IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Systematic Literature Review of Dialectal Arabic: Identification and Detection

Ashraf Elnagar [1,5], Sane Yagi [2,5], Ali Bou Nassif [3,5], Ismail Shahin [4,5], and Said A. Salloum [5]

[1]Department of Computer Science, University of Sharjah, Sharjah, UAE
[2]Department of Foreign Language, University of Sharjah, Sharjah, UAE
[3]Computer Engineering Department, University of Sharjah, Sharjah, UAE
[4]Electrical Engineering Department, University of Sharjah, Sharjah, UAE
[5]Machine Learning and NLP Research Group, University of Sharjah, Sharjah, UAE

Corresponding author: Ashraf Elnagar (e-mail: ashraf@sharjah.ac.ae).

This work was supported in part by the UoS under Grant 1702141151-P.

ABSTRACT It is becoming increasingly difficult to know who is working on what and how in computational studies of Dialectal Arabic. This study comes to chart the field by conducting a systematic literature review that is intended to give insight into the most and least popular research areas, dialects, machine learning approaches, neural network input features, data types, datasets, system evaluation criteria, publication venues, and publication trends. It is a review that is guided by the norms of systematic reviews. It has taken account of all the research that adopted a computational approach to dialectal Arabic identification and detection and that was published between 2000 and 2020. It collected, analyzed, and collated this research, discovered its trends, and identified research gaps. It revealed, inter alia, that our research effort has not been directed evenly between speech and text or between the vernaculars; there is some bias favoring text over speech, regional varieties over individual vernaculars, and Egyptian over all other vernaculars. Furthermore, there is a clear preference for shallow machine learning approaches, for the use of n-grams, TF-IDF, and MFCC as neural network features, and for accuracy as a statistical measure of validation of results. This paper also pointed to some glaring gaps in the research: (1) total neglect of Mauritanian and Bahraini in the continuous Arabic language area and of such enclave varieties as Anatolian Arabic, Khuzistan Arabic, Khurasan Arabic, Uzbekistan Arabic, the Subsaharan Arabic of Nigeria and Chad, Djibouti Arabic, Cypriot Arabic and Maltese; (2) scarcity of city dialect resources; (3) rarity of linguistic investigations that would complement our research; (4) and paucity of deep machine learning experimentation.

INDEX TERMS Arabic dialects; Arabic Natural Language Processing; Dialect identification; Modern Standard Arabic; Systematic review.

I. INTRODUCTION

Arabic was adopted as an official language of the United Nations by the General Assembly in its 28th session on 18 December 1973. Resolution 3190 [1] put into effect Arabic as an official and working language of the General Assembly and its Main Committees in recognition of the fact that it was the language of nineteen Members of the United Nations and a working language in specialized UN agencies.

Arabic is the national language of more than 422 million people [2] and is ranked as the fifth most extensively used language in the world. It has two primary varieties: Modern Standard Arabic (MSA), the formal written language, and Dialectal Arabic (DA), the informal spoken language that varies greatly across regions and countries. MSA is exclusively used in news bulletins, publications, official speeches, film subtitles, and religious rites and ceremonies [3]. MSA is ipso facto the lingua franca of Arabs, often resorted to in speech to ensure mutual intelligibility. DA, on the other hand, is the variety spoken at home, with friends, at the marketplace, and in all other informal contexts. It is the intimate variety that speakers feel most comfortable with.

In the past decade, there has been an unprecedented surge of interest in DA which translated in a flurry of natural language processing (NLP) research [4]. This is primarily attributable to political considerations that made funding available to researchers. Technically also, the extensive use

of Arabic dialects in social media made data abundant [3], [5]. This coupled with advances in machine learning made it all alluring to researchers.

Automatic processing of Arabic is challenging if for nothing, for the fact that it is written in a non-Roman script and written from right to left. As it is non-European, its lexis hardly has cognates in any Indo-European languages. Morphologically, Arabic is root-based with introflexive, fusional morphology, and inflectional syntax. The challenges of processing it are outlined in [6], [7]. An excellent book-size explanation of the issues peculiar to MSA is [6]; it briefly alludes to phenomena in DA. It highlights all those issues that the NLP community needs to be cognizant of, whether those relating to orthography, morphological analysis, part of speech tagging, or machine translation, etc. Likewise, [7] explains, in an article, the nature of MSA, identifies the challenges it poses to natural language processing, explains how researchers have been dealing with these challenges, and attempts to suggest some solutions that might guide current research.

Authors of [3], [8] briefly define DA, list its features, discuss the relationship between the Standard language and its regional varieties, and give a classification of the major regional dialects.

A concerted effort has been found by [4], [5] to be put into DA corpus and lexicon construction, speech recognition, and dialect identification (e.g., [9], [10]). Less effort has been made in morphological analysis and machine translation. It appears that DA syntactic analysis has been largely neglected.

Approaches to the speech recognition of DA are addressed in [11]. The authors recognize the sparsity of DA speech resources, so they describe how existing MSA speech data can be utilized in DA speech recognition and outline how acoustic models may be adapted for DA speech recognition.

In terms of morphological analysis of DA, there have been a few studies (e.g., [12]–[14]) which primarily leveraged MSA resources for the study of DA. For instance, [12] retargeted an existing MSA morphology modeling tool, namely MADA – Morphological Analysis and Disambiguation of Arabic, for the analysis of Egyptian Arabic. Similarly, [13] adapted Al-Khalil, the MSA morphological analyzer, to Tunisian Arabic, by creating a Tunisian lexicon that is based on an MSA lexicon and MSA derivation patterns and by adding Tunisian roots and patterns to it. Authors of [14] extended the database of an MSA morphological analyzer by adding a set of handwritten rules and affixes that were peculiar to Levantine, Egyptian, and Iraqi Arabic. The same is true in [15] where Buckwalter's MS morphological (BAMA) was adapted to two dialects in Algeria. This involved modifying BAMA's three tables (i.e., stems, suffixes, and prefixes) and three compatibility tables that define relations between these word parts. With the aim of paving the way for efficient morphological analysis of Moroccan Arabic, an MSA-Moroccan dictionary of 18,000 entries was built in [10] by first manually translating the standard Arabic words into their regional equivalents and then by enriching the dictionary with Moroccan words from the internet.

Several studies [16]–[19] were motivated to either create dialectal resources to improve machine translation or use machine translation to create resources for DA. In [16], a parallel Levantine Arabic-English and an Egyptian Arabic-English parallel corpora were constructed, first by selecting passages with a relatively high percentage of dialectal words from a monolingual Arabic corpus, then benefiting from crowdsourcing to classify them by dialect, segment them into sentences, and translate them into English. This was followed by the development of a Dialectal Arabic MT system. It was established that a small percentage of dialectal data could dramatically impact the performance of an MT system.

A full-fledged dialect to standard Arabic MT system was developed in [17]. Elissa is a rule-based translation system that relies on morphological analysis of DA, complex morphological transfer rules, and dictionaries when it translates from Levantine, Egyptian, Iraqi, and to a lesser degree Gulf Arabic into MSA. The author of [18] used an MSA finite state machine morphological analyzer, a DA-MSA sentence-aligned parallel corpus, and machine learning techniques to convert DA into MSA then submits the output texts to an MT system, a hybrid statistical and rule-based system to translate them into English. Machine translation was also conducted in [19] to translate from DA to MSA. This research developed a Parallel Arabic Dialect Corpus of 6400 sentences in the Arabic dialects of Algeria, Annaba, Southern Tunisia, Syrian, Palestinian, and MSA, then experimented with machine translating texts from the five dialects to MSA and obtained encouraging results.

With the upsurge in dialect research, it soon became difficult to know how dialect research was evolving. In 2015, a review of the literature was carried out by [5]. It found and reviewed 89 studies that were carried out on DA until then. It classified their contributions into four categories: (1) basic language analysis; (2) resource building; (3) dialect identification; and (4) semantic analysis as expressed in machine translation and sentiment analysis. Another review appeared in [20]. It focused only on new research, 74% of which was published between 2015 and 2018 but was not exclusively focused on DA but rather on Classical Arabic, MSA, and DA in both Arabic script and Roman script. It reviewed 90 studies that focused on NLP in general.

Our consideration of the work that has been done so far on DA encouraged us to conduct a systematic review that transcends the limitations of these two reviews. The review in [5] is too old to be of value to current research; Furthermore, it makes no claim that it was exhaustive in its selection of articles, and neither does it explain the principles that it was guided by. Similarly, the review in [20] does not make any assertions that it was systematic or exhaustive. It reviewed 90 studies, three quarters of which were published

in three years (2015-2018). Furthermore, it does not claim to have been focused exclusively on dialect studies or on the DA language variety. In fact, it states in the title that it was a review of NLP work. In any case, neither [5] nor [20] took stock of the tools and resources that were developed for DA in the reviewed literature.

It is timely now to establish our bearings in the midst of this flurry of research on DA. This paper will report on a systematic review of all the computational literature on DA that was published between 2000 and 2020. It seeks to upgrade and remedy previous reviews by being transparent and methodical in its selection of literature for review, exhaustive and comprehensive, classificatory, and thoroughly analytical. It will show the direction that the research community is taking in relation to computational dialect research of both speech and text modalities, identify gaps in the literature, and answer the following specific research questions:

RQ1: What are the key research areas in Arabic computational dialect studies?

RQ2: What are the dialects of concern in the reviewed articles?

RQ3: What are the machine learning algorithms used in Arabic dialect studies?

RQ4: What are the input features used in Arabic dialect studies?

RQ5: Which types of data are the most widely used in Arabic dialect identification studies?

RQ6: What are the datasets most often utilized in Arabic dialect studies?

RQ7: What are the trends across time in Arabic dialect identification?

RQ8: What are the evaluation criteria of machine learning techniques that were used in Arabic dialect identification?

RQ9: Where do the results of Arabic dialect research get published?

A. Arabic dialects

Arabic has been classified by [21] as a morphologically introflexive, fusional language and we know it is syntactically synthetic and its word order is free, with significant bias towards a Verb-Subject-Object order. It has been viewed by Arabic-speaking NLP specialists as especially challenging [3], [6], [7], etc., the benchmark being English. Edward Sapir [22] classified English as a mixed-relational fusional language. Syntactically, it is analytic and its word order is Subject-Verb-Object. As the two languages use different orthographies and they differ in their linguistic typology, computer scientists find challenging the adaptation of technologies made for English. If we focus on the writing direction alone and observe how English writes left to right but Arabic adopts a right to left writing orientation, we will immediately realize that tools developed for the processing of English are not going to work for Arabic without much tweaking and possibly radical alteration if not total replacement. We acknowledge that the development of resources and tools for the processing of Arabic is involving and at times daunting. However, determination and consistency of efforts have removed numerous hurdles and have culminated in successful adaptation of English resources and tools and often in the development of native grown solutions. No sooner have NLP specialists developed solutions for the successful automatic processing of MSA, than they realized the limitations in processing the contemporary Arabic of social media.

Contemporary Arabic of the Social Media (CASM) is problematic for NLP because it may use Arabic or Latin orthography and often mixes MSA with DA. The two varieties are different as rightly identified in [3], [8]; however, the differences are grossly exaggerated either for political reasons that NLP specialists are oblivious to but submissively follow, or for research-justification and paper publication purposes. The differences between DA and MSA are differences between the spoken and written modes of expression. These differences are recognized in all languages. The claim that MSA is no one's native language is as much true as Received Pronunciation is no one's native tongue; that is why it is called 'Received'. You receive it in school, just like MSA is received in school. DA is the spoken variety that is acquired at home but the written variety, MSA, is learned at school. The Standard variety of any language is no one's native tongue. It is an ideal. It is not associated with a geographical region either. Standard English is not the native language of London, for there are several spoken varieties (e.g. Cockney) that no one identifies with Standard English. Furthermore, the regional varieties are also on a continuum such that it would be difficult to be definitive about, say Saudi Arabic, because there is a multitude of varieties in Saudi Arabia. Even if one would want to specify varieties by city claiming, for instance, that there is a Jedda or Makkah variety, it would be difficult because of variation due to social class, gender, age, profession, etc. In other words, the decision to consider significant some linguistic differences is politically motivated. If the intention is to divide, then surely the differences would constitute enough justification to give labels to some variation and ignore some others. Take the differences between Urdu and Hindi as an example. They are varieties of the same language, yet for political reasons they are considered different languages and they even use different orthographies. If the intention is to unite, on the other hand, then the differences would be ignored. Take the example of Greek's Katharevusa and Dhimotiki. Because Greece wanted to reconcile itself with its past, it created Katharevusa artificially from Classical Greek and treated it as the written variety. Katharevusa is not spoken by anyone, it is not the native language of anyone, yet it is a variety of Greek.

The co-existence of language varieties is a phenomenon recognized by linguists as diglossia. Ferguson [23] defined

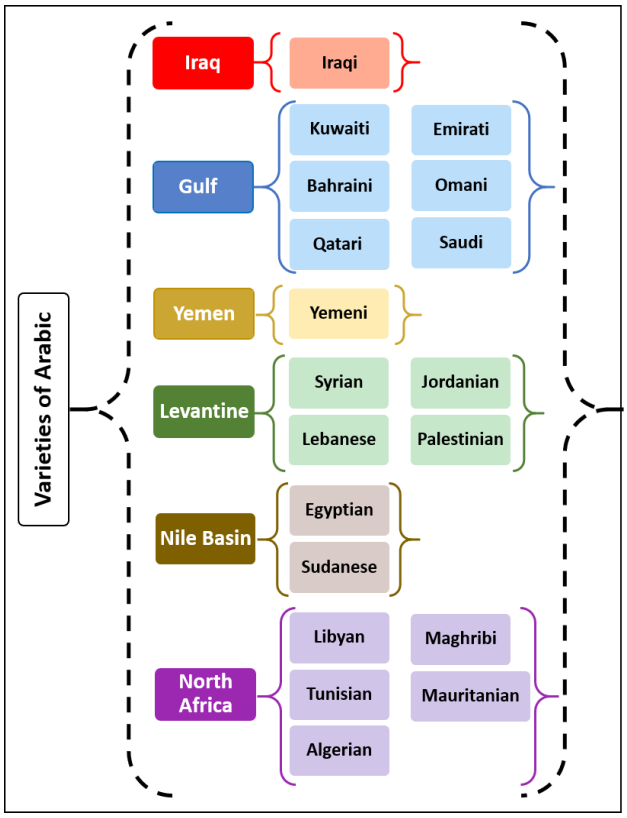


FIGURE 1 **Arabic country-dialects in the continuous language-area.**

this phenomenon as "a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any section of the community for ordinary conversation" (p.336). He added further that "Diglossia is apparently not limited to any geographical region or language family" (p.337) and that "Arabic diglossia seems to reach as far back as our knowledge of Arabic goes, and the superposed 'Classical' language has remained relatively stable" (p.327).

Thus, the Arabic that Abu Jahl spoke with Abi Sufyan 1450 years ago is not the same as the Fusha in the poetry of their contemporary Hassaan Bin Thabit. They all spoke at home differently. Then the claim made by some NLP specialists that the differences between MSA and DA are akin to the differences between the Romance languages must be an exaggeration.

B. Arabic dialect identification

Dialect identification (DI) is the automatic recognition and classification of language varieties by comparing new data to previously annotated or classified old data using some similarity measures. The latest developments of communication technologies and extensive use of social media have made it imperative to develop technologies like language/dialect identification. DI technologies have been used in the monitoring of health and safety [24]–[27], real-time disaster operation management [28], and human mobility assessment [29]–[31]. Moreover, the application of such technology has allowed automatic filtration of foreign texts [32], and has facilitated the acquisition of multilingual information from a range of data sources including the web [33] and has supported machine translation [16], [34], [35]. In this paper, we will review all the recent research in the automatic identification and classification of the dialects of the Arabic language.

Arabic is spoken as a first language in the geographical region that extends between the Atlantic Ocean in the west and the Persian Gulf in the East, including all the countries south and south east of the Mediterranean Sea, west of the Sahara Desert, around the Nile, in Malta, in the Arabian Peninsula, and in enclaves in Iran and central Asia. It is also taught and spoken as a second language or a lingua franca in all countries with Muslim communities. FIGURE 1 displays country-dialects in what Watson [8] calls 'continuous Arabic language area', the uninterrupted land of Arab countries in the Middle East, the Nile Basin, and North and West Africa; thus, it excludes Nigeria, Somalia, Djibouti, and Comoros.

Arabic has a multitude of dialects, i.e., regional varieties. Notwithstanding the dialect distribution continuum, every one of the 22 Arab countries may claim to speak a dialect of its own. Thus, researchers talk of Sudanese, Algerian, and Tunisian Arabic, etc. In addition, there are also Classical Arabic (CA), the language of scholarship up until the Arab renaissance that was triggered by Napoleon's invasion of Egypt in 1798, and Modern Standard Arabic (MSA), the language of scholarship since then, the language variety taught at school, and used at formal occasions and in publications. Both CA and MSA are written modes of the language while the dialects are the spoken colloquial varieties that are used at home and in informal contexts. Differences between CA and MSA are primarily in the vocabulary as some words or senses of these words became archaic over time; otherwise, structures in MSA are a subset of structures in CA [36].

Dialects differ substantially from both CA and MSA, if for nothing else than phonology, vocabulary, and grammar. This is widely acknowledged [35], [37], [38] though grossly exaggerated; learning CA and MSA has been likened to learning a foreign language [19]. The debate here is more politically than linguistically motivated. Similarly, it has been claimed that the differences between DA varieties are like the differences between the Romance languages [11]; other people liken the differences to those between Norwegian, Swedish, and Danish or the differences between Czech, Slovak, and Polish [8].

What is unequivocal is that CA and MSA are the same since they share the same grammar, morphology, and lexicon. The bulk of their vocabulary is shared, conceding some variation in form, and differences in pronunciation, sense, and context of use [36], [39]. The distinction between them is to capture this lexical difference. DA varieties, on the other hand, are descendants of spoken varieties of classical eras [40]. They share with CA and MSA enough vocabulary, morphology, and grammar to be called Arabic, but they do differ from them and from one another significantly in pronunciation, word structure, sentence grammar, and meaning [41]. Badawi, in [42],places these varieties on a continuum that he labels as 'language level continuum', with Classical Arabic at the top, followed by MSA, formal spoken Arabic, colloquial DA of the literate, and colloquial DA of the illiterate at the bottom. The boundaries between these varieties are fuzzy. That is why dialect classifiers would fail to find texts that are exclusively dialectal in vocabulary. Most often the mention of a single dialectal word in a sentence would bias the classifier towards treating that sentence as dialectal. Furthermore, it is quite difficult to be definitive in the assignment of a sentence to one dialect as linguistic features are often shared. Instead, it is found easier to classify a whole text because then the spelling, lexical items, and grammatical structures would all be taken into account.

Dialects may be found to be on a time, a space, a social, and a religious and ethnic continuum. Cognizant of the fuzziness of boundaries, therefore, Arabic dialects may be classified on the time continuum into proto-Arabic, classical, middle, modern, and contemporary; on the space continuum, into South Arabia, Arabian Peninsula, Levantine, Mesopotamia, Nile, North Africa, Sub-Sahara, and the periphery; on the social continuum into Bedouin, ruralite, and urbanite; and on the religious and ethnic continuum into Muslim, Christian, Jewish, Sunni, Shia, Druze, Alawite, Malekite, Ibadi, Arab, Berber, etc. In automatic dialect classification, however, the focus thus far has been on geographical classification. Five categories of dialects were identified in [43]. The computer science literature abounds with references to Egyptian, Gulf, Iraqi, Levantine, and Maghribi.

C. Knowledge gap

A research gap in the existing literature on Arabic dialects may be identified by performing a bibliometric mapping analysis with assistance of the VOSviewer. In the Scopus database between 2000 and 2020, there has been a total of 940 articles with titles that include "Arabic dialects", "Colloquial Arabic" or "Arabic vernaculars". Upon closer inspection, it becomes clear that the keywords most extensively employed in the research of concern to us are "Arabic Dialect Identification", "Dialect Recognition", and "Dialect Classification" (see the results of the bibliometric analysis in APPENDIX A). It is also observed that merely 15% of these articles highlight "Arabic Dialect Identification". The outcome of bibliometric analysis shown in FIGURE 2 also points out that the phrase "Dialect Recognition" was part of the title of 4% of these articles, of which 27 were journal articles. Of these, "Survey of Arabic Dialect" was the topic of eight articles.

Two of these surveys were the most significant. The first was an extensive literature review of natural language processing of DA, [5]. It reviewed the literature from 2004 up to 2015 and identified four areas of research: basic language analysis, language resource development, semantic analysis and synthesis, and dialect identification. It recognized the emergence of a trend in dialectal Arabic NLP, observed strong concentration on the Egyptian vernacular, noticed that research aimed at the development of DA corpora and corpus annotation, and detected a gap in research, the absence of DA syntactic analysis.

The second most significant survey was [20]. It reviewed research, three quarters of which were published between 2015-2018. It was not an exclusive review of DA. It reviewed work on Classical Arabic and MSA as well. Furthermore, like its predecessor that it came to complement (i.e., [5]), it was not concerned with dialect studies per se but rather with NLP in general. It broadened the canvas, though, to include not only DA in native Arabic orthography but also that in Arabizi, Arabic written in Roman orthography. It had the additional advantage of presenting the resources and tools associated with the reviewed literature. Three of its major findings were that (1) only few works were concerned with Classical Arabic, (2) Arabizi was quickly emerging as a research area, and (3) none of the resources developed in the reviewed literature is yet publicly accessible.

Our current review differs from all its predecessors by being (1) more up to date, (2) systematic in its inclusion and exclusion of literature, (3) exhaustive in its coverage, (4) and exhaustive in its appraisal of tools and resources involved in DA research. It reviews all the literature published between 2000 and 2020.

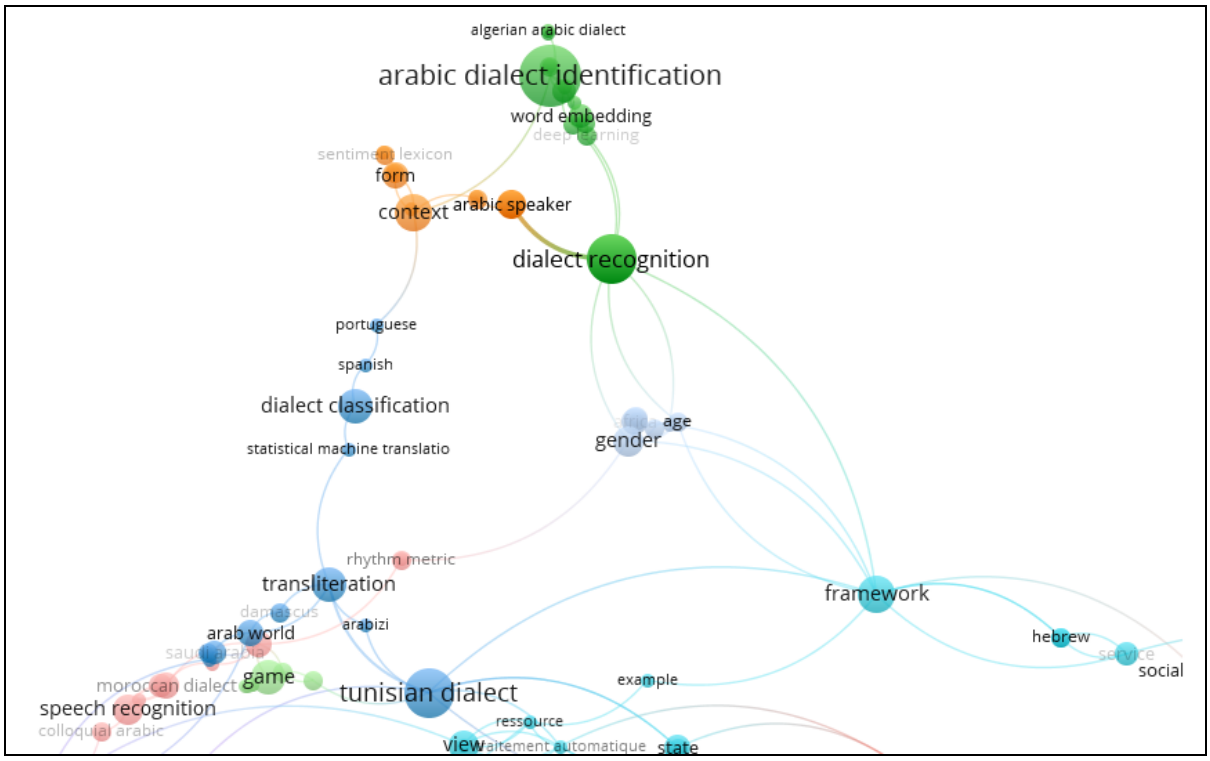


FIGURE 2 **Most used author keywords in Arabic dialects.**

This is the first review of its kind for DA studies. It is a systematic literature review (SLR) that adheres to Kitchenham and Charters's guidelines for SLRs [44] instead of performing orthographic analysis or including articles based on the detection of basic language analysis like "Morphological Analysis & POS Tagging" or "Syntax & Parsing". From among 710 articles that were acquired through the use of search strategy, this strict inclusion criterion was fulfilled by only 130 publications that were ultimately included in the SLR. The inadequacy of the current taxonomy of Arabic Dialect Identification/ detection (ADI) techniques in explaining the newly emerging detection techniques was proven as a result of the analysis of current Arabic Dialect Identification/ detection (ADI) methods. Hence, to address this issue, an improved and better taxonomy has been put forward in this paper. This taxonomy allows researchers to gain insight into the novel identification or detection methods or the existing unexplored identification/ detection methods.

This review equips the researchers with the comprehensive vision and offers insight into the shortcomings of existing justifications in opposition to ADI, thus contributing to the research area. This manuscript contributes to develop this research area by offering complete insight into ADI and the identification/ detection techniques associated with it. The significance of gaining an insight into associated issues for the comprehension of research trend in context of Arabic language processing concerning the researches pertaining to Arabic Dialect Identification/ detection has been highlighted in this systematic literature review. This SLR also highlighted the issue of lack of clarity regarding the factors of Arabic Dialect Identification/ detection. This issue has been declared as the most crucial one by IS scholars [5], [6], [20].

II. Works on Arabic Dialects

All studies reviewed in this review are listed in TABLE A of the Appendix; each of which has been analyzed in terms of the following criteria, among others: regional dialect(s), research area, type of resources utilized, ML algorithms used, and the distinctive features of the patterns that the ML algorithms were trained on.

Therefore, we will group the reviewed articles into two major categories: those that addressed resource development and others that were concerned with dialect recognition and identification. Then each category will be divided further into subgroups.

A. Building Resources

A survey of freely available Arabic corpora and language resources is available in [45]. This survey identified few speech-based DA corpora: the Tunisian Dialect Corpus (TuDiCoI) of transcribed speech which consists of 1465 railway staff utterances and 1615 client utterances [46], and the Arabic Multi-Dialect Text Corpus with its 48M tokens that were collected from 55K webpages and distributed unevenly over four major dialect areas: Gulf, Levantine, Egyptian, and North African, [47]. Furthermore, authors of

[45] identified another free corpus the Multi-Dialect Arabic Speech Parallel Corpus that is composed of 1291 sentences in MSA, 1069 of which were translated into Gulf, Levantine, and Egyptian Arabic, resulting in a total of less than 5000 sentences in the four varieties [48]. Each sentence was recorded by several male, female, young, and old speakers of the four varieties, resulting in a total of 67,132 recorded files. The transcripts of these sentences were stored in a four-language-variety parallel corpus.

An attempt reported in [49] was made to compile a large text-based corpus for typological linguistic analysis. It resulted in the Med-Typ Database a typologically annotated corpus for Mediterranean languages.

B. Building Lexicons

One of the earliest parties to develop resources for DA is the Linguistic Data Consortium. To facilitate the description and modeling of DA, Graff et al. reported in [44] the development of a lexicon of Iraqi Dialectal Arabic that specifies the pronunciation, morphology, part of speech, and English gloss of 120,000 word tokens. In another study, [50], an Egyptian Cairene Arabic lexicon was produced for natural language processing purposes. It had MSA synonyms and part of speech tags to facilitate mapping onto Cairene entries. It also has a tag for the top-ranked meaning that is acquired from the internet. In [51], the authors created a spelling corrector aimed at the Iraqi dialect. With the help of an orthographic density metric, entrant words were able to have a fine-grained ranking. An effort on updating three bilingual dictionaries aimed for English-speakers studying Iraqi, Syrian, and Moroccan dialects is described in [52]. The authors of [53] presented a Tunisian dialect text corpus and how to build a bilingual dictionary. The objective is to utilize a language model in a speech recognition system to be used by Tunisian Broadcast News. A Levantine lexicon is made while utilizing transductive learning via half annotated text, [54]. A dedicated lexicon aimed at idioms and slang sentimental keywords so that the social network data can be sentimentally analyzed is presented in [55]. A study on building Iraqi Word Net that takes into consideration an English-Iraqi dictionary, the English WordNet, and the MSA WordNet is described in [56]. Similarly, the work presented in [57] presented a Tunisian dialect WordNet, which was initially a Tunisian corpus.

C. Building Corpora and Treebanks

An NLP task to build a corpus (with multiple genres) aimed at Egyptian Arabic is carried out in [58]. Online knowledge market services, forums, blogs, and Twitter were considered for compiling the corpus data. Unsupervised parts of speech tagging, linguistically hypercorrecting, vowel-based spelling variation, dialect identification, function-based web harvesting, base phrase chunking for dialectal Arabic, and other such factors within a dialectal Arabic corpus were addressed by the study. In a similar study, [47], the question of how to collect information from the internet by building multi-dialect Arabic corpora aimed at North African, Egyptian, Levantine, and Gulf dialects is answered. The creation of a lexicon for the Tunisian dialect is developed from Tunisian broadcast news,[53]. A corpus (with multiple dialects and genres) aimed at Iraqi, Maghrebi, Levantine, Gulf, and Egyptian dialects is presented in [59]. An additional multi-dialectal corpus that considers Twitter data and aimed at seven unique dialects is described in [60]. Similarly, a corpus (having 43 thousand words) for the Palestinian dialect is developed, [61]. With machine translation in mind, [15] suggested a parallel corpus for both MSA and Algerian dialect. In [62], which presents a pilot Levantine Arabic Treebank, syntactic and morphological data were used for annotating an informal telephone speech having close to 26K words. Another treebank for the Egyptian dialect can be found in [63]. Many researchers have focused on the annotation process quality as it is a prerequisite for most high performing language tasks. Various systems for developing NLP resources aimed at Iraqi, Moroccan, Egyptian, Levantine, and other such Arabic dialects are described in [37]. When it comes to the systems, there was utilization of both MAGEAD, [64], and Buckwalter morphological analyzer and generator (BAMA), [65]. The COLABA information retrieval system was used for evaluating how well the COLOBA can handle Arabic dialects? A web application that can annotate Moroccan, Levantine, Iraqi, and Egyptian dialects is presented in [66]. The researchers focus on not only efficiency, accuracy, and speed optimization but also the data integrity and security. An Arabic online commentary dataset having 52 million words and great dialectal content is developed in [67]. There was also a discussion on the long-term annotation efforts for identifying every sentence's dialect level. Proper instructions on the detection of Arabic code switching in regard to tokens and words have been suggested in [68]. With the help of these instructions, the annotation of a corpus with a lot of Iraqi, Levantine, and Egyptian dialects that have frequent code switching to MSA was made possible. In [69], instructions were developed for the identification of how dialectal a specific text is. 'Dialectalness' was classified into three categories: a Dialectal lexeme, MSA words having dialect morphology, and MSA having non-standard orthography. Classifying and annotating Egyptian expressions (with multiple words) in a specific computational lexicon is detailed in [70]. A graphical tool to annotate Tweets in Moroccan is available in [71]. A detailed set of instructions on the annotation of an Arabic corpus with Qatari dialect is described in [45]; QALB (short for Qatar Arabic Language Bank) is the name of the corpus. The manual correction has been the epicenter of this work and learning-based Arabic error correction mechanisms should be able to get training data from it.

D. Dialect Identification and Recognition

Arabic identification has been the focus of many works such as [72]–[75], which have been suggested by [72], [73], [75], respectively. These studies help with identifying multiple dialects and MSA. With that being said, Section 2.2 comprehensively discusses dialect identification as it is the primary objective of these works.

Dialect Identification in Text

There has been implicit inclusion of dialect identification elements in a few earlier referenced research works related to machine translation [76] or text annotation [37], [67]. Standard annotation instructions for identifying when a written text switches from MSA to a Levantine or an Egyptian dialect is presented in [69]. With the help of these instructions, large data collections can be annotated to train and test NLP tasks. A supervised method aimed at sentences for distinguishing an MSA dialect from an Egyptian one is suggested in [77]. Sentence-level features can be derived using token-level labels and, alongside other meta and primary features, can be utilized for training a generative classifier aimed at predicting the right label for all sentences of the provided input text. Using this tool for the Moroccan, Levantine, and Iraqi dialects is carried out in [78]. The training and evaluation of automatic classifiers with the help of a large annotated dataset to identify Arabic dialects is described in [9]. In terms of an Arabic sentence, the diversity of its Arabic is determined as part of the task. The usage of gulf, Iraqi, Levantine, Egyptian, Maghrebi, and MSA contribute to its diversity. There was a recent suggestion of a native Bayes classifier that considers the character bi-gram model for identifying eighteen unique Arabic dialects, [79]. The usage of phonological, morphological, and lexical data for identifying the Egyptian dialect is shown in [80]. A comprehensive monolingual dataset alongside annotated dialects for identifying the Maghrebi, Iraqi, Egyptian, Gulf, and Levantine dialects is discussed [81]. As for the cross dialectical research, [82] aimed at identifying various Maghrebi dialects as well as Palestinian and Syrian Arabic.

Dialect Recognition in Speech

A factor analysis-based modeling procedure for describing what composes the super vector as per the Gauss Mixture Model aimed at identifying dialects, [83]. The data's transcript file contains information knowledge types that are used by this procedure, which works with the Syrian, Palestinian, Iraqi, Egyptian, and Emirati dialects. A system for automatic identification of a speaker's Arabic dialect (MSA, Egyptian, Levantine, Iraqi, and Gulf) using their speech sample is presented in [84]. The performance of the newly-created language recognition methods that use the speech recognition models for discriminating Arabic dialects has been researched in [85]. However, [86] suggests an automatic recognition system aimed at Arabic dialects, which include Gulf, Iraqi, Yemeni, Lebanese, Syrian, Egyptian, Algerian, Moroccan, and Tunisian. The Gaussian Mixture Models and platform Alize have been used for analyzing the standard deviation of consonantal intervals and the vocalic intervals percentage. An evaluation of the differences in super vector pre-processing aimed at identifying dialects taking into consideration phone-recognition support vector machines is presented in [87]. The study also tackled how super vector dimensions are normalized in the pre-squashing phase, how squashing functions produce difference, and how N-gram is selected for reducing supervector dimensionality. The study included Levantine, Egyptian, Gulf, and Iraqi dialects. Speech recognition aimed at the Egyptian, Tunisian, and Saudi Arabic dialects appeared in [53], [88], [89], respectively. Egyptian conversational dialect detection can be improved using MSA acoustic data, [88]. For simplifying the task, there is an automatic conversion of the MSA data into vowels before it mixes with the Egyptian conversational dialect data. Developing language models for a speech recognition system aimed at the Tunisian Broadcast News, a corpus was developed in [53]. Word error rate can also be reduced through micro-blog data with the help of the Egyptian speech recognition system suggested in [90]. A speech database that includes native speakers all over Saudi Arabia is described in [89]. It is possible for researchers to develop a speech database with unattainable dialect maps by selecting samples out of a population. A speech recognition mechanism was trained using the acquired corpus. The findings gained from the Orien-Tel project, which is a European project aimed at developing telephony databases from different parts of the Middle East and Northern Africa is detailed in [91].

III. Method

It is becoming increasingly difficult to know who is working on what and how in computational DA. This study comes to chart the field by conducting a systematic literature review that is guided by [92] and [93] and molded after [94]–[100]. It takes account of all the research that adopted a computational approach to the identification and detection of Arabic dialects and was published between 2000 and 2020. The key words used to retrieve articles for this review are: (1) 'Arabic' to exclude other languages that might be subject of investigation; (2) 'dialect' to include regional language variation and exclude variation due to age, gender, race, or profession; and (3) 'detection' or 'identification' to limit the search to computational studies that are focused on the discovery of dialects; linguistics is more focused on explanation of variation in terms of geography, age, gender, race, and profession than on spotting and classifying when an utterance belongs to a certain dialect. In this section is an outline of the methodology followed in the review, which mirrors the phrases of this research. At first, there is a description of the source databases and the adopted search strategies; then the criteria of publication inclusion or exclusion; followed by the procedure of quality assessment

of the publications; and finally, the coding and publication analysis.

TABLE I
SEARCH KEYWORDS

Query Terms
"Arabic dialects" & "Identification" or "Detection"
"Colloquial Arabic" & "Identification" or "Detection"
"Arabic vernaculars" & "Identification" or "Detection"

A. Data sources and search strategies

The sources of articles on computational Arabic dialect studies were these major databases: ACM Digital Library, Google Scholar, IEEE, ProQuest, Springer, and ScienceDirect. The search took place in September 2020. Since the search for articles is dependent on its query terms [101], ours consisted of iterations of these keywords: Arabic; dialect, colloquial, vernacular; and detection, identification (see TABLE I). The query terms retrieved a total of 710 articles as detailed in the next section.

B. Inclusion/exclusion criteria

To decide which studies to review here and which to ignore, the criteria in TABLE II were applied. The inclusion criteria were: (1) The studies must have been concerned with 'Arabic dialects'; (2) They must involve language resources whether in terms of corpora that might have been subjected to analysis or developed for that purpose; (3) The language of the paper must be English since it is the lingua franca of the natural language processing community; and (4) The date of publication must have been between April 2000 and September 2020,when dialect studies became a concern for the Arabic NLP community.

A total of 710 articles were retrieved from the keywords given above. Out of these, 123 articles were found to be duplicates, and hence, they were removed. This meant that a total of 587 articles were left for the systematic review. For every study, the inclusion and exclusion criteria were checked by the authors. It was found that 130 articles and their distribution according to most popular databases they belong to is presented in TABLE III fulfilled the inclusion criteria, and so they were included in the analysis process. The Preferred Reporting Items for Systematic Reviews and

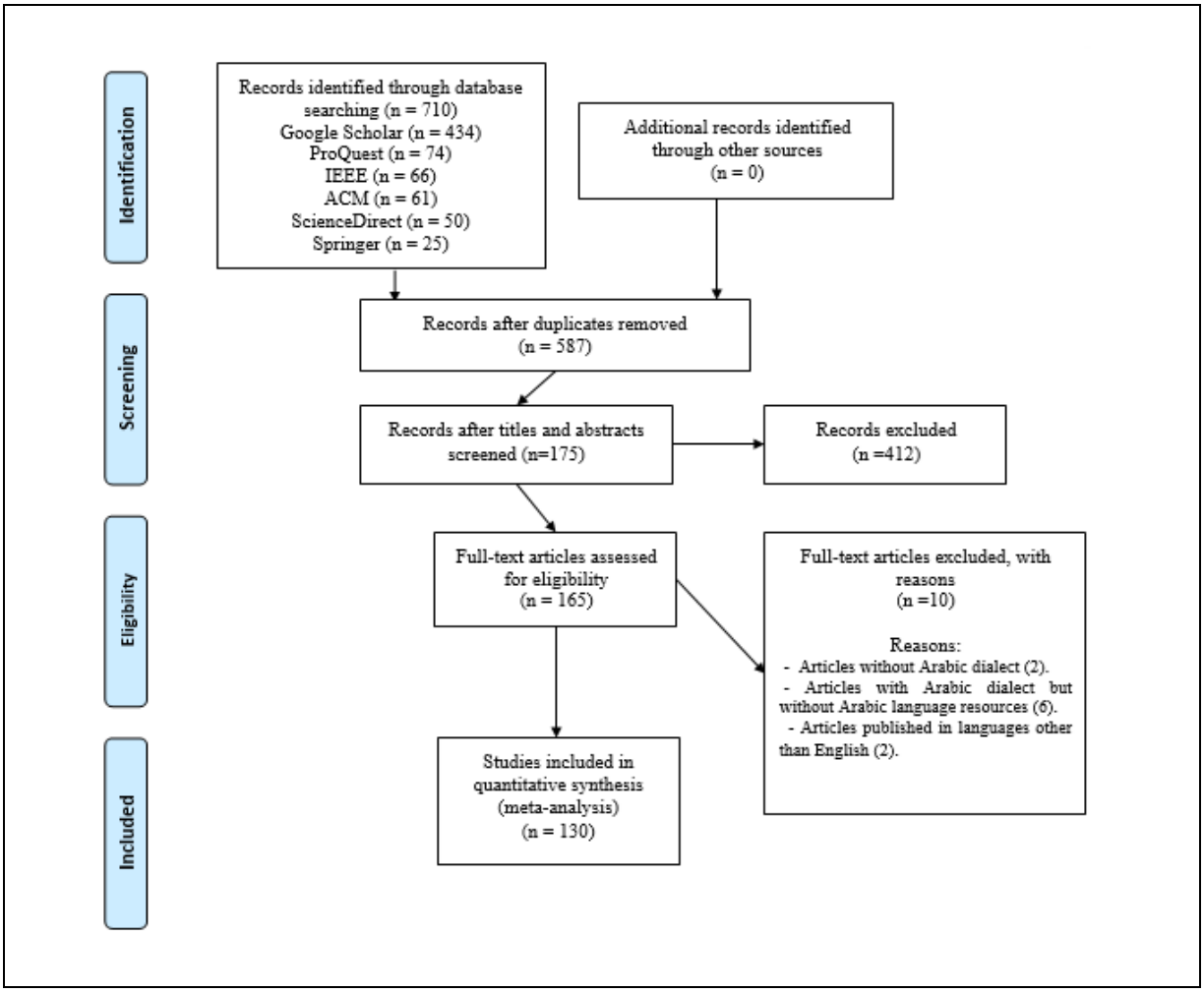


FIGURE 3 PRISMA flow diagram.

Meta-Analysis (PRISMA) was adhered to in the search and refinement phases of the review study [102]. The PRISMA flowchart can be seen in FIGURE 3.

TABLE II
EXCLUSION CRITERIA

Exclusion Criteria
Arabic dialects are not in focus.
No utilization or development of Arabic language resources.
The language medium of communication is not English.
Published before 2000.

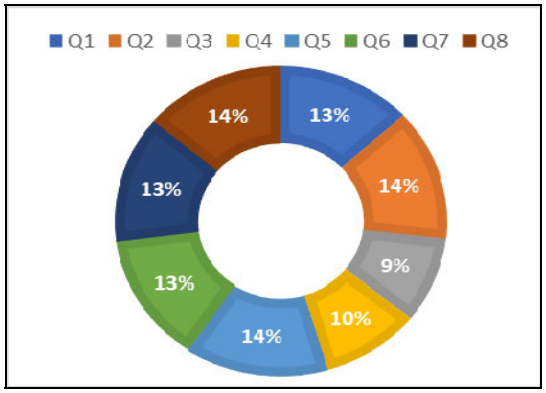


FIGURE 4 Number of studies (%) addressing quality assessment criteria.

TABLE III
FINAL SEARCH RESULTS ACROSS MOST POPULAR DATABASES.

No.	Database	Count
1	ACM	9
2	Google Scholar	69
3	IEEE	21
4	ProQuest	8
5	Springer	9
6	Science Direct	14
	Total	130

C. Quality assessment

In addition to the inclusion and exclusion criteria, the articles candidate for inclusion were subjected to quality assessment. A checklist of eight criteria was adapted from those put forward by [93] and was used to give a quality score to each of the (N=130) articles to be covered by this review. TABLE IV shows this quality assessment checklist. The purpose of the checklist was not to serve as a way of criticizing the work of any scholar [93] but rather to give the reader assurance that each of them meets the survey requirements. We used a three-point scale to give a score to each criterion, which is formulated as a question in TABLE IV. "Yes" was given 1 point, "No" was given 0 point, and "Partially" was given 0.5 point. Therefore, every article would get a score between 0 and 8; the score signifies the degree of confidence that an article meets the criteria. The quality assessment outcomes

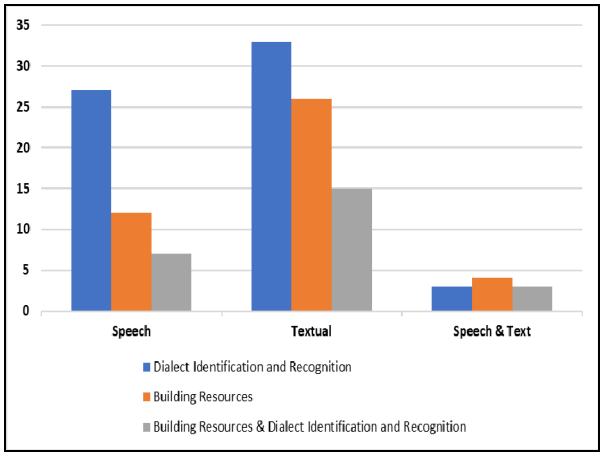


FIGURE 5 Distribution of studies per language mode.

for the 130 studies are shown in FIGURE 4. The findings show that the quality criteria have been fulfilled by all the studies; i.e., all 130 studies are qualified for further analysis.

TABLE IV
QUALITY ASSESSMENT CRITERIA [44].

No.	Question
1	Are the research aims clearly specified?
2	Was the study designed to achieve these aims?
3	Are the techniques/algorithms considered by the study clearly specified?
4	Is the study context/discipline clearly specified?
5	Are the data collection methods adequately detailed?
6	Is the machine learning technique measured and reported?
7	Do the results add to the literature?
8	Does the study add to your knowledge or understanding?

D. Data coding and analysis

The following features were coded for each article: (a) the main research area within dialect studies; (b) investigated regional dialect (e.g., Algerian, Gulf, Iraqi, North African, etc.); (c) research techniques and algorithms (e.g., artificial neural networks, Logistic Regression, etc.); (d) key machine learning input features; (e) data type; (f) datasets used; (g) year of publication; (h) evaluation criteria; and (i) place of publication.

IV. Results

The 130 studies on Arabic dialect detection and identification that were published between 2000 and 2020 constitute the corpus that we will analyze here. This section will be organized around the nine research questions that we posed at the beginning of this review. Please note that the frequency mentioned here is not in a one to one correspondence with paper titles since an article might use two machine learning algorithms, for example, and get counted as an instance of each of them. It might investigate Egyptian vis-à-vis Levantine and Gulf Arabic and be counted as an instance of each of the three regional varieties.

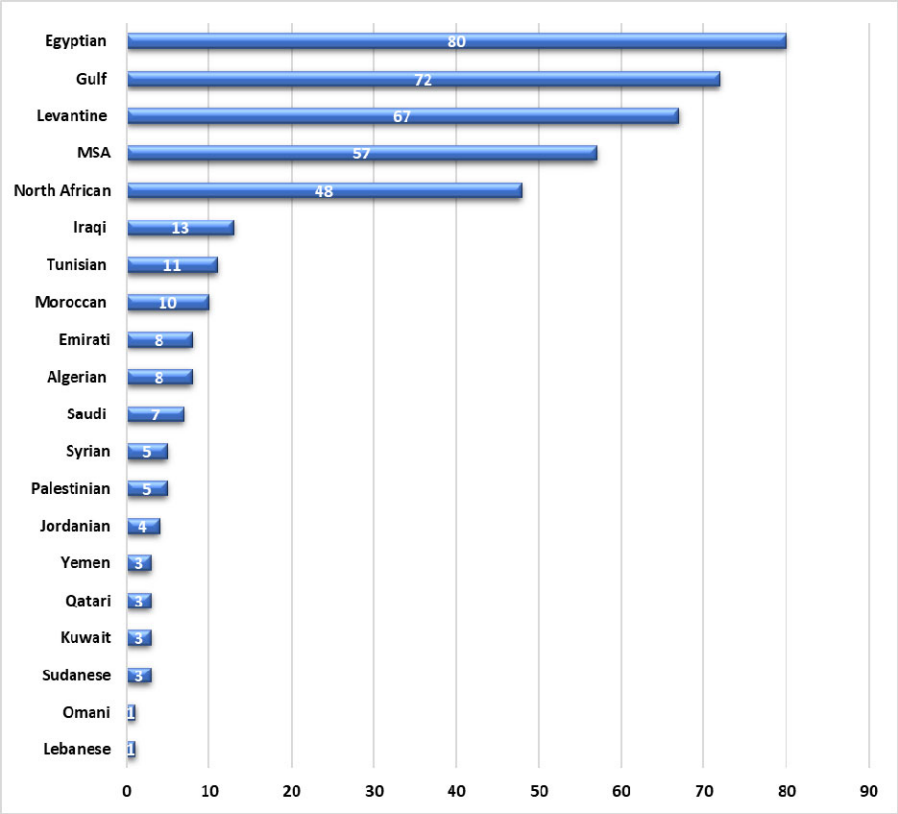


FIGURE 6 Research per country/regional dialect.

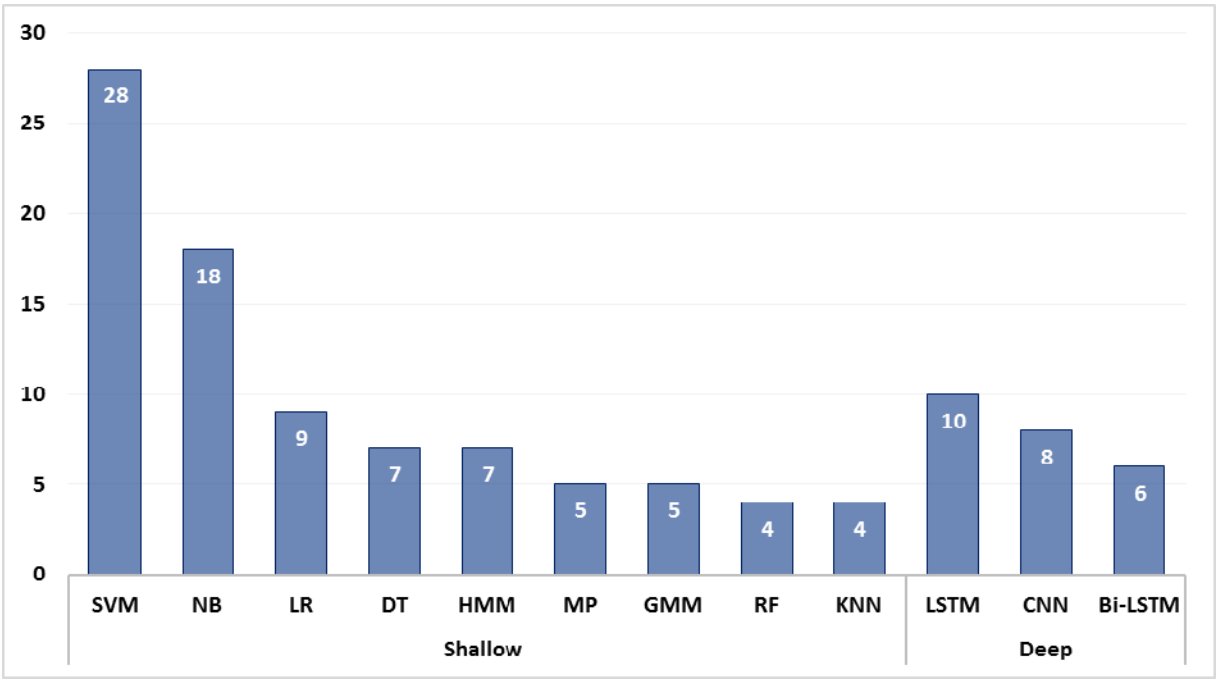


FIGURE 7 Popular machine learning algorithms.

1. *What are the key research areas in Arabic computational dialect studies (RQ1)?*

More than half the Arabic computational dialect research focused exclusively on the vernacular in texts, 38% in speech, while 7% investigated the two modes simultaneously, as shown in FIGURE 5.

TABLE V shows that less than one third of the research conducted since 2000 has been dedicated to resource building, while half of it focused on dialect identification. There is a conspicuous preference for resource development for the written language, with almost twice as many textual resources as speech.

TABLE V
STUDIES PER RESEARCH AREA.

Research Area	Speech	Textual	Text & acoustic	Count
Dialect Identification and Recognition	27	33	3	63
Building Resources	12	26	4	42
Building Resources & Dialect Identification and Recognition	7	15	3	25

2. What are the dialects of concern in the reviewed articles (RQ2)?

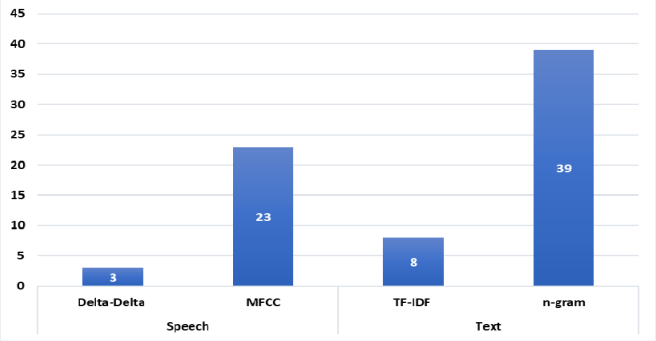


FIGURE 8 Features used in ML studies.

Dialect research is an emerging scientific pursuit in Arabic. In fact, most conservative monolingual Arabic specialists would not acknowledge such an endeavor as legitimate. Furthermore, dialect research is often framed with reference to the standard variety due to the richness of its codification and documentation. This is why Modern Standard Arabic attracted almost 14% of the instances of dialect study, often by way of comparison with country or regional dialects (see FIGURE 6). By far, Egyptian attracted the highest proportion of dialect studies (almost one fifth of those in our corpus), succeeded by two multiregional varieties (Gulf and Levantine at almost 18% and16%, respectively). The reader may be alerted to some redundancy in our classification of dialect use: Gulf Arabic usually includes Saudi and Emirati Arabic, but we had to create this category to label those studies that were concerned with the variety of the Gulf region rather than the dialects of the individual countries that make up the Gulf. The same is true of North African vis-à-vis Tunisian, Algerian, and Moroccan Arabic, and of Levantine Arabic vis-à-vis Syrian, Palestinian, Jordanian, and Lebanese.

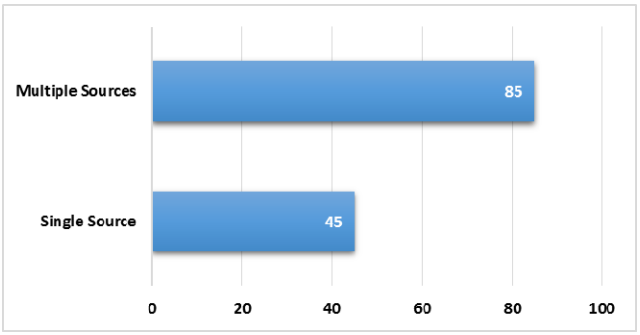


FIGURE 9 Type of data-sources used in dialect identification.

3. What are the machine learning algorithms used in Arabic dialect studies (RQ3)?

FIGURE 7 displays the popularity of machine learning algorithms in the reviewed articles. Some utilized several machine learning algorithms, hence, the frequencies in this figure stand for the instances of algorithm use, rather than number of papers. Eighty-four research techniques were identified in the corpus, but we plotted in this figure the algorithms that were used in four articles at minimum. Three quarters of the instances of ML adopted a shallow ML approach. Of all shallow models, the most widely used in Arabic computational dialect studies was "Support Vector Machine (SVM)" with one quarter of the studies adopting it. The second in popularity was "Naive Bayes (NB)" being utilized in 16% of all instances of use, followed by Logistic Regression at 8%; Decision Tree and Hidden Markov Model at 8% each; Multilayer Perceptron and Gaussian Mixture Model at 5% each; and finally, Random Forest and K-Nearest neighbor at 4% each. As for deep learning networks that were most popular in Arabic dialect research, Long Short-Term Memory was the most frequently adopted (in 9% of the instances of ML algorithms), followed by Convolutional Neural Network (in 7%) and Bidirectional Long Short-Term Memory (in 5%). See Appendix A for details.

4. What are the input features used in Arabic dialect studies (RQ4)?

FIGURE 8 shows the most popular features used in Arabic machine learning dialect studies. In textual dialect identification, the most frequently used features are n-grams, the contiguous sequences of n-items in a corpus. N-grams constitute more than four fifths of the feature instances in the reviewed literature. Second in popularity is TF-IDF, Term Frequency–Inverse Document Frequency, with 17% of the instances of use in the papers we reviewed. TF-IDF reflects the importance of a word to a document in the textual corpus.

In speech corpora, on the other hand, the features most prevalent are (1) the mel-frequency cepstral coefficient (MFCC), which was used in 88% of the times machine learning was applied to speech; and (2) Delta-Delta coefficient, which was used the rest of time.

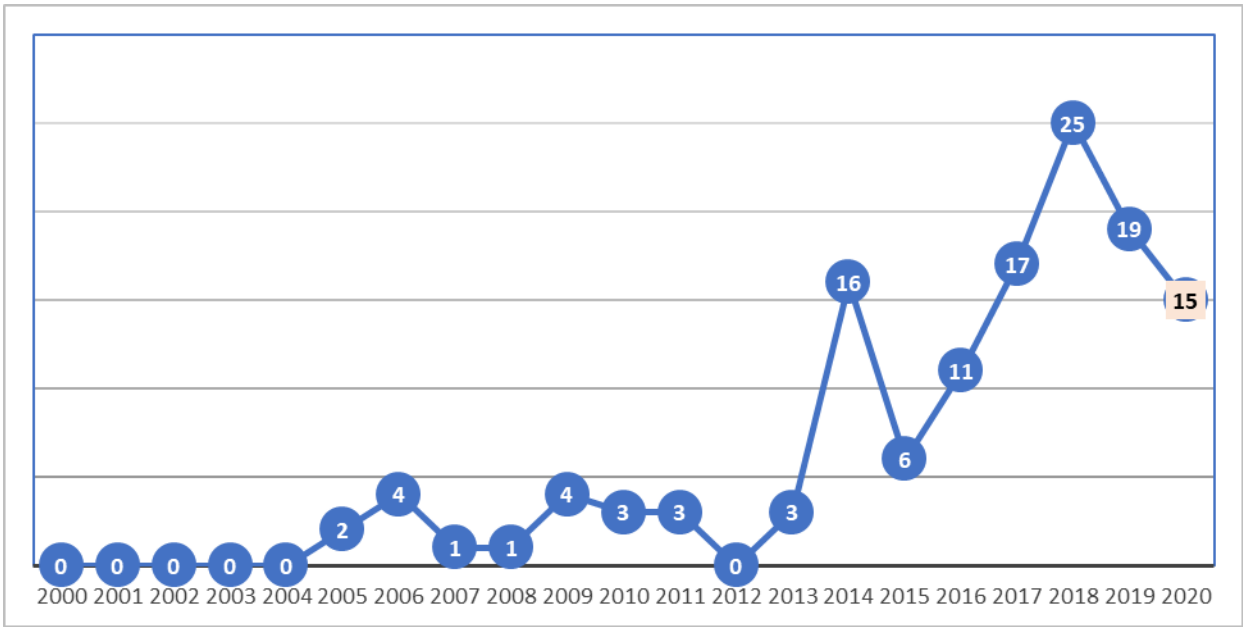


FIGURE 10 **Distribution of Arabic dialects studies per publication year.**

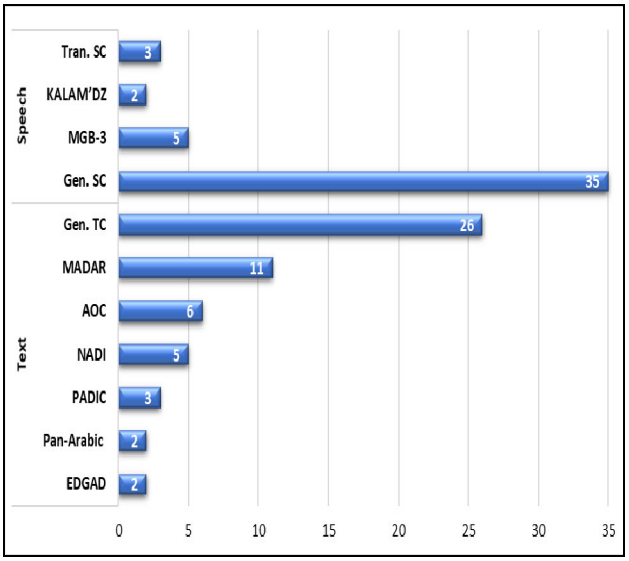


FIGURE 11 **Corpora used in the reviewed papers.**

Note that we only display here the top 4 features. Any feature that was used less than 3 times has been excluded; otherwise the graph would be too big to fit on the page.

5. *Which types of data-sources are the most widely used in Arabic dialect identification studies (RQ5)?*

It is curious to know which types of discourse were the most used in the course of dialect identification. Most studies did not bother about the discourse genre they experimented with; they were only interested in sampling dialectal language.
Others were more selective. FIGURE 9 displays the types of data utilized in the reviewed articles. Notice how data that came from multiple sources is more than two-fold the data that came from one source.
Datasets used were collected from a variety of sources. Academic articles, entertainment, health, technology, sports, and politics were used in one study each. Business and YouTube videos were used in 3 studies each. Other sources such as commentary, travel, twitter, news, and social media were used in multiple studies ranging from 7 to 14 for each. While 85 studies used datasets from multiple sources, 45 studies used a single source.

6. *What are the datasets most often utilized in Arabic dialect studies (RQ6)?*

In this section, we describe popular benchmark datasets that were used for dialectal Arabic detection. FIGURE 11 depicts corpora/datasets used in the reviewed papers. General datasets are collected by the authors of each study. For instance, we have BRAD (book reviews) [103], HARD (hotel reviews) [104], ADI17 for Fine-grained Arabic Dialect Identification (ADI) [105], Habibi (a multi Dialect multi National Arabic Song Lyrics Corpus) [106], and ArapTweet (A Large Multi-Dialect Twitter Corpus) [(LREC 2018, Miyazaki, Japan (7-12 May 2018)] [107]. However, some popular datasets have been adopted by several studies as benchmark datasets. We describe these below.

AOC:

AOC is the Arabic Online Commentary dataset, [43], that consists of the textual content of reader commentary from the online versions of three Arabic newspapers: AlGhad from Jordan, Al-Riyadh from Saudi Arabia, and Al-Youm Al-Sabe' from Egypt, newspapers from three different dialectal

regions. The comments are labeled, by crowdsourcing, for the dialectal variety that each represents. AOL comprises, in total, more than 1.1M words that make up 63K sentences in MSA and 0.85M words that make up 44K sentences in dialectal Arabic.

PADIC:

PADIC is the Parallel Arabic DIalect Corpus, [19]. It consists, in its original form, of aligned sentences in five city dialects from four Arab countries (Annaba and Algiers in Algeria, Sfax in Tunisia, Damascus in Syria, and Gaza in Palestine), in addition to MSA. It was created from hand-transcribed recordings of everyday life conversations and movie and TV dialogue scripts that were translated into the six varieties. It contains more than 37K tokens, roughly 10K word types in each of the five dialects and in their MSA version.

MADAR:

MADAR is the Multi Arabic Dialect Applications and Resources corpus, [108]. It is a parallel text corpus of 25 city dialects in 15 Arab countries. It consists of the dialectal and MSA translations of a selection of items from the Basic Travel Expression Corpus (BTEC); originally a Japanese-English parallel spoken language corpus of sentence pairs that travel phrasebooks abound with [109]. MADAR translated English sentences and expressions from this list into French, Modern Standard Arabic, and into the dialectal Arabic of the five regional varieties: Maghrebi, Nile Basin, the Levant, Gulf, and Yemen. Absent from this corpus is the dialectal Arabic of Bahrain, Comoros, Djibouti, Kuwait, Mauritania, Somalia, and the United Arab Emirates. The dataset has 12K sentences each of the Cairo, Doha, Tunisia, Rabat, and Beirut varieties but 2K sentences for the remaining 21 city varieties.

NADI:

NADI is the Nuanced Arabic Dialect Identification shared task dataset, [110]. It is a sub-country level province-labeled set of naturally occurring dialectal Arabic tweets. It consists of 21K tweets that were made by users who consistently and exclusively tweeted during a 10-month period from a single location in the geographical area that the dataset represents.

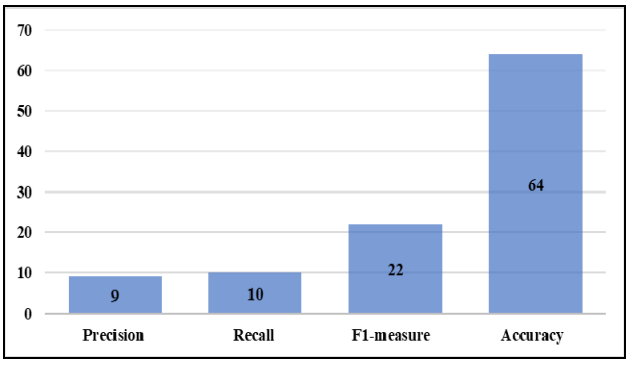


FIGURE 12 **Popularity of various evaluation criteria.**

7. *What are the trends across time in Arabic dialect identification (RQ7)?*

Computational dialect studies started in the early 2000s. The first article to appear was by [91] which reported on the compilation of the OrienTel-Telephony Database that was sponsored by the European Commission and coordinated by ScanSoft of Xerox. Ever since then, the NLP community has been developing dialectal Arabic resources. Of the pioneering contributions were speech corpora [88] , text corpora [54], [111]–[113], a tagger for Egyptian [114], and a tagger for Levantine [64], [115]. Early DA research was propped with MSA resources. For instance, [63] and [87] continued to use MSA for comparison or scaffolding.

It is curious to note the research gap in 2012. The flow of studies since 2013 has been growing steadily with no disruption at all. FIGURE 10 illustrates the distribution of Arabic dialect studies in terms of publication year. Notice that the number shown for year 2020 is provisional since the year was not over at the time of writing.

8. *What are the evaluation criteria of machine learning techniques that were used in Arabic dialect identification (RQ8)?*

It is customary now for NLP research to validate results and to test the reliability of systems and resources. Studies of dialectal Arabic mostly used one or more of these evaluation statistics: accuracy, recall, precision, and F1 [98].

Accuracy in binary dialect classification evaluates the correct identification or exclusion of dialectal language by calculating the proportion of correctly identified dialect instances (i.e., both true positives and true negatives) to the total number of instances considered (i.e., true positives, true negatives, false positives, and false negatives). It is such a popular measure that 62% of the reviewed studies adopted it.

The least popular evaluation statistic is precision, with 9% of the reviewed articles adopting it. Precision is an index of the number of correctly identified instances (i.e., true positives) divided by the total number of true positives and false positives. Slightly more popular than precision is recall,

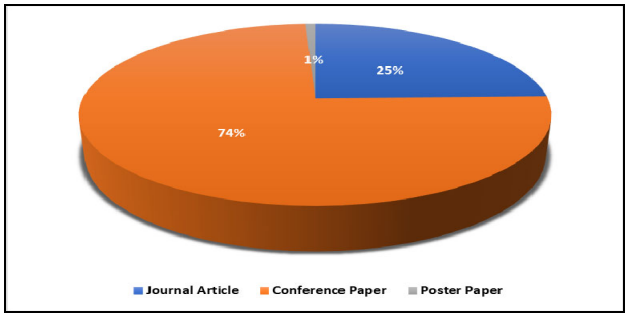


FIGURE 13 **Research Dissemination Venues.**

with 10% of studies adopting it. Recall is the number of true positives divided by the sum of true positives and false negatives.

F1 is twice as popular as recall or precision, with 20% of the reviewed articles adopting it. Perhaps this is due to the fact that it is a score that captures the harmonic mean of both precision and recall. It is an attempt at capturing measures where maximizing one would result in minimizing the other. FIGURE 12 shows the popularity of various evaluation metrics.

9. Where do the results of Arabic dialect research get published (RQ9)?

Like the general trend in computer science, three quarters of Dialectal Arabic research gets disseminated through professional conferences, rather than journal articles (see FIGURE 13). This is expected as conference publication first affords the researcher speedy dissemination of findings and opportunities for brainstorming and personal interaction with peers. Journal publication second affords more impact and citation.

V. CONCLUSIONS

Research effort has not been directed evenly to speech and text; there is some bias favoring text, but it is not as alarming in dialect identification and recognition as it is in resource development. It appears that researchers find it easier to build written language resources than spoken language ones. The reason might be related to all the requirements and setup procedures that are involved in speech resource development.

What received the most attention are regional varieties of Arabic vis-à-vis individual local vernaculars. The Arabic vernaculars spoken in the continuous Arabic language area received varying degrees of attention from computational dialect researchers. The Egyptian variety drew the most interest since almost one third of the speakers of Arabic reside in Egypt. However, what calls for attention in this language area are the Mauritanian and Bahraini varieties because they were not used in dialect identification or in resource building. Another clear gap in Arabic dialectal research is in relation to what we might term 'Enclave Arabic' or what Watson [8] calls 'Sprachinseln', i.e., language islands where a minority of Arabic speakers are surrounded by speakers of other languages. This would include Anatolian Arabic, Khuzistan Arabic, Khurasan Arabic, Uzbekistan Arabic, the Subsaharan Arabic of Nigeria and Chad, Djibouti Arabic, Somali Arabic, Comoros Arabic, and Cypriot Arabic. Maltese Arabic might also warrant research.

It would also be linguistically interesting if speech resources were developed for city dialects, similar to that in MADAR and NADI, as this could excite the compilation of a linguistic atlas for Arabic and for its individual spoken varieties. A linguistic atlas of this type would constitute a 'museum', as Jastrow [117] put it, that would encourage diachronic research and comparative studies in Semitic at large, where the questions posed by Watson [8] could be answered: "Do all modern Arabic dialects share a single unified ancestor, or do they have many different, but related, ancestors? And if they share a single ancestor, how is this ancestor related to Classical Arabic or to the ʕarabiyya and are these latter one and the same language?".

In its treatment of dialect identification and recognition, computational Arabic dialect research utilized machine learning algorithms heavily, with three quarters of the research adopting a shallow ML approach. SVM and NB were the most popular. Deep learning models, as an emerging technology, has been used, which appears to be a promising subject of research.

Another observation is that researchers seem to have favored certain linguistic features over others in dialect identification and classification neural networks. There is dominance of n-grams and substantial presence of TF-IDF in textual dialectal Arabic and there is preponderance of MFCC use as a feature in speech DA. Are these the only most rewarding measurable features? Could other language properties be as informative and discriminating?

To validate results and evaluate system reliability, the most highly used statistic in Arabic dialect identification was accuracy, which means that researchers were concerned with the degree of proximity of their classification to reality. Other important metrics used include precision, recall, and F1-score which are used particularly with imbalanced datasets. While precision measures the percentage of relevant results, recall is concerned with the percentage of total relevant results that are correctly classified. The F1-score is the harmonic mean of precision and recall.

In terms of the data used in dialect identification, it seems that there is preference for the collection of data from multiple sources over single source data. Tweets, in particular, have been favored over all other data. This might be due to the volume and the ease of data extraction that is afforded by the Twitter APIs. Also, the datasets are problematic! The text-based datasets are more common in research compared to speech-based datasets. A hybrid method that combines both speech and text would make a good research line. The use of speech is a prerequisite for developing a robust system.

All computational dialect research at the inception of this research area was concerned with speech; interest in text was subsequent to it. There is a steady growth in dialectal research since 2013. One would expect future research, though, to take up additional themes. For instance, a question that remains unanswered is what is the degree of similarity or difference between MSA, on the one hand, and the various dialectal varieties? What relationship holds between one vernacular and another? Contrastive research would also require (1) settling the issue of spelling and transcription of the vernaculars since it introduces variability; (2) studying the defining phonological, morphological, syntactic, and semantic features of each vernacular; (3) identifying the

isoglosses between dialects; (4) using current knowledge about contemporary vernaculars for historical dialectology; (5) investigating the effect of such social factors as region, age, gender, ethnicity, social class, and profession on language variation.

REFERENCES

[1] General Assembly resolution A/RES/3190(XXVIII), *Inclusion of Arabic among the official and the working languages of the General Assembly and its Main Committees*. 1973.

[2] UNESCO, "World Arabic Language Day," Paris, 2012.

[3] M. Diab and N. Habash, "Arabic dialect processing tutorial," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts*, 2007, pp. 5–6.

[4] S. Harrat, K. Meftouh, and K. Smaïli, "Maghrebi Arabic dialect processing: an overview," 2018.

[5] A. Shoufan and S. Alameri, "Natural language processing for dialectical Arabic: A Survey," in *Proceedings of the second workshop on Arabic natural language processing*, 2015, pp. 36–48.

[6] N. Y. Habash, "Introduction to Arabic natural language processing," *Synth. Lect. Hum. Lang. Technol.*, vol. 3, no. 1, pp. 1–187, 2010.

[7] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, no. 4, p. 14, 2009.

[8] J. Watson, "Arabic dialects (general article)," 2011.

[9] O. F. Zaidan and C. Callison-Burch, "Arabic dialect identi_cation," *Microsoft Res.*, 2012.

[10] R. Tachicart, K. Bouzoubaa, and H. Jaafar, "Building a Moroccan dialect electronic dictionary (MDED)," in *5th International Conference on Arabic Language Processing*, 2014, pp. 216–221.

[11] M. Elmahdy, R. Gruhn, and W. Minker, *Novel techniques for dialectal arabic speech recognition*. Springer Science & Business Media, 2012.

[12] N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh, "Morphological analysis and disambiguation for dialectal Arabic," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 426–432.

[13] I. Zribi, M. E. Khemakhem, and L. H. Belguith, "Morphological analysis of Tunisian dialect," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 992–996.

[14] W. Salloum and N. Habash, "ADAM: Analyzer for dialectal Arabic morphology," *J. King Saud Univ. Inf. Sci.*, vol. 26, no. 4, pp. 372–378, 2014.

[15] S. Harrat, K. Meftouh, M. Abbas, and K. Smaili, "Building resources for algerian arabic dialects," 2014.

[16] R. Zbib *et al.*, "Machine translation of Arabic dialects," in *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2012, pp. 49–59.

[17] W. Salloum and N. Habash, "Elissa: A dialectal to standard Arabic machine translation system," in *Proceedings of COLING 2012: Demonstration Papers*, 2012, pp. 385–392.

[18] H. Sawaf, "Arabic dialect handling in hybrid machine translation," in *Proceedings of the conference of the association for machine translation in the americas (amta), denver, colorado*, 2010.

[19] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, and K. Smaili, "Machine translation experiments on PADIC: A parallel Arabic dialect corpus," 2015.

[20] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: an overview," *J. King Saud Univ. Inf. Sci.*, 2019.

[21] V. Velupillai, *An introduction to linguistic typology*. John Benjamins Publishing Company AmsterdamPhiladelphia, 2012.

[22] P. Swiggers, *The collected works of edward sapir I: General linguistics*. 2008.

[23] C. A. Ferguson, "Diglossia," WORD, vol. 15, no. 2, pp. 325–340, Jan. 1959.

[24] A. J. Yepes, A. MacKinlay, and B. Han, "Investigating public health surveillance using twitter," in *Proceedings of BioNLP 15*, 2015, pp. 164–170.

[25] Q. C. Nguyen *et al.*, "Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity," *JMIR public Heal. Surveill.*, vol. 2, no. 2, p. e158, 2016.

[26] Q. C. Nguyen *et al.*, "Twitter-derived neighborhood characteristics associated with obesity and diabetes," *Sci. Rep.*, vol. 7, no. 1, p. 16425, 2017.

[27] M. M. Abdul-Mageed, A. Buffone, H. Peng, J. Eichstaedt, and L. Ungar, "Recognizing pathogenic empathy in social media," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[28] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 851–860.

[29] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, "Geo-located Twitter as proxy for global mobility patterns," *Cartogr. Geogr. Inf. Sci.*, vol. 41, no. 3, pp. 260–271, 2014.

[30] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth, "Understanding human mobility from Twitter," *PLoS One*, vol. 10, no. 7, p. e0131469, 2015.

[31] M. Lenormand *et al.*, "Comparing and modelling land use organization in cities," *R. Soc. open Sci.*, vol. 2, no. 12, p. 150449, 2015.

[32] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, and K. Lindén, "Automatic language identification in texts: A survey," *arXiv*, vol. 65. arXiv, pp. 675–782, Apr-2018.

[33] M. Araújo, A. Pereira, and F. Benevenuto, "A comparative study of machine translation for multilingual sentence-level sentiment analysis," *Inf. Sci. (Ny).*, vol. 512, pp. 1078–1102, 2020.

[34] S. Harrat, K. Meftouh, K. Abidi, and K. Smaïli, "Automatic

identification methods on a corpus of twenty five fine-grained Arabic dialects," in *International Conference on Arabic Language Processing*, 2019, pp. 79–92.

[35] A. Tawfik, M. Emam, K. Essam, R. Nabil, and H. Hassan, "Morphology-Aware Word-Segmentation in Dialectal Arabic Adaptation of Neural Machine Translation," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 11–17.

[36] K. C. Ryding, "Modern Standard Arabic," in *The semitic languages : an international handbook*, no. Bd 36 =, S. Weninger, G. Khan, M. P. Streck, and J. C. E. Watson, Eds. Berlin ; Boston: De Gruyter Mouton, 2011, pp. 844–850.

[37] M. Diab, N. Habash, O. Rambow, M. Altantawy, and Y. Benajiba, "COLABA: Arabic dialect annotation and processing," in *Lrec workshop on semitic language processing*, 2010, pp. 66–74.

[38] M. Abdul-Mageed, "Subjectivity and Sentiment Analysis of Arabic as a Morphologically-Rich Language," Indiana University, Bloomington, 2015.

[39] C. Holes, *Modern Arabic : structures, functions, and varieties*, Rev. Washington, D.C.: Georgetown University Press, 2004.

[40] J. Owens, *A linguistic history of Arabic*. Oxford ; New York: Oxford University Press, 2009.

[41] C. Holes, *Dialect, culture, and society in Eastern Arabia*. Leiden, The Netherlands ; Boston, Massachusetts: Brill, 2016.

[42] E.-S. M. Badawi, *Mustawayāt al-ʻArabīyah al-muʻāṣirah fī Miṣr*. Miṣr: Dār al-Maʻārif, 1973.

[43] O. Zaidan and C. Callison-Burch, "The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content. BT - The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 Ju." pp. 37–41, 2011.

[44] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[45] W. Zaghouani, "Critical survey of the freely available Arabic corpora current situation of the freely available," in *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme, LREC*, 2014.

[46] M. Graja, M. Jaoua, and L. Hadrich Belguith, "Lexical study of a spoken dialogue corpus in tunisian dialect," in *The international arab conference on information technology (acit), benghazi–libya*, 2010.

[47] K. Almeman and M. Lee, "Automatic building of arabic multi dialect text corpora by bootstrapping dialect words," in *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2013, pp. 1–6.

[48] K. Almeman, M. Lee, and A. A. Almiman, "Multi dialect Arabic speech parallel corpora," in *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2013, pp. 1–6.

[49] A. Sansò, "MED-TYP: A Typological Database for Mediterranean Languages.," in *LREC*, 2004.

[50] R. Al-Sabbagh, R., & Girju, "Mining the web for the induction of a dialectical arabic lexicon," *Lr.*, 2010.

[51] W. Rytting, C. A., Zajic, D. M., Rodrigues, P., & S. C., Hettick, C., Buckwalter, T., and C. C. Blake, "Spelling correction for dialectal arabic dictionary lookup," *ACM Trans. Asian Lang. Inf. Process. (TALIP), 10(1), 3.*, 2011.

[52] D. Graff and M. Maamouri, "Developing LMF-XML Bilingual Dictionaries for Colloquial Arabic Dialects.," in *LREC*, 2012, pp. 269–274.

[53] R. Boujelbane, M. E. Khemekhem, S. BenAyed, and L. H. Belguith, "Building bilingual lexicon to create Dialect Tunisian corpora and adapt language model," in *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, 2013, pp. 88–93.

[54] K. Duh and K. Kirchhoff, "Lexicon acquisition for dialectal Arabic using transductive learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 399–407.

[55] A. R. Hedar and M. Doss, "Mining social networks arabic slang comments," in *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2013.

[56] V. Cavalli-Sforza, H. Saddiki, K. Bouzoubaa, L. Abouenour, M. Maamouri, and E. Goshey, "Bootstrapping a wordnet for an arabic dialect from other wordnets and dictionary resources," in *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, 2013, pp. 1–8.

[57] R. Bouchlaghem, A. Elkhlifi, and R. Faiz, "Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014, pp. 104–113.

[58] R. Al-Sabbagh and R. Girju, "YADAC: Yet another Dialectal Arabic Corpus.," in *LREC*, 2012, pp. 2882–2889.

[59] R. Cotterell and C. Callison-Burch, "A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic.," in *LREC*, 2014, pp. 241–245.

[60] H. Mubarak and K. Darwish, "Using Twitter to collect a multi-dialectal corpus of Arabic," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014, pp. 1–7.

[61] M. Jarrar, N. Habash, F. Alrimawi, D. Akra, and N. Zalmout, "Curras: an annotated corpus for the Palestinian Arabic dialect," *Lang. Resour. Eval.*, vol. 51, no. 3, pp. 745–775, 2017.

[62] M. Maamouri *et al.*, "Developing and Using a Pilot Dialectal Arabic Treebank.," in *LREC*, 2006, pp. 443–448.

[63] M. Maamouri, A. Bies, S. Kulick, M. Ciul, N. Habash, and R. Eskander, "Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development.," in *LREC*, 2014, pp. 2348–2354.

[64] N. Habash and O. Rambow, "MAGEAD: a morphological analyzer and generator for the Arabic dialects," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 681–688.

[65] T. Buckwalter, "Buckwalter arabic morphological analyzer version 2.0. linguistic data consortium, University of Pennsylvania, 2002. ldc cat alog no," Ldc2004l02. Technical

report, 2004.

[66] Y. Benajiba and M. Diab, "A web application for dialectal Arabic text annotation," in *Editors & Workshop Chairs*, 2010, p. 91.

[67] O. F. Zaidan and C. Callison-Burch, "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: short papers-Volume 2*, 2011, pp. 37–41.

[68] H. Elfardy and M. T. Diab, "Simplified guidelines for the creation of Large Scale Dialectal Arabic Annotations.," in *LREC*, 2012, pp. 371–378.

[69] N. Habash, O. Rambow, M. Diab, and R. Kanjawi-Faraj, "Guidelines for annotation of Arabic dialectness," in *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, 2008, pp. 49–53.

[70] A. Hawwari, M. Attia, and M. Diab, "A framework for the classification and annotation of multiword expressions in dialectal arabic," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014, pp. 48–56.

[71] S. Tratz, D. Briesch, J. Laoudi, and C. Voss, "Tweet conversation annotation tool with a focus on an arabic dialect, moroccan darija," in *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 2013, pp. 135–139.

[72] A. Ali *et al.*, "Automatic dialect detection in arabic broadcast speech," *arXiv Prepr. arXiv1509.06928*, 2015.

[73] M. El-Haj, P. Rayson, and M. Aboelezz, "Arabic dialect identification in the context of bivalency and code-switching," in *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.*, 2018, pp. 3622–3627.

[74] S. Shon, A. Ali, and J. Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," *arXiv Prepr. arXiv1803.04567*, 2018.

[75] R. Tachicart, K. Bouzoubaa, S. L. Aouragh, and H. Jaafa, "Automatic identification of Moroccan colloquial Arabic," in *International Conference on Arabic Language Processing*, 2017, pp. 201–214.

[76] H. Soltau, L. Mangu, and F. Biadsy, "From modern standard arabic to levantine asr: Leveraging gale for dialects," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 2011, pp. 266–271.

[77] H. Elfardy and M. Diab, "Sentence level dialect identification in Arabic," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, vol. 2, pp. 456–461.

[78] H. Elfardy and M. Diab, "Aida: Automatic identification and glossing of dialectal arabic," in *Proceedings of the 16th eamt conference (project papers)*, 2012, p. 83.

[79] F. Sadat, F. Kazemi, and A. Farzindar, "Automatic identification of arabic dialects in social media," in *Proceedings of the first international workshop on Social media retrieval and analysis*, 2014, pp. 35–40.

[80] K. Darwish, H. Sajjad, and H. Mubarak, "Verifiably effective arabic dialect identification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1465–1468.

[81] O. F. Zaidan and C. Callison-Burch, "Arabic dialect identification," *Comput. Linguist.*, vol. 40, no. 1, pp. 171–202, 2014.

[82] S. Harrat, K. Meftouh, M. Abbas, S. Jamoussi, M. Saad, and K. Smaili, "Cross-dialectal arabic processing," in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2015, pp. 620–632.

[83] Y. Lei and J. H. L. Hansen, "Factor analysis-based information integration for Arabic dialect identification," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4337–4340.

[84] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, 2009, pp. 53–61.

[85] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer, and A. Mandal, "Effective Arabic dialect classification using diverse phonotactic models," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[86] M. Belgacem, G. Antoniadis, and L. Besacier, "Automatic Identification of Arabic Dialects.," in *LREC*, 2010.

[87] Q. Zhang, H. Bořil, and J. H. L. Hansen, "Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7363–7367.

[88] K. Kirchhoff and D. Vergyri, "Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition," *Speech Commun.*, vol. 46, no. 1, pp. 37–51, 2005.

[89] M. Alghamdi, F. Alhargan, M. Alkanhal, A. Alkhairy, M. Eldesouki, and A. Alenazi, "Saudi accented Arabic voice bank," *J. King Saud Univ. Inf. Sci.*, vol. 20, pp. 45–64, 2008.

[90] A. Ali, H. Mubarak, and S. Vogel, "Advances in dialectal arabic speech recognition: A study using twitter to improve egyptian asr," in *International Workshop on Spoken Language Translation (IWSLT 2014)*, 2014.

[91] D. J. Iskra *et al.*, "OrienTel-Telephony Databases Across Northern Africa and the Middle East.," in *LREC*, 2004.

[92] K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes, "Five steps to conducting a systematic review," *J. R. Soc. Med.*, vol. 96, no. 3, pp. 118–121, 2003.

[93] S. Kitchenham, B, Charters, "Guidelines for performing systematic literature reviews in software engineering.," *Softw. Eng. Group, Sch. Comput. Sci. Math. Keele Univ. 1–57.*, 2007.

[94] O. G. Iroju and J. O. Olaleke, "A systematic review of natural language processing in healthcare," *Int. J. Inf. Technol. Comput. Sci.*, vol. 8, pp. 44–50, 2015.

[95] L. Gutiérrez and B. Keith, "A Systematic Literature Review on Word Embeddings," in *International Conference on Software Process Improvement*, 2018, pp. 132–141.

[96] W. Alabbas, H. M. Al-Khateeb, and A. Mansour, "Arabic text

classification methods: Systematic literature review of primary studies," in *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, 2016, pp. 361–367.

[97] M. A. Ibrahim and N. Salim, "OPINION ANALYSIS FOR TWITTER AND ARABIC TWEETS: A SYSTEMATIC LITERATURE REVIEW.," *J. Theor. Appl. Inf. Technol.*, vol. 56, no. 3, 2013.

[98] H. Al-Mahmoud and M. Al-Razgan, "Arabic text mining a systematic review of the published literature 2002-2014," in *2015 International Conference on Cloud Computing (ICCC)*, 2015, pp. 1–7.

[99] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: a systematic review," *Comput. Sci. Rev.*, vol. 29, pp. 21–43, 2018.

[100] S. A. Pitchay and F. Ridzuan, "A Systematic Review Analysis for Quran Verses Retrieval," *J. Eng. Appl. Sci.*, vol. 100, no. 3, pp. 629–634, 2016.

[101] V. Costa and S. Monteiro, "Knowledge processes, absorptive capacity and innovation: A mediation analysis," *Knowl. Process Manag.*, vol. 23, no. 3, pp. 207–218, 2016.

[102] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *Ann. Intern. Med.*, vol. 151, no. 4, pp. 264–269, 2009.

[103] A. Elnagar and O. Einea, "Brad 1.0: Book reviews in arabic dataset," in *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, 2016, pp. 1–8.

[104] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel Arabic-reviews dataset construction for sentiment analysis applications," in *Intelligent Natural Language Processing: Trends and Applications*, Springer, 2018, pp. 35–52.

[105] "No Title." [Online]. Available: https://groups.csail.mit.edu/sls/downloads/adi17/.

[106] M. El-Haj, "Habibi-a multi Dialect multi National Arabic Song Lyrics Corpus."

[107] W. Zaghouani and A. Charfi, "Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification," *arXiv Prepr. arXiv1808.07674*, 2018.

[108] H. Bouamor *et al.*, "The madar Arabic dialect corpus and lexicon," *Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval.*, pp. 3387–3396, 2019.

[109] T. Takezawa, Q. Huo, B. Ma, E.-S. Chng, and H. Li, "Multilingual Spoken Language Corpus Development for Communication Research," 2006, pp. 781–791.

[110] M. Abdul-Mageed, C. Zhang, H. Bouamor, and N. Habash, "NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task," in *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020), Barcelona, Spain*, 2020.

[111] C. David, M. Diab, N. Habash, O. Rambow, and S. Shareef, "Parsing Arabic dialects," in *Proceedings of EACL*, 2006, pp. 369–376.

[112] D. Graff, T. Buckwalter, M. Maamouri, and H. Jin, "Lexicon Development for Varieties of Spoken Colloquial Arabic.," in *LREC*, 2006, pp. 999–1004.

[113] M. Maamouri *et al.*, "Developing and Using a Pilot Dialectal Arabic Treebank. BT - Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006." pp. 443–448, 2006.

[114] K. Duh and K. Kirchhoff, "POS tagging of dialectal Arabic," no. June, p. 55, 2005.

[115] N. Habash and O. Rambow, "Morphophonemic and orthographic rules in a multi- dialectal morphological analyzer and generator for Arabic verbs," *ISCAL-2007 1st Int. Sysmposium Comput. Arab. Lang.*, no. February 2015, 2007.

[116] M. Elmahdy, R. Gruhn, W. Minker, and S. Abdennadher, "Modern standard Arabic based multilingual approach for dialectal Arabic speech recognition," in *2009 Eighth International Symposium on Natural Language Processing*, 2009, pp. 169–174.

[117] "Jastrow, O. 2002 Arabic dialectology: the state of the art. In: Shlomo Izre'el (ed.). Semitic linguistics: The state of the art at the turn of the 21st century (Israel Orienal Studies 20. Winona Lake: Eisenbrauns) 347363. - Google Search." .

[118] K. Almeman, "Automatically Building VoIP Speech Parallel Corpora for Arabic Dialects," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 1, p. 4, 2017.

[119] A. Masmoudi, M. E. Khmekhem, M. Khrouf, and L. H. Belguith, "Transliteration of Arabizi into Arabic Script for Tunisian Dialect," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 2, p. 32, 2019.

[120] A. M. Qaroush, A. Hanani, B. Jaber, M. Karmi, and B. Qamhiyeh, "Automatic Spoken Customer Query Identification for Arabic Language," in *Proceedings of the 2016 8th International Conference on Information Management and Engineering*, 2016, pp. 41–46.

[121] L. Lulu and A. Elnagar, "Automatic Arabic Dialect Classification Using Deep Learning Models," *Procedia Comput. Sci.*, vol. 142, pp. 262–269, 2018.

[122] M. N. Al-Gedawy, "Detecting Egyptian Dialect Microblogs using a Boosted PSO-based Fuzzier," *Egypt. Comput. Sci. J.*, vol. 39, no. 1, 2015.

[123] S. Wray and A. Ali, "Crowdsource a little to label a lot: Labeling a speech corpus of dialectal Arabic," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[124] H. Bouamor, N. Habash, and K. Oflazer, "A Multidialectal Parallel Corpus of Arabic.," in *LREC*, 2014, pp. 1240–1245.

[125] A. E. Bulut, Q. Zhang, C. Zhang, F. Bahmaninezhad, and J. H. L. Hansen, "UTD-CRSS submission for MGB-3 Arabic dialect identification: Front-end and back-end advancements on broadcast speech," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 360–367.

[126] Y. Samih *et al.*, "A neural architecture for dialectal Arabic segmentation," in *Proceedings of the Third Arabic Natural Language Processing Workshop*, 2017, pp. 46–54.

[127] G. Liu, Y. Lei, and J. H. L. Hansen, "Dialect identification: Impact of differences between read versus spontaneous speech," in *2010 18th European Signal Processing Conference*, 2010, pp. 2003–2006.

[128] W. Alabbas, H. M. al-Khateeb, A. Mansour, G. Epiphaniou, and I. Frommholz, "Classification of colloquial Arabic tweets in real-time to detect high-risk floods," in *2017 International Conference On Social Media, Wearable And Web Analytics (Social Media)*, 2017, pp. 1–8.

[129] C. R. Voss, S. Tratz, J. Laoudi, and D. M. Briesch, "Finding Romanized Arabic Dialect in Code-Mixed Tweets.," in *LREC*, 2014, pp. 2249–2253.

[130] F. Biadsy and J. Hirschberg, "Using prosody and phonotactics in arabic dialect identification," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[131] F. Biadsy, H. Soltau, L. Mangu, J. Navratil, and J. B. Hirschberg, "Discriminative phonotactics for dialect recognition using context-dependent phone classifiers," 2010.

[132] J. Younes and E. Souissi, "A quantitative view of Tunisian dialect electronic writing," in *5th International Conference on Arabic Language Processing*, 2014, pp. 63–72.

[133] M. Elmahdy, M. Hasegawa-Johnson, and E. Mustafawi, "A transfer learning approach for under-resourced Arabic dialects speech recognition," in *Workshop on Less Resourced Languages, new technologies, new challenges and opportunities (LTC 2013)*, 2013, pp. 60–64.

[134] S. C. Tratz, "Accurate arabic script language/dialect classification," ARMY RESEARCH LAB ADELPHI MD, 2014.

[135] E. J. Harfash and A. H. Abdul-kareem, "Automatic Arabic Dialect Classification," *Int. J. Comput. Appl.*, vol. 975, p. 8887.

[136] R. Ziedan, M. Micheal, A. Alsammak, M. Mursi, and A. Elmaghraby, "A Unified Approach for Arabic Language Dialect Detection," in *29th International Conference on Computers Applications in Industry and Engineering (CAINE 2016)*, 2016.

[137] A. Alshutayri and E. Atwell, "Creating an Arabic Dialect Text Corpus by Exploring Twitter, Facebook, and Online Newspapers," in *Proceedings of OSACT'2018 Open-Source Arabic Corpora and Processing Tools*, 2018.

[138] A. Alshutayri and E. Atwell, "A social media corpus of Arabic dialect text," *Comput. Commun. Soc. Media Corpora. Clermont-Ferrand Press. Univ. Blaise Pascal*, 2019.

[139] A. Alshutayri and E. Atwell, "Arabic dialects annotation using an online game," in *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, 2018, pp. 1–5.

[140] K. Lounnas, M. Abbas, and M. Lichouri, "Building a Speech Corpus based on Arabic Podcasts for Language and Dialect Identification," in *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, 2019, pp. 54–58.

[141] A. Alshutayri and E. Atwell, "Classifying Arabic dialect text in the Social Media Arabic Dialect Corpus (SMADC)," in *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, 2019, pp. 51–59.

[142] B. Talafha, A. Fadel, M. Al-Ayyoub, Y. Jararweh, A.-S. Mohammad, and P. Juola, "Team JUST at the MADAR Shared Task on Arabic Fine-Grained Dialect Identification," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 285–289.

[143] K. Meftouh, K. Abidi, S. Harrat, and K. Smaïli, "The SMarT Classifier for Arabic Fine-Grained Dialect Identification," 2019.

[144] A. Hanani, A. Qaroush, and S. Taylor, "Classifying ASR transcriptions according to Arabic dialect," in *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, 2016, pp. 126–134.

[145] L. Beltaifa-Zouari and A. Chayeh, "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY SPEAKER RECOGNITION OF MAGHREB DIALECTS."

[146] A. O. O. Alshutayri and E. Atwell, "Exploring Twitter as a source of an Arabic dialect corpus," *Int. J. Comput. Linguist.*, vol. 8, no. 2, pp. 37–44, 2017.

[147] M. Al-Badrashiny, H. Elfardy, and M. Diab, "Aida2: A hybrid approach for token and sentence level dialect identification in arabic," in *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 2015, pp. 42–51.

[148] H. Bouamor *et al.*, "The MADAR Arabic dialect corpus and lexicon," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[149] M. Abdul-Mageed, H. Alhuzali, and M. Elaraby, "You tweet what you speak: A city-level dataset of arabic dialects," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[150] I. Alsarsour, E. Mohamed, R. Suwaileh, and T. Elsayed, "Dart: A large dataset of dialectal arabic tweets," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[151] S. Wray, "Classification of closely related sub-dialects of Arabic using support-vector machines," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[152] S. Khurana and A. Ali, "QCRI advanced transcription system (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 292–298.

[153] O. Obeid, M. Salameh, H. Bouamor, and N. Habash, "ADIDA: Automatic Dialect Identification for Arabic," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 6–11.

[154] M. Hassine, L. Boussaid, and H. Massaoud, "Tunisian dialect recognition based on hybrid techniques.," *Int. Arab J. Inf. Technol.*, vol. 15, no. 1, pp. 58–65, 2018.

[155] K. Alrifai, G. Rebdawi, and N. Ghneim, "Arabic Tweeps Gender and Dialect Prediction.," in *CLEF (Working Notes)*, 2017.

[156] C. Tillmann, S. Mansour, and Y. Al-Onaizan, "Improved sentence-level arabic dialect classification," in *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, 2014, pp. 110–119.

[157] F. Sadat, F. Kazemi, and A. Farzindar, "Automatic identification of arabic language varieties and dialects in social media," in *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, 2014, pp. 22–27.

[158] G. de Francony, V. Guichard, P. Joshi, H. Afli, and A. Bouchekif, "Hierarchical Deep Learning for Arabic Dialect Identification," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 249–253.

[159] A. Ragab *et al.*, "Mawdoo3 AI at MADAR Shared Task: Arabic Fine-Grained Dialect Identification with Ensemble Learning," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 244–248.

[160] P. Mishra and V. Mujadia, "Arabic Dialect Identification for Travel and Twitter Text," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 234–238.

[161] D. Ghoul and G. Lejeune, "MICHAEL: Mining Character-level Patterns for Arabic Dialect Identification (MADAR Challenge)," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 229–233.

[162] Y. Fares *et al.*, "Arabic Dialect Identification with Deep Learning and Hybrid Frequency Based Features," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 224–228.

[163] P. Přibáň and S. Taylor, "ZCU-NLP at MADAR 2019: Recognizing Arabic Dialects," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 208–213.

[164] M. Elaraby and M. Abdul-Mageed, "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)*, 2018, pp. 263–274.

[165] E. Michon, M. Q. Pham, J. M. Crego, and J. Senellart, "Neural network architectures for Arabic dialect identification," in *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, 2018, pp. 128–136.

[166] M. Ali, "Character level convolutional neural network for Arabic dialect identification," in *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, 2018, pp. 122–127.

[167] A. M. Butnaru and R. T. Ionescu, "UnibucKernel Reloaded: First place in Arabic dialect identification for the second year in a row," *arXiv Prepr. arXiv1805.04876*, 2018.

[168] S. Bougrine, A. Chorana, A. Lakhdari, and H. Cherroun, "Toward a Web-based Speech Corpus for Algerian Dialectal Arabic Varieties," in *Proceedings of the Third Arabic Natural Language Processing Workshop*, 2017, pp. 138–146.

[169] S. Malmasi and M. Zampieri, "Arabic dialect identification using iVectors and ASR transcripts," in *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 2017, pp. 178–183.

[170] M. Eldesouki, F. Dalvi, H. Sajjad, and K. Darwish, "Qcri@ dsl 2016: Spoken arabic dialect identification using textual features," in *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, 2016, pp. 221–226.

[171] R. T. Ionescu and M. Popescu, "UnibucKernel: An approach for Arabic dialect identification based on multiple string kernels," in *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, 2016, pp. 135–144.

[172] S. Malmasi and M. Zampieri, "Arabic dialect identification in speech transcripts," in *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, 2016, pp. 106–113.

[173] W. Adouane, N. Semmar, R. Johansson, and V. Bobicev, "Automatic detection of arabicized berber and arabic varieties," in *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, 2016, pp. 63–72.

[174] A. S. M. B. A. WAZIR and J. H. CHUAH, "Spoken Arabic Digits Recognition Using Deep Learning," in *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, 2019, pp. 339–344.

[175] M. Moftah, M. W. Fakhr, and S. El Ramly, "Arabic dialect identification based on motif discovery using GMM-UBM with different motif lengths," in *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, 2018, pp. 1–6.

[176] S. Shon, A. Ali, and J. Glass, "MIT-QCRI Arabic dialect identification system for the 2017 multi-genre broadcast challenge," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 374–380.

[177] S. Bougrine, H. Cherroun, and D. Ziadi, "Prosody-based spoken Algerian Arabic dialect identification," *Procedia Comput. Sci.*, vol. 128, pp. 9–17, 2018.

[178] C. Zhang, Q. Zhang, and J. H. L. Hansen, "Semi-supervised Learning with Generative Adversarial Networks for Arabic Dialect Identification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5986–5990.

[179] M. Najafian, S. Khurana, S. Shan, A. Ali, and J. Glass, "Exploiting convolutional neural networks for phonotactic based dialect identification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5174–5178.

[180] M. Al-Ayyoub, M. K. Rihani, N. I. Dalgamoni, and N. A. Abdulla, "Spoken Arabic dialects identification: The case of Egyptian and Jordanian dialects," in *2014 5th International Conference on Information and Communication Systems (ICICS)*, 2014, pp. 1–6.

[181] Y. A. Alotaibi, A. H. Meftah, S.-A. Selouani, and Y. M. Seddiq, "Speaker environment classification using rhythm metrics in Levantine Arabic dialect," in *2014 9th International Symposium on Communication Systems, Networks & Digital Sign (CSNDSP)*, 2014, pp. 706–709.

[182] M. A. Al-Walaie and M. B. Khan, "Arabic dialects classification using text mining techniques," in *2017 International Conference on Computer and Applications (ICCA)*, 2017, pp. 325–329.

[183] A. Alshutayri and H. Albarhamtoshy, "Arabic spoken language identification system (aslis): A proposed system to identifying modern standard arabic (msa) and egyptian dialect," in *International Conference on Informatics Engineering and Information Science*, 2011, pp. 375–385.

[184] J. Younes, H. Achour, and E. Souissi, "Constructing linguistic

resources for the Tunisian dialect using textual user-generated contents on the social web," in *International Conference on Web Engineering*, 2015, pp. 3–14.

[185] M. Hassine, L. Boussaid, and H. Messaoud, "Maghrebian dialect recognition based on support vector machines and neural network classifiers," *Int. J. Speech Technol.*, vol. 19, no. 4, pp. 687–695, 2016.

[186] N. Al-Twairesh *et al.*, "Suar: Towards building a corpus for the Saudi dialect," *Procedia Comput. Sci.*, vol. 142, pp. 72–82, 2018.

[187] S. Hussein, M. Farouk, and E. Hemayed, "Gender identification of Egyptian dialect in Twitter," *Egypt. Informatics J.*, vol. 20, no. 2, pp. 109–116, 2019.

[188] E. Zarrouk, Y. BenAyed, and F. Gargouri, "Graphical Models for Multi-Dialect Arabic Isolated Words Recognition," *Procedia Comput. Sci.*, vol. 60, pp. 508–516, 2015.

[189] M. A. Menacer, O. Mella, D. Fohr, D. Jouvet, D. Langlois, and K. Smaïli, "Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect," *Procedia Comput. Sci.*, vol. 117, pp. 81–88, 2017.

[190] K. Almeman, "The Building and Evaluation of a Mobile Parallel Multi-Dialect Speech Corpus for Arabic," *Procedia Comput. Sci.*, vol. 142, pp. 166–173, 2018.

[191] M. Lichouri, M. Abbas, A. A. Freihat, and D. E. H. Megtouf, "Word-Level vs Sentence-Level Language Identification: Application to Algerian and Arabic Dialects," *Procedia Comput. Sci.*, vol. 142, pp. 246–253, 2018.

[192] B. Mouaz, B. H. Abderrahim, and E. Abdelmajid, "Speech Recognition of Moroccan Dialect Using Hidden Markov Models," *Procedia Comput. Sci.*, vol. 151, pp. 985–991, 2019.

[193] I. Shahin, A. B. Nassif, and M. Bahutair, "Emirati-accented speaker identification in each of neutral and shouted talking environments," *Int. J. Speech Technol.*, vol. 21, no. 2, pp. 265–278, 2018.

[194] I. Shahin, A. B. Nassif, and S. Hamsa, "Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network," *IEEE Access*, vol. 7, pp. 26777–26787, 2019.

[195] I. Shahin and A. B. Nassif, "Emirati-Accented Speaker Identification in Stressful Talking Conditions," *arXiv Prepr. arXiv1909.13070*, 2019.

[196] I. Shahin and M. N. Ba-Hutair, "Emarati speaker identification," in *2014 12th International Conference on Signