, which is deeply rooted in Classical Arabic [14–23]. However, the adaptation of these NLP resources to DA introduces significant challenges. These challenges stem from the substantial linguistic divergences between the various dialects and MSA [24,25]. This situation highlights a critical gap in the applicability of existing tools to the linguistic realities of the Arab region, underlining the need for targeted research and development efforts in the field of NLP to bridge these disparities.

Arabic dialects demonstrate significant divergences from MSA in several linguistic domains, including morphology, phonology, and syntax [4,26,27]. Furthermore, the lexicon across various Arabic dialects varies considerably, and most dialects do not employ standardized orthographies [6,28]. While the development of resources for NLP tasks in DA remains nascent in comparison to MSA [29,30], there are ongoing initiatives to construct dialect-specific tools and resources. These efforts include the creation of annotated corpora and morphological analyzers tailored to particular dialects [31]. Notably, the development of NLP tools for the Egyptian and Levantine dialects has progressed more substantially [32–34]. Despite the extensive presence of dialectal content online, Gulf Arabic still experiences a significant deficiency in NLP tools and resources, highlighting a critical area for further linguistic research [4,6,35,36].

Arabic dialects are predominantly classified into several principal groups based on geographical regions, including Gulf, Egyptian, Levantine, North African (Maghrebi), Iraqi, Yemeni, and Sudanese dialects [37–40]. The Gulf dialect comprises the linguistic variants spoken in countries adjacent to the Arabian Gulf, such as Saudi Arabia, Kuwait, Qatar, United Arab Emirates (UAE), Bahrain, and Oman, see Figure 1a. The Egyptian dialect encompasses the linguistic varieties found primarily in Egypt and select areas of Sudan. The Levantine dialect is primarily spoken in the Levant region, including Palestine, Syria, Lebanon, and Jordan. The Maghrebi dialect includes a range of dialects spoken across North Africa, excluding Egypt, and covers countries such as Morocco, Algeria, Tunisia, and Libya. The Iraqi dialect refers to the linguistic variety prevalent in Iraq [41], and the Yemeni dialect is used in Yemen. In Saudi Arabia, a country characterized by its substantial geographical diversity, there is a significant variation in linguistic dialects across different regions; see Figure 1b [4,42]. Historically, there has been a tendency to not recognize the Saudi dialect as a distinct linguistic entity, typically categorizing it within the broader Gulf dialects [43].




(**a**) Arabian Gulf countries (**b**) Saudi Arabia

**Figure 1.** The regional dialectal map displays: (**a**) the six countries of the Arabian Gulf, and (**b**) the distinct dialectal subregions of Saudi Arabia.

The profound social and political transformations occurring in the Gulf region, particularly in Saudi Arabia, underscore the need for a dedicated Saudi Dialect Corpus. Saudi Arabia, characterized by its conservative and autocratic nature, is home to a predominantly young population, with about 50% under the age of 25. The transformative government-

sponsored reforms have sparked significant debates within the society, prominently featured on social media platforms such as Twitter due to restrictions on public debate and protests [10].

The need for a dedicated corpus is exemplified by the intense discussions on Twitter that followed the Saudi government's decision to lift the driving ban for women in September 2017. This landmark policy shift ignited widespread debates and garnered support, as documented in [44]. Researchers gathered and analyzed tweets and hashtags in the local Saudi dialect from the initial days after the ban was lifted, providing valuable insights into public sentiment on this crucial issue. This work underscores the necessity of a specialized corpus that accurately captures the unique linguistic nuances of the Saudi dialect, which is vital for precise sentiment analysis and in-depth cultural research.

A proper Saudi Dialect Corpus would thus not only facilitate a deeper understanding of the public discourse and sentiment in the region but also enhance the development of tailored NLP tools. These tools are essential for accurately interpreting and responding to the nuanced language used in social media, which is often imbued with cultural and regional specificities not covered by standard Arabic corpora. This makes the development and availability of a dedicated Saudi Dialect Corpus crucial for researchers working on Arabic NLP, especially in applications involving sentiment analysis, social monitoring, and cultural studies.

In Saudi Arabia, the dialectical landscape is distinguished by the prevalence of specific regional dialects: Hijazi in the western region, Najdi in the central region, Southern dialect in the southern territories, Northern dialect in the northern areas, and Eastern dialect in the eastern part of the Arabian peninsula. This research paper outlines our efforts in constructing an extensively annotated corpus of Gulf dialectical Arabic. We utilized data sourced from X, formerly known as Twitter (throughout this paper, we will continue to use the terms "Twitter" and "tweets", as there are no widely accepted substitutes for the latter). The data extraction process was improved by incorporating advanced techniques for textual analysis, notably feature extraction. This approach leveraged geographical metadata embedded within user profiles to ensure the accurate representation of diverse dialectical variations across the specified regions. This methodology enhances the precision and robustness of our data processing and annotation efforts, contributing significantly to the depth and quality of linguistic analysis. These techniques are crucial for advancing AI-driven frameworks, optimizing data utilization, and enhancing the reliability of AI applications across various sectors.

This study introduces two pivotal resources: the Gulf Arabic Corpus (GAC-6) and the Saudi Dialect Corpus (SDC-5). The GAC-6 is an expansive corpus comprising approximately 1.7 million Arabic tweets that reflect the dialectical nuances of the Gulf region, encompassing Saudi Arabia, Bahrain, Kuwait, Oman, Qatar, and the Emirates. Concurrently, the SDC-5 emerges as a comprehensive annotated corpus dedicated to Saudi dialects, encapsulating linguistic variations across five key Saudi regions: Hijazi, Najdi, Southern, Northern, and Eastern. Our methodology includes the meticulous manual annotation of a data subset with specific dialect labels, establishing a gold standard for dialect identification. The resources elucidated in this paper hold significant potential not only for advancing NLP applications such as machine translation but also for facilitating in-depth linguistic analysis of Gulf and Saudi dialects. This contributes to a broader understanding of regional linguistic diversity, aiding both academic research and practical applications in computational linguistics.

The objective of this research is to augment the domain of language resources by developing comprehensive linguistic datasets for both Gulf and Saudi dialects. This initiative leverages the substantial repository of Arabic textual content available on social media platforms, with a specific focus on Twitter. The contributions of this paper are manifold and include the following:

- The development of an extensive corpora of Gulf and Saudi dialects, which are sourced from Twitter and automatically tagged with dialect labels. These corpora are designed

to significantly support and enable a wide range of Arabic NLP research endeavors in the future.

- The employment of native speakers for the manual annotation of the segments of the datasets, thereby validating the accuracy of the automatically assigned dialect labels. This step ensures that the linguistic characteristics of each dialect are accurately captured.
- The evaluation of the corpora's quality through the measurement of inter-annotator agreement, quantified using the Kappa statistic. This assessment ensures the reliability and consistency of the annotation process, confirming that the data are robust for academic and practical applications in computational linguistics.

The structure of this paper is organized as follows: Section 2 reviews the relevant literature and related studies. Section 3 details the methodology used in compiling, preprocessing, and annotating the corpora. Section 4 provides an in-depth discussion of the compiled corpora, including general statistics, and assesses the quality of the annotations. Finally, Section 5 offers concluding remarks and outlines potential directions for future research.

## 2. Related Work

This section provides an overview of the literature pertaining to Arabic dialectal corpora, emphasizing the significant contributions and key findings within this research domain.

The exploration of Arabic dialects has garnered considerable interest in recent years. A myriad of Arabic dialectal corpora has been developed, serving as invaluable resources for the study and linguistic analysis of these dialects. In recent times, there has been an augmented effort toward the aggregation of datasets and the formulation of comprehensive corpora from a variety of sources, thereby facilitating in-depth investigations into Arabic dialectology.

A seminal piece of research in the realm of corpus development was conducted by [45], who introduced a pioneering dataset dedicated to Dialectal Arabic, known as the Arabic Online Commentary (AOC) Dataset. The AOC Dataset represents the inaugural dialectal corpus made available to the academic community, comprising approximately 52 million words derived from the comment sections of online Arabic news platforms.

Several studies, such as those documented by [46], have utilized datasets that were manually annotated through crowdsourcing efforts. In particular, Alsarsour et al. [46] introduced the Dialectal Arabic Tweets (DART) dataset, a collection comprising approximately 25,000 Arabic tweets that were manually annotated via crowdsourcing. This dataset is characterized by its balanced representation across five major Arabic dialect groups: Egyptian, Maghrebi, Levantine, Gulf, and Iraqi. Furthermore, Zaghouani and Charfi [47] developed a multidialectal corpus, also annotated manually through crowdsourcing, which includes around 2.4 million tweets originating from 11 Arab regions, including North Levant, South Levant, Egypt, Gulf, Morocco, Tunisia, Algeria, Yemen, Iraq, Libya, and Sudan. Sadat et al. [48] contributed a sentence-level manually annotated dataset containing about 62,000 sentences sourced from online blogs across various Arab nations.

Khalifa et al. [33] compiled a substantial Gulf Arabic Corpus, consisting of 1200 forum novels, with annotations at the document level derived from the novels' titles and authors' names. Alshutayri and Atwell [49] outlined the methodology for constructing an Arabic dialect text corpus from social media platforms like Twitter and Facebook, as well as comments from newspapers, with each entry receiving a dialect tag through crowdsourcing. Some scholars have also embarked on creating parallel corpora, which consist of sentences translated into Arabic dialects from other datasets. An exemplar of such a corpus is presented in [50], where the authors unveiled the Multidialectal Parallel Corpus of Arabic (MPCA), a compilation of approximately 2000 sentences that were manually translated into various dialects from Egyptian Arabic.

Moreover, Bouamor et al. [2] introduced the Multi Arabic Dialect Applications and Resources (MADAR) corpus, a parallel corpus encompassing 25 dialects from Arabic-

speaking cities across 15 Arab countries. This corpus was developed through the manual translation of selected sentences from the Basic Travel Expression Corpus (BTEC) [51]. Notably, the corpus lacks representation of certain Gulf Arabic dialects, specifically those from Bahrain, Kuwait, and the United Arab Emirates.

Other studies, such as the next one, have employed a semi-automatic approach for corpus annotation. Mubarak and Darwish [52] developed a multidialectal Twitter corpus of Arabic, consisting of 6.5 million tweets. This corpus is annotated based on distinct dialects and the geographical locations of users. Similarly, Abu Kwaik et al. [24] introduced the Shami Dialects Corpus (SDC), which encompasses four dialects (Palestinian, Jordanian, Lebanese, and Syrian) and contains approximately 118,000 sentences. This corpus is compiled from Arabic tweets, with automatic annotations derived from the Twitter API's geographical location feature, supplemented by manual annotations for web content.

Furthermore, there are datasets that have been collected and labeled entirely through automated processes. Abdul-Mageed et al. [3] unveiled a vast corpus of tweets representing city-level dialects from 29 Arab cities across 10 Arab countries, with diverse dialectical features. The annotation for this corpus was conducted automatically using a Python geocoding library. Additionally, Abdelali et al. [25] created a balanced, non-genre-specific, country-level Arabic dialectal tweet corpus. This corpus was generated using a series of filters; user accounts were selected based on country-specific keywords, and tweets were further filtered to exclude users predominantly using MSA. The final corpus comprises 540,000 tweets from 2525 Twitter users, along with a test set consisting of 182 tweets per country, which were manually classified by native Arabic speakers.

Conversely, certain dialectal corpora have concentrated on specific dialects, with a focus on morphological annotation. An exemplar is the work by [53], who introduced the Saudi corpus for NLP Applications and Resources (SUAR), targeting the Saudi dialect. This corpus encompasses 104,000 words sourced from various online social media platforms, and it underwent morphological annotation via the MADAMIRA tool [54]. This initial automated annotation was subsequently subjected to manual review to ensure accuracy and validate the analysis.

Alowisheq et al. [55] unveiled the Multi-domain Arabic Resources for Sentiment Analysis (MARSA), a sentiment-annotated corpus specific to the Gulf dialect. This corpus comprises 61,000 tweets, each manually annotated with sentiment labels by two independent annotators, ensuring the reliability of sentiment assessment.

Further, Elgibreen et al. [56] introduced the King Saud University Saudi Corpus (KSUSC), a comprehensive new corpus containing over 161 million sentences harvested from a variety of sources. While this corpus is extensive and spans multiple domains, it lacks annotations, presenting a vast but unstructured resource for linguistic analysis. This study also conducted a review of existing Arabic corpora, highlighting the need for more in-depth research into corpora representing the Saudi dialect.

Moreover, Alruily [57] developed a dialectal Saudi Twitter corpus and provided an analysis of its linguistic peculiarities, such as compounding, abbreviation, spelling discrepancies, and the emergence of neologisms, shedding light on the unique challenges associated with processing this dialect. Lastly, Al-Ghadir and Azmi [58] capitalized on the vibrant social media environment of Saudi Arabia to investigate the posting patterns of local users, delineating these behaviors by gender and educational attainment. Concentrating on author profiling in this milieu, this study provides an understanding of demographic trends, thereby enhancing the comprehension of dialectical subtleties as manifested on social media platforms.

## 3. Our Methodology

Social media platforms and microblogging sites, notably Twitter, have emerged as significant repositories of natural language textual data, offering a rich vein of content for research purposes [59,60]. Twitter, in particular, is integrated into the daily routines of millions, positioning it as one of the foremost social media networks currently in op-

eration [61]. The platform facilitates the aggregation of a substantial volume of text, contributed by a diverse array of Arab speakers who often express their thoughts and opinions in their respective Arabic dialects [62].

Tweets are typically short and informal, frequently composed in the users' own dialects, and reveal a plethora of spoken language features. This makes them an invaluable resource for the study of Arabic dialects, offering insights into the linguistic nuances and vernacular expressions prevalent within the Arab-speaking community [63,64].

Our goal is to develop two distinct corpora: the first, GAC-6, targets the dialects commonly spoken in the Arabian Gulf, encompassing Saudi Arabia, Bahrain, Kuwait, Oman, Qatar, and the UAE. The second, SDC-5, focuses on the five principal dialects within Saudi Arabia—Hijazi, Najdi, Southern, Northern, and Eastern. Both corpora are compiled from data collected from Twitter.

In the development of the GAC-6 and the SDC-5, we maintained stringent ethical standards and emphasized privacy protection throughout the data collection phase. All data used in this study were derived from publicly available Twitter posts, adhering to Twitter's data usage policies. We collected no private information, such as direct messages or personal identifiers, beyond what is publicly visible on user profiles. To enhance privacy protection, all usernames and profile details were anonymized and omitted from the dataset. Additionally, the data were aggregated to eliminate the possibility of tracing back to individual users. We also took precautions to ensure that no sensitive data, which could potentially disclose the identities of individuals or communities, were gathered.

Next, we will explore the details of the corpora development process. This includes discussions on compilation and preprocessing techniques, annotation methodologies, and the protocols used for corpus validation, ensuring a comprehensive understanding of the corpora's foundational integrity.

### *3.1. Compilation and Preprocessing of the Corpora*

Twitter provides an excellent Application Programming Interface (API) and development platform, featuring a comprehensive suite of tools and meticulously crafted documentation. This framework enables researchers to tap into the vast reservoir of social media content and extract additional metadata, such as geographical information [65,66]. Consequently, Twitter was chosen as the primary data source for the construction of the envisaged corpus. The collection of data was facilitated by the Twitter REST APIs, which are accessible through Twitter user credentials via Open Authentication (OAuth) [67]. Leveraging the Twitter API streaming library "Tweepy", a Python library dedicated to tweet retrieval, we amassed thousands of tweets characterized by the use of dialectal expressions typical of Gulf Arabic and Saudi Dialect speakers [68,69]. Our methodology in this study was predicated on the extraction of dialect-specific tweets through the application of filters based on seed words pertinent to each dialect. This approach circumvents the necessity of sifting through an extensive volume of Arabic tweets and subsequently undertaking a labor-intensive annotation process. Algorithm 1 outlines our approach for the automated construction of an annotated dialectal corpus.

The initial step in our corpus construction involved the identification of seed words for each dialect. Seed words are defined as terms that are frequently and predominantly used within a particular dialect and not found in others [37,46]. We focused on unique expressions characteristic of each dialect, used exclusively by its native speakers, to compile the corpus. These dialect-specific terms were employed as query parameters, coupled with the "lang = ar" filter to confine the search to Arabic tweets, facilitating the collection of a real-time tweet stream from speakers of the targeted dialects [49]. Each tweet was then annotated with the user's name and location.

To enhance the reliability of the dialect attribution, we incorporated geographical information from the Twitter profiles associated with each tweet, ensuring that the tweets originated from within the designated dialectal regions. The corpus was further expanded by aggregating additional tweets from each user's profile. We compiled a list of user pro-

files pertinent to each of the six Gulf dialect regions and the five Saudi dialects by identifying tweets containing specific seed words and dialectal expressions exclusive to a single dialect. Only users with a minimum of 1000 tweets were considered, from which we downloaded up to 500 tweets per account. During the collection process, duplicate tweets and retweets were excluded to maintain the corpus's uniqueness.

**Algorithm 1:** A high-level algorithm for the automatic construction of the proposed corpus. During construction, we rely on four components of a tweet: the text, author, timestamp, and location.

**Output:** Collected tweets in corpus $C$ in CSV format.

**begin**
  Initialize $DL$ (dialects list), $SL$ (seed words list), and $LL$ (locations list)
  Identify list of dialects saving it into $DL$
  **foreach** $d \in DL$ **do**
    Identify list of seed words and locations saving into $SL[d]$ and $LL[d]$, respectively
  **end**
  **foreach** $d \in DL$ **do**
    **foreach** dialectal term $s \in SL[d]$ **do**
      Retrieve tweets $T$ containing $s$
      Ensure location of each $t \in T$ belongs to $LL[d]$; otherwise, remove it from $T$
      Label tweets with dialect $d$
      Save tweets into corpus $C$
    **end**
  **end**
  Expand corpus $C$ by collecting 500 tweets from each author
  Filter out retweets and duplicates in $C$
  Preprocess tweets in $C$
**end**

The dialect classification methodology was twofold: initially, it relied on the presence of dialect-specific terms within the tweets, and subsequently, it required a match between the user's profile location and the designated dialect region. The general architecture for our corpus construction is presented in Figure 2.

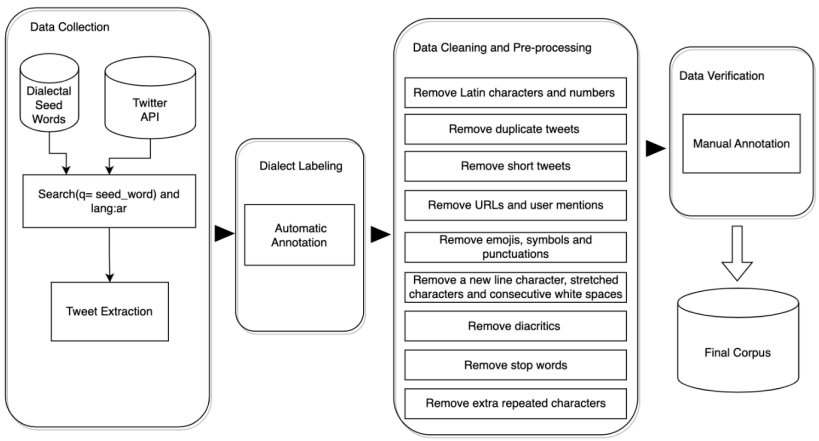



**Figure 2.** The general architecture of the proposed approach.

Our methodology was fundamentally anchored in the utilization of dialect-specific terminology. For instance, the term "watermelon" is articulated as جح in the Saudi dialect and as رقي in the Kuwaiti dialect. Employing this strategy enabled us to refine our tweet selection, ensuring a focus on the specific dialects of the Gulf region and Saudi Arabia. The compilation of seed words and dialectal expressions for each Saudi dialect was compiled from https://lahajat.blogspot.com/p/blog-page_7.html (accessed on 9 June 2023). Conversely, the compilation of seed words for the Gulf dialects, encompassing Kuwait, Bahrain, Qatar, the United Arab Emirates, Saudi Arabia, and Oman, was derived from Mo3jam (https://en.mo3jam.com/, a dictionary of colloquial Arabic/Arabic slang, accessed on 24 October 2023). A selection of these seed words from each Gulf dialect, accompanied by representative tweet examples, is listed in Table 1. Furthermore, Table 2 presents illustrative tweets in the Saudi dialect, demonstrating the application of dialect-specific terms in authentic social media discourse.

Given the noise and extraneous information typical of data sourced from social media, the processes of data cleaning and preprocessing are crucial [53,70]. The primary objective of preprocessing the extracted data was to cleanse it of noise and irrelevant content, thereby enhancing the dataset's quality for more precise dialect identification tasks [46].

The initial phase of preprocessing involved manually discarding tweets we identified as advertisements. Subsequent to this manual filtering, the tweets were subjected to a systematic preprocessing regimen, as shown in Algorithm 2 [71]. This algorithmic approach to preprocessing facilitated the refinement of the dataset, culminating in a polished corpus. Consequently, the Gulf Arabic Corpus was reduced to an approximate total of 1.7 million tweets, while the Saudi Dialect Corpus was cut to about 790,000 tweets, thereby ensuring a cleaner, more focused dataset for subsequent analysis.

**Table 1.** Sample seed words for each dialect in the Arabian Gulf are provided alongside a sample tweet from that specific dialect. Seed words are distinctive terms exclusive to a particular dialect. For convenience, an English translation (tr) is also included.

| | Seed Words | | Sample Tweet | |
|---|---|---|---|---|
| **Region** | **Arabic** | **English (tr)** | **Arabic** | **English (tr)** |
| Saudi Arabia | ارتعت، تكفون | I was horrified, please | احتاج عصفر ارتعت روعه كان قلبي بيوقف تعالو اقرو علي تكفون | I need a yellow safflower. I was horrified. My heart was stopping. Come and tell me that you will be satisfied. |
| Bahrain | سهده، صج، رزنامة، اليمعة | Quiet, true, calendar, Friday | الظاهر رزنامه متأخره عندنا اليوم الخميس وعندهم يوم اليمعه صج سهده الشوارع | Apparently, we have a late calendar. Today is Thursday and there are, on Friday, truly quiet streets. |
| Kuwait | ماكو، تنطرون | Nothing, waiting for | صدقوني ماكو احلى من انكم تنطرون الشروق عالبحر | Believe me, there is nothing better than waiting for the sunrise over the sea. |
| Oman | احسش، جزاتش، سح | I feel, you deserve, dates (fruit) | حبيبتي احسش قفلتي على قلبي بقفل وجزاتش اهديش سطل سح | My dear, I feel like you have locked my heart with a lock and you deserve to be provided with a bucket of dates. |
| Qatar | دلاغ، جوتي | Sock, shoes | ستايل ولدي اليوم ذكرني بالتعاون بين غوتشي وأديداس يعني ما يصير تلبس دلاغ طويل أزرق وفيه رسمة سيارات ملونة والجوتي | My son's style today reminded me of the collaboration between Gucci and Adidas. I mean, you can't wear a long blue sock that has a drawing of colorful cars, and those shoes. |
| Emirates | رقاد، ما اروم | Sleepy, can't | الله ياخذ الرياضه سويت رياضه وللحين منسدحه فيني رقاد ما اروم اتحرك ادرس | May God take away exercise. I did exercise, and for now, I am lying down, sleepy, and can't move or study. |

**Table 2.** Sample seed words accompanied by a tweet example from regional Saudi dialects. We also included an English translation.

| | Seed Words | | Sample Tweet | |
|---|---|---|---|---|
| **Region** | **Arabic** | **English (tr)** | **Arabic** | **English (tr)** |
| Hijazi | التاني، هديك، اشبك | The other, that, what is going on | ابدا مافي تفاهم كل واحد يعطي التاني هديك النظرة اللي اشبك | There is absolutely no understanding. Each one gives the other that look. That is what is going on. |
| Najdi | تسان علمتين طملة، واخزياه | Why don't, tell me, wet, shame | رحت أمر أختي بالكلية وأنا عند الباب قالت ترى معي بنت عمي طيب تسان علمتين قبل والله ي السيارة طملة من المطر واخزياه | I went to pick up my sister at the college. While I was at the door, she said "See, I'm with my cousin". So why don't you tell me before? I swear that the car is wet from the rain and shame. |
| Southern | يطعني، عنش | Protect me, About you | يابعد الدنيا يسوير الله يطعني عنش ياروحي | My darling swayer, may God protect me from you, my dear. |
| Northen | شنوح | Why | والله زعلانه شنوح ما فزنا احسب بناخذ كاس العالم | I swear I'm upset because why didn't we win? I thought we'd take the World Cup. |
| Eastern | هطف | Dumb | منظر يغيض ويغبن كل هطف وهطفه | A sight that enrages and belittles every little person. |

**Algorithm 2:** Preprocessing the tweets

**Input:** Raw tweets $t$
**Output:** Cleaned tweets $t'$

**begin**
- Discard short tweets (those with single words, e.g., نعم "yes")
- Filter out retweets
- Exclude duplicate tweets
- Exclude tweets that contain non-Arabic words
- Remove all stop-words from $t$
- Filter out each word $w$ in tweets such as numerals, diacritical marks, emojis, symbols including newline characters, punctuation, URL links, hashtags, and user mentions (i.e.,@Username)
- Filter out stretched characters (kashida) and consecutive white spaces
- Normalize speech effects in $t$ (e.g., "Helloo", "Helloooo", etc., are mapped to "Hello")

**end**

### *3.2. Annotating the Corpus*

In this section, we describe the methodology adopted for the annotation of the corpus, detailing the procedural steps and tools employed for annotating each tweet within the corpus, followed by the strategies implemented to assess the quality of the annotation task.

The genesis of a corpus extends beyond mere data accumulation; it encompasses a rigorous process of data verification and validation to ensure the corpus's reliability and applicability [72–74]. The primary goal of this endeavor is to forge a dialectal corpus of tweets, distinguished by high-quality annotations, to serve as a resource for scholars engaged in the study of Arabic dialects within the domain of Arabic NLP. This includes, but is not limited to, investigations pertaining to Arabic dialect identification.

The success of dialect identification is intrinsically linked to the precision of annotation outcomes [75]. In order to evaluate the quality of our corpus, constructed and annotated via the proposed algorithm, we conducted a manual annotation exercise. From each

region represented in the Arabian Gulf corpus, we randomly selected 2000 tweets, resulting in a total of 12,000 tweets. Similarly, for the Saudi Corpus, a total of 10,000 tweets were chosen for manual annotation. For this task, we utilized Label Studio, (https://labelstud.io/, accessed on 28 September 2023), an open-source data labeling platform renowned for its versatility in annotating, labeling, and preparing diverse datasets.

Within the Label Studio environment, we established a project dedicated to the annotation of tweets. Annotators were presented with tweets and instructed to assign a dialect label to each one. For the GAC-6, the tweets were categorized into one of six dialect labels corresponding to the Gulf regions: Saudi Arabia, Bahrain, Kuwait, Oman, Qatar, and UAE. In the case of the SDC-5, annotators classified each tweet under one of the five designated Saudi labels: Hejazi, Najdi, Southern, Northern, and Eastern. This meticulous process of manual annotation serves as a cornerstone for ensuring the integrity and utility of the corpus for research in Arabic dialect identification and other NLP applications.

Almuzaini and Azmi [76] employed crowdsourcing to annotate a large volume of data in MSA. However, they encountered quality issues and lapses due to incompetent or dishonest annotators, particularly when the tasks were poorly defined or required specialized knowledge. Considering the anonymity of crowdsourcers, we chose to interact directly with the annotators to minimize these risks.

For the annotation process, we employed two primary annotators, supported by a third annotator responsible for resolving any discrepancies or ambiguities that may arise during the initial annotation phase. To ensure high-quality and consistent output, all annotators underwent comprehensive training. This training encompassed detailed annotation guidelines, which included definitions, examples, and specific linguistic features characteristic of each dialect. Additionally, the annotators participated in sessions to familiarize themselves with the Label Studio annotation platform, and they practiced annotating sample data, receiving feedback to fine-tune their understanding and application of the guidelines.

To guarantee the accuracy of the annotations, the third annotator reviewed instances where the initial annotations differed. The criteria for annotation focused on identifying distinct lexical items, phonological variations, and syntactic constructions unique to each dialect. Annotators were trained to recognize explicit markers—specific words or phrases characteristic of a dialect—as well as contextual clues like cultural references or idiomatic expressions, ensuring a thorough and nuanced analysis. Figure 3 illustrates sample instances of the annotation task.

Throughout this process, annotators were instructed to identify the dialect of the text presented to them and mark their selection using the provided checkboxes corresponding to each dialect. It was imperative that the annotators possess native-level proficiency in Arabic to ensure the accuracy of the dialect labeling.

All annotators involved were native Arabic speakers, each with a collegiate level of education. Specifically, for the annotation of the Saudi Dialect Corpus (SDC-5), annotators were recruited from within Saudi Arabia to ensure an intrinsic understanding of the regional dialects. Similarly, the annotation of the Gulf Arabic Corpus (GAC-6) was entrusted to native speakers from the Gulf countries, ensuring an authentic representation of the dialectal nuances.


**Figure 3.** The graphical Arabic user interface for manually annotating the Gulf Arabic Corpus (GAC-6). The annotator receives a sample tweet (displayed in the top line) from the GAC-6 and is instructed to classify the text into one of the designated Gulf Arabic dialects, with the six options displayed at the botton.

## 4. Results and Discussion

### *4.1. Overview of Our Compiled Corpora*

In this study, we have constructed a large-scale corpus consisting of Gulf dialect Arabic text, derived from data sourced from Twitter. For the purposes of this research, the following two corpora have been developed:

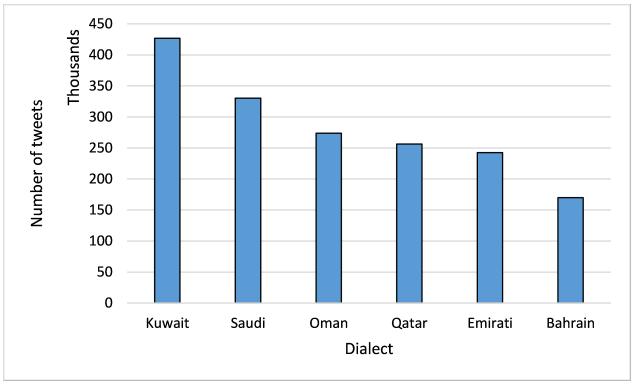
- The Gulf Arabic Corpus (GAC-6), which encompasses the dialects prevalent in the Gulf region, specifically Saudi Arabia, Bahrain, Kuwait, Oman, Qatar, and the Emirates (UAE). This corpus contains approximately 1.7 million Arabic tweets, offering a broad representation of the linguistic diversity within the Gulf countries.
- The Saudi Dialect Corpus (SDC-5), comprising around 790,000 tweets in Saudi Arabic, representing the five main dialects found within Saudi Arabia: Hijazi, Najdi, Southern, Northern, and Eastern. This corpus provides a focused insight into the linguistic variations across different regions of Saudi Arabia.

In the construction of the GAC-6, our initial collection comprised 2.6 million tweets. Following a meticulous cleaning process, which included the removal of redundant tweets, the corpus was reduced to 1.7 million tweets. The final composition of the corpus featured a diverse distribution of tweets across various Gulf dialects: 330,408 tweets were categorized under the Saudi Arabian dialect, 169,977 tweets under the Bahraini dialect, 426,771 tweets under the Kuwaiti dialect, 273,920 tweets under the Omani dialect, 256,377 tweets under the Qatari dialect, and 242,590 tweets under the Emirati dialect. The distribution of tweets by dialect within the Gulf Arabic Corpus is shown in Figure 4a. Notably, the Saudi and Kuwaiti dialects are represented by a larger volume of tweets compared to the Omani, Qatari, and Bahraini dialects. This discrepancy in tweet volumes may be attributed to the higher popularity of Twitter in Saudi Arabia and Kuwait.

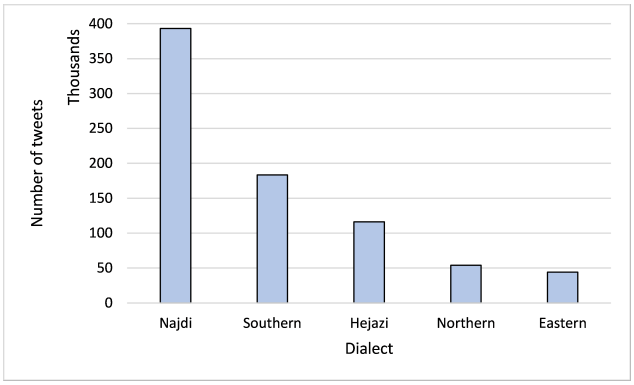
For the development of the SDC-5, our objective was to encompass all five dialects prevalent within Saudi Arabia. Prior corpora focusing on Saudi dialects [42] have often omitted the Southern and Northern dialects, attributing this exclusion to the relative scarcity and limited usage of these dialects in comparison to the Najdi, Hijazi, and Eastern dialects. Contrary to these precedents, our research endeavored to construct a corpus that inclusively represents the five Saudi dialects: Hijazi, Najdi, Southern, Northern, and Eastern. As shown in Figure 4b, the Saudi Dialect Corpus comprises a total of 790,000 tweets, distributed as follows: 116,117 tweets in the Hijazi dialect, 393,342 tweets in the Najdi dialect, 183,487 tweets in the Southern dialect, 53,883 tweets in the Northern dialect, and

44,029 tweets in the Eastern dialect. This distribution underscores our commitment to providing a comprehensive representation of the linguistic diversity within Saudi Arabia.

The comprehensive statistical overview of both the GAC-6 and the SDC-5 is shown in Table 3. This summary delineates the number of tweets and the extracted word count for each dialect within both corpora. Additionally, it presents an analysis of the tweet lengths across the dialects, specifying the minimum, maximum, and average number of words per tweet.




(**a**)




(**b**)

**Figure 4.** The number of tweets, expressed in thousands, amassed for each dialect across the two corpora. (**a**) Gulf Arabic Corpus (GAC-6). (**b**) Saudi Dialect Corpus (SDC-5).

**Table 3.** The general statistics of our two corpora, the Gulf Arabic Corpus (GAC-6) and the Saudi Dialect Corpus (SDC-5). We report the size of the tweet in words.

| | | | | **Tweet Size** | | |
|---|---|---|---|---|---|---|
| **Corpus** | **Dialects** | **# Tweets** | **# Words** | **Min** | **Max** | **Avg** |
| GAC-6 | Kuwaiti | 426,771 | 4,738,460 | 3 | 48 | 11.13 |
| | Saudi | 330,408 | 3,638,971 | 3 | 44 | 11.35 |
| | Omani | 273,920 | 3,060,739 | 4 | 45 | 11.73 |
| | Qatari | 256,377 | 3,676,452 | 4 | 46 | 14.34 |
| | Emirati | 242,590 | 2,562,961 | 5 | 40 | 10.56 |
| | Bahraini | 169,977 | 2,051,254 | 4 | 35 | 12.67 |
| SDC-5 | Najdi | 393,342 | 3,960,246 | 3 | 44 | 10.68 |
| | Southern | 183,487 | 2,994,528 | 4 | 38 | 16.32 |
| | Hejazi | 116,117 | 1,123,901 | 3 | 47 | 9.67 |
| | Northern | 53,883 | 550,647 | 4 | 39 | 10.29 |
| | Eastern | 44,029 | 441,850 | 4 | 35 | 10.35 |

Within the GAC-6, the average tweet length stands at 11.96 words. The corpus features a tweet in the Kuwaiti dialect as the longest, containing 48 words, while the shortest tweets, found within both the Kuwaiti and Saudi dialects, comprise merely three words. Conversely, in the SDC-5, the Hejazi dialect boasts the longest tweet, encompassing 47 words. The shortest tweets, each consisting of just three words, are observed in both the Najdi and Hejazi dialects, highlighting the variance in expression and conciseness across the different dialects represented in the corpora.

While the development of GAC-6 and SDC-5 contributes valuable resources to Arabic NLP, it is essential to acknowledge several limitations of this study. Data collection was limited to publicly available Twitter posts, which might not capture the complete spectrum of dialectal variations due to user demographics and the informal nature of Twitter content, such as the brevity of tweets. Additionally, the representation of specific dialects, particularly the Southern and Northern Saudi dialects, is somewhat underrepresented in the corpora, potentially limiting the generalizability of our findings to all dialects. Lastly, despite the use of both manual and automated annotation techniques aimed at ensuring high-quality data, variability in inter-annotator agreement for some dialects persists, suggesting a need for further refinement of the annotation guidelines and processes.

### *4.2. Evaluating the Quality of Corpus Annotations*

Ensuring the quality of annotations is crucial for the reliability of the corpus and its subsequent utility in developing precise dialect identification models. To assess the annotation quality within our study, we employed inter-annotator agreement (IAA) measures on the dialect annotations of the tweets. IAA measures provide insights into the consistency of annotator choices regarding dialect annotations, underpinning the validity of the annotated data.

The premise is that if annotators exhibit discrepancies in their annotations, it could be indicative of potential challenges for a dialect identification model to accurately classify those instances [77]. In our research, we quantified inter-annotator agreement using Cohen's Kappa coefficient ($\kappa$), a widely recognized statistical measure designed to evaluate the level of agreement between two annotators beyond chance in classification tasks. Cohen's Kappa is calculated using the following equation, which accounts for the observed agreement and the expected agreement by chance, thereby offering a normalized measure of annotator concordance:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{1}$$

where $P_o$ is the observed agreement among annotators, and $P_e$ can be defined as the expected agreement obtained by the random assignment of labels by annotators during the annotation process. The $P_e$ is given by

$$P_e = \frac{1}{N^2} \sum_{T} n_{T_1} \cdot n_{T_2}, \tag{2}$$

where $n_{T_1}$ and $n_{T_2}$ is the number of tokens labeled with tag $T$ by annotator 1 and annotator 2, respectively, and $N$ is the total number of annotated tokens.

The analysis of annotation quality through Cohen's Kappa revealed a commendable degree of agreement among the annotators for both corpora under study. For the GAC-6, the average Cohen's Kappa was approximately 78%, indicating a robust level of concurrence. The SDC-5 exhibited an even higher level of annotator consensus, with an average Kappa value of 90%.

The detailed breakdown of inter-annotator agreement for each region within both corpora, as presented in Table 4, showcases varying degrees of agreement across the dialects. For the GAC-6, the Bahraini, Omani, and Qatari dialect annotations fell within the "satisfactory" range, with Kappa values between 0.6 and 0.8. In contrast, the Saudi, Kuwaiti, and

Emirati dialects demonstrated "really good" agreement levels, with Kappa values ranging from 0.8 to 1, underscoring a high reliability in these annotations.

In the context of the SDC-5, the Hijazi dialect annotations achieved a notably high Kappa value of 92%, reflecting near-perfect annotator alignment. The Najdi and Southern dialects also showcased excellent agreement, with Kappa values around 91%, while the Northern and Eastern dialects exhibited very good agreement, with Kappa values at 87% and 89%, respectively. These results underscore the high quality of the annotation process, bolstering the reliability of the corpora for dialect identification research and applications.

**Table 4.** Cohen's Kappa ($\kappa$) for the individual dialects within each corpora, GAC-6 and SDC-5, along with the overall average.

| Corpus | Dialects | $\kappa$ |
|---|---|---|
| GAC-6 | Kuwaiti | 0.89 |
| | Saudi | 0.91 |
| | Omani | 0.65 |
| | Qatari | 0.67 |
| | Emirati | 0.88 |
| | Bahraini | 0.69 |
| | Average | 0.78 |
| SDC-5 | Najdi | 0.91 |
| | Southern | 0.91 |
| | Hejazi | 0.92 |
| | Northern | 0.87 |
| | Eastern | 0.89 |
| | Average | 0.90 |

## 5. Conclusions

In this study, we have outlined the methodologies utilized in compiling and developing two substantial linguistic resources for the Arabic language: the Gulf Arabic Corpus (GAC-6) and the Saudi Dialect Corpus (SDC-5). These datasets, richly annotated and multidialectal, encapsulate a broad spectrum of linguistic nuances from the Gulf and Saudi regions. GAC-6 comprises approximately 1.7 million labeled tweets from six Gulf dialects, while SDC-5 features around 790,000 tweets that reflect the predominant dialects within Saudi Arabia: Hijazi, Najdi, Southern, Northern, and Eastern.

The annotation strategy combined manual and automated methods, leveraging user location data and dialect-specific seed words to achieve precise dialect identification. A portion of the corpus was subjected to thorough manual annotation, with inter-annotator agreement metrics confirming the quality of the annotations and the overall reliability of the datasets. These resources are set to greatly advance Arabic Natural Language Processing, supporting sophisticated inquiries into dialect identification, sentiment analysis, author profiling, machine translation, and morphological analysis.

Looking ahead, we plan to expand these corpora by incorporating diverse textual data from additional online platforms, thus broadening the scope of dialectal representation. We aim to further refine our annotation methods and enhance the robustness of our techniques. Moreover, by making these comprehensive resources accessible to the global research community, we contribute to the fields of Arabic and computational linguistics. This initiative aligns with the movement toward employing sophisticated data-driven AI technologies to dissect and understand the intricate linguistic framework of the Arabic language, thereby facilitating the development of next-generation AI applications.

**Author Contributions:** Conceptualization, N.A.-S. and A.M.A.; methodology, N.A.-S.; investigation, N.A.-S.; data curation, N.A.-S.; writing—original draft, N.A.-S.; writing—review and editing, A.M.A.; supervision, A.M.A. and M.H.; funding acquisition, A.M.A. All authors have read and agreed to the published version of this manuscript.

**Funding:** The authors acknowledge funding from the Research, Development, and Innovation Authority (RDIA), Saudi Arabia, Saudi Basic Science Initiative—Basic Science Grants (BSG), number 10529.

**Data Availability Statement:** The data will be provided upon request. Please contact the corresponding author for access.

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

## Abbreviations

The following abbreviations (ordered alphabetically) are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| DA | Dialectal Arabic |
| GAC-6 | Gulf Arabic Corpus |
| IAA | inter-annotator agreement |
| MENA | Middle East and North Africa |
| MSA | Modern Standard Arabic |
| NLP | Natural Language Processing |
| SDC-5 | Saudi Dialect Corpus |

## References

1. Azmi, A.; Alsaiari, A. Arabic typography: A survey. *Int. J. Electr. Comput. Sci.* **2010**, *9*, 16–22.
2. Bouamor, H.; Habash, N.; Salameh, M.; Zaghouani, W.; Rambow, O.; Abdulrahim, D.; Obeid, O.; Khalifa, S.; Eryani, F.; Erdmann, A.; et al. The MADAR Arabic Dialect Corpus and Lexicon. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018.
3. Abdul-Mageed, M.; Alhuzali, H.; Elaraby, M. You Tweet What You Speak: A City-Level Dataset of Arabic Dialects. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018.
4. Azmi, A.M.; Aljafari, E.A. Universal web accessibility and the challenge to integrate informal Arabic users: A case study. *Univers. Access Inf. Soc.* **2018**, *17*, 131–145. [CrossRef]
5. Jarrar, M.; Habash, N.; Alrimawi, F.; Akra, D.; Zalmout, N. Curras: An annotated corpus for the Palestinian Arabic dialect. *Lang. Resour. Eval.* **2017**, *51*, 745–775. [CrossRef]
6. Azmi, A.M.; Aljafari, E.A. Modern information retrieval in Arabic–catering to standard and colloquial Arabic users. *J. Inf. Sci.* **2015**, *41*, 506–517. [CrossRef]
7. Haff, K.E.; Jarrar, M.; Hammouda, T.; Zaraket, F. Curras + Baladi: Towards a levantine corpus. *arXiv* **2022**, arXiv:2205.09692.
8. Jarrar, M.; Zaraket, F.A.; Hammouda, T.; Alavi, D.M.; Wahlisch, M. LISAN: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Copora with Morphological Annotations. In Proceedings of the 2023 20th ACS/IEEE International Conference on Computer Systems and Applications, Giza, Egypt, 4–7 December 2022.
9. Shoufan, A.; Alameri, S. Natural language processing for dialectical Arabic: A survey. In Proceedings of the Second Workshop on Arabic Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 36–48.
10. Azmi, A.M.; Al-Ghadir, A.I. Using Twitter as a digital insight into public stance on societal behavioral dynamics. *J. King Saud Univ. Comput. Inf. Sci.* **2024**, *36*, 102078. [CrossRef]
11. AlShenaifi, N.; Azmi, A. Arabic dialect identification using machine learning and transformer-based models. In Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP 2022), Abu Dhabi, United Arab Emirates, 8 December 2022.
12. Alshargi, F.; Dibas, S.; Alkhereyf, S.; Faraj, R.; Abdulkareem, B.; Yagi, S.; Kacha, O.; Habash, N.; Rambow, O. Morphologically annotated corpora for seven Arabic dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 1–2 August 2019; pp. 137–147.
13. Lichouri, M.; Abbas, M.; Freihat, A.A.; Megtouf, D.E.H. Word-level vs. sentence-level language identification: Application to Algerian and Arabic dialects. *Procedia Comput. Sci.* **2018**, *142*, 246–253. [CrossRef]
14. Azmi, A.; Al-Thanyyan, S. Ikhtasir—A user selected compression ratio Arabic text summarization system. In Proceedings of the 2009 International Conference on Natural Language Processing and Knowledge Engineering, Dalian, China, 24–27 September 2009; pp. 1–7.
15. Al-Jouie, M.F.; Azmi, A.M. Automated Evaluation of School Children Essays in Arabic. In Proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing 2017), Dubai, United Arab Emirates, 5–6 November 2017; Volume 117, pp. 19–22.
16. Mohammed, E.A.; Aziz, M.J.A. English to Arabic machine translation based on reordring algorithm. *J. Comput. Sci.* **2011**, *7*, 120–128. [CrossRef]

17. Alnefaie, R.; Azmi, A.M. Automatic minimal diacritization of Arabic texts. In Proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing 2017), Dubai, United Arab Emirates, 5–6 November 2017; Volume 117, pp. 169–174.
18. Tarmom, T.; Teahan, W.; Atwell, E.; Alsalka, M. Compression vs Traditional Machine Learning Classifiers to Detect Code-switching in Varieties and Dialects: Arabic as a Case Study. *J. Nat. Lang. Eng.* **2020**, *26*, 663–676. [CrossRef]
19. Alhussain, A.; Azmi, A.M. Beyond Event-Centric Narratives: Advancing Arabic Story Generation with Large Language Models and Beam Search. *Mathematics* **2024**, *12*, 1548. [CrossRef]
20. Al-Thanyyan, S.S.; Azmi, A.M. Simplification of Arabic text: A hybrid approach integrating machine translation and transformer-based lexical model. *J. King Saud Univ. Comput. Inf. Sci.* **2023**, *35*, 101662. [CrossRef]
21. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 23–24 June 2014; pp. 55–60.
22. Alwaneen, T.H.; Azmi, A.M. Stacked dynamic memory-coattention network for answering why-questions in Arabic. *Neural Comput. Appl.* **2024**, *36*, 8867–8883. [CrossRef]
23. Mannaa, Z.M.; Azmi, A.M.; Aboalsamh, H.A. Computer-assisted i'raab of Arabic sentences for teaching grammar to students. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *10*, 8909–8926. [CrossRef]
24. Abu Kwaik, K.; Saad, M.; Chatzikyriakidis, S.; Dobnik, S. Shami: A Corpus of Levantine Arabic Dialects. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), Miyazaki, Japan, 7–12 May 2018.
25. Abdelali, A.; Mubarak, H.; Samih, Y.; Hassan, S.; Darwish, K. QADI: Arabic dialect identification in the wild. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kiev, Ukraine, 19 April 2021; pp. 1–10.
26. Alharbi, A.; Lee, M. Kawarith: An Arabic Twitter corpus for crisis events. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kiev, Ukraine, 19 April 2021; pp. 42–52.
27. Darwish, K.; Sajjad, H.; Mubarak, H. Verifiably effective Arabic dialect identification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1465–1468.
28. Al-Kabi, M.; Al-Ayyoub, M.; Alsmadi, I.; Wahsheh, H. A prototype for a standard Arabic sentiment analysis corpus. *Int. Arab J. Inf. Technol.* **2016**, *13*, 163–170.
29. Ahmed, A.; Ali, N.; Alzubaidi, M.; Zaghouani, W.; Abd-alrazaq, A.A.; Househ, M. Freely available Arabic corpora: A scoping review. *Comput. Methods Programs Biomed. Update* **2022**, *2*, 100049. [CrossRef]
30. Mubarak, H. Dial2MSA: A tweets corpus for converting dialectal Arabic to modern standard Arabic. In Proceedings of the 3rd Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT3), Miyazaki, Japan, 8 May 2018; pp. 49–53.
31. Zaghouani, W. Critical Survey of the Freely Available Arabic Corpora. In Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme, Reykjavik, Iceland, 27 May 2014; pp. 1–8.
32. Al-Sabbagh, R.; Girju, R. YADAC: Yet another Dialectal Arabic Corpus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, 21–27 May 2012; pp. 2882–2889.
33. Khalifa, S.; Habash, N.; Abdulrahim, D.; Hassan, S. A Large Scale Corpus of Gulf Arabic. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, Portorož, Slovenia, 23–28 May 2016; pp. 4282–4289.
34. Khalifa, S.; Habash, N.; Eryani, F.; Obeid, O.; Abdulrahim, D.; Al Kaabi, M. A morphologically annotated corpus of Emirati Arabic. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018.
35. Abdulrahim, D.; Inoue, G.; Shamsan, L.; Khalifa, S.; Habash, N. The Bahrain Corpus: A Multi-genre Corpus of Bahraini Arabic. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 2345–2352.
36. Al-Mulla, S.; Zaghouani, W. Building a corpus of Qatari Arabic expressions. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 11–16 May 2020; pp. 24–31.
37. Alshutayri, A.O.O.; Atwell, E. Exploring Twitter as a source of an Arabic dialect corpus. *Int. J. Comput. Linguist. (IJCL)* **2017**, *8*, 37–44.
38. Elaraby, M.; Abdul-Mageed, M. Deep models for Arabic dialect identification on benchmarked data. In Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), Santa Fe, NM, USA, 20 August 2018; pp. 263–274.
39. Harrat, S.; Meftouh, K.; Smaili, K. Creating parallel Arabic dialect corpus: Pitfalls to avoid. In Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING), Budapest, Hungary, 17–23 April 2017.
40. Alshutayri, A.; Atwell, E. Classifying Arabic dialect text in the social media Arabic dialect corpus (SMADC). In Proceedings of the 3rd Workshop on Arabic Corpus Linguistics, Cardiff, UK, 22 July 2019; pp. 51–59.
41. Alshutayri, A.; Atwell, E. Arabic dialects annotation using an online game. In Proceedings of the IEEE 2nd International Conference on Natural Language and Speech Processing (ICNLSP), Algiers, Algeria, 25–26 April 2018; pp. 1–5.
42. Bayazed, A.; Torabah, O.; AlSulami, R.; Alahmadi, D.; Babour, A.; Saeedi, K. SDCT: Multi-dialects corpus classification for Saudi Tweets. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 216–223. [CrossRef]
43. Darwish, K.; Magdy, W. Arabic information retrieval. *Found. Trends Inf. Retr.* **2014**, *7*, 239–342. [CrossRef]

44. Al-Razgan, M.; Alrowily, A.; Al-Matham, R.N.; Alghamdi, K.M.; Shaabi, M.; Alssum, L. Using diffusion of innovation theory and sentiment analysis to analyze attitudes toward driving adoption by Saudi women. *Technol. Soc.* **2021**, *65*, 101558. [CrossRef]
45. Zaidan, O.; Callison-Burch, C. The Arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19 June 2011; pp. 37–41.
46. Alsarsour, I.; Mohamed, E.; Suwaileh, R.; Elsayed, T. Dart: A large dataset of dialectal Arabic tweets. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018.
47. Zaghouani, W.; Charfi, A. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. *arXiv* **2018**, arXiv:1808.07674.
48. Sadat, F.; Kazemi, F.; Farzindar, A. Automatic identification of Arabic dialects in social media. In Proceedings of the First International Workshop on Social Media Retrieval and Analysis (SoMeRA'14), Gold Coast, QLD, Australia, 11 July 2014; Association for Computing Machinery: New York, NY, USA; pp. 35–40.
49. Alshutayri, A.; Atwell, E. Creating an Arabic dialect text corpus by exploring Twitter, Facebook, and online newspapers. In Proceedings of the OSACT3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools, LREC, Miyazaki, Japan, 8 May 2018; pp. 54–61.
50. Bouamor, H.; Habash, N.; Oflazer, K. A Multidialectal Parallel Corpus of Arabic. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, 26–31 May 2014; pp. 1240–1245.
51. Takezawa, T.; Kikui, G.; Mizushima, M.; Sumita, E. Multilingual spoken language corpus development for communication research. *Int. J. Comput. Linguist. Chin. Lang. Process.* **2007**, *12*, 303–324.
52. Mubarak, H.; Darwish, K. Using Twitter to collect a multi-dialectal corpus of Arabic. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar, 25 October 2014; pp. 1–7.
53. Al-Twairesh, N.; Al-Matham, R.; Madi, N.; Almugren, N.; Al-Aljmi, A.H.; Alshalan, S.; Alshalan, R.; Alrumayyan, N.; Al-Manea, S.; Bawazeer, S.; et al. SUAR: Towards building a corpus for the Saudi dialect. *Procedia Comput. Sci.* **2018**, *142*, 72–82. [CrossRef]
54. Pasha, A.; Al-Badrashiny, M.; Diab, M.; El Kholy, A.; Eskander, R.; Habash, N.; Pooleery, M.; Rambow, O.; Roth, R. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014; pp. 1094–1101.
55. Alowisheq, A.; Al-Twairesh, N.; Altuwaijri, M.; Almoammar, A.; Alsuwailem, A.; Albuhairi, T.; Alahaideb, W.; Alhumoud, S. MARSA: Multi-domain Arabic resources for sentiment analysis. *IEEE Access* **2021**, *9*, 142718–142728. [CrossRef]
56. Elgibreen, H.; Faisal, M.; Al Sulaiman, M.; Abdou, S.; Mekhtiche, M.A.; Moussa, A.M.; Alohali, Y.A.; Abdul, W.; Muhammad, G.; Rashwan, M.; et al. An incremental approach to corpus design and construction: Application to a large contemporary Saudi corpus. *IEEE Access* **2021**, *9*, 88405–88428. [CrossRef]
57. Alruily, M. Issues of dialectal Saudi twitter corpus. *Int. Arab J. Inf. Technol.* **2020**, *17*, 367–374. [CrossRef]
58. Al-Ghadir, A.; Azmi, A. A Study of Arabic Social Media Users—Posting Behavior and Author's Gender Prediction. *Cogn. Comput.* **2019**, *11*, 71–86. [CrossRef]
59. Cotterell, R.; Callison-Burch, C. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014; pp. 241–245.
60. Ibrahim, H.S.; Abdou, S.M.; Gheith, M. MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis. In Proceedings of the IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), Kolkata, India, 9–11 July 2015; pp. 353–358.
61. Hegazi, M.O.; Al-Dossari, Y.; Al-Yahy, A.; Al-Sumari, A.; Hilal, A. Preprocessing Arabic text on social media. *Heliyon* **2021**, *7*, e06191. [CrossRef]
62. Charfi, A.; Zaghouani, W.; Mehdi, S.H.; Mohamed, E. A fine-grained annotated multi-dialectal Arabic corpus. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 3–12 September 2019; pp. 198–204.
63. Althobaiti, M.J. Creation of annotated country-level dialectal Arabic resources: An unsupervised approach. *Nat. Lang. Eng.* **2022**, *28*, 607–648. [CrossRef]
64. Alshutayri, A.; Atwell, E. A social media corpus of Arabic dialect text. In *Building Computer-Mediated Communication Corpora for Sociolinguistic Analysis*; Stemle, E., Wigham, C.R., Eds.; Presses Universitaires Blaise Pascal: Clermont-Ferrand, France, 13 June 2019; pp. 1–23.
65. Refaee, E.; Rieser, V. An Arabic twitter corpus for subjectivity and sentiment analysis. In Proceedings of the 9th International Language Resources and Evaluation Conference, Reykjavik, Iceland, 26–31 May 2014; pp. 2268–2273.
66. Gugliotta, E.; Dinarelli, M. TArC: Tunisian Arabish Corpus First complete release. *arXiv* **2022**, arXiv:2207.04796.
67. Alabbas, W.; al Khateeb, H.M.; Mansour, A.; Epiphaniou, G.; Frommholz, I. Classification of colloquial Arabic tweets in real-time to detect high-risk floods. In Proceedings of the IEEE International Conference On Social Media, Wearable And Web Analytics (Social Media), London, UK, 19–20 June 2017; pp. 1–8.
68. Baly, R.; Khaddaj, A.; Hajj, H.; El-Hajj, W.; Shaban, K.B. Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in Arabic levantine tweets. *arXiv* **2019**, arXiv:1906.01830.
69. Zaghouani, W.; Charfi, A. Guidelines and annotation framework for Arabic author profiling. *arXiv* **2018**, arXiv:1808.07678.

70. Al-Laith, A.; Shahbaz, M.; Alaskar, H.F.; Rehmat, A. AraSenCorpus: A semi-supervised approach for sentiment annotation of a large Arabic text corpus. *Appl. Sci.* **2021**, *11*, 2434. [CrossRef]
71. Zahir, J. IADD: An Integrated Arabic Dialect Identification Dataset. *Data Brief* **2022**, *40*, 107777. [CrossRef]
72. Kwaik, K.A.; Chatzikyriakidis, S.; Dobnik, S.; Saad, M.; Johansson, R. An Arabic Tweets Sentiment Analysis Dataset (ATSAD) using Distant Supervision and Self Training. In Proceedings of the 4th Workshop on Open-source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 12 May 2020; pp. 1–8.
73. Mahany, A.; Khaled, H.; Nouh, E.; Aljohani, N.; Ghoniemy, S. Annotated Corpus with Negation and Speculation in Arabic Review Domain: NSAR. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 38–46. [CrossRef]
74. Guellil, I.; Azouaou, F.; Chiclana, F. ArAutoSenti: Automatic Annotation and New Tendencies for Sentiment Classification of Arabic Messages. *Soc. Netw. Anal. Min.* **2020**, *10*, 75. [CrossRef]
75. Almuqren, L.; Alzammam, A.; Alotaibi, S.; Cristea, A.; Alhumoud, S. A review on corpus annotation for Arabic sentiment analysis. In *Social Computing and Social Media: Applications and Analytics (SCSM 2017), Part II, LNCS 10283*; Springer: Cham, Switzerland, 2017; pp. 215–225.
76. Almuzaini, H.A.; Azmi, A.M. An unsupervised annotation of Arabic texts using multi-label topic modeling and genetic algorithm. *Expert Syst. Appl.* **2022**, *203*, 117384. [CrossRef]
77. Al-Twairesh, N.; Al-Khalifa, H.; Al-Salman, A.; Al-Ohali, Y. Arasenti-tweet: A corpus for Arabic sentiment analysis of Saudi tweets. *Procedia Comput. Sci.* **2017**, *117*, 63–72. [CrossRef]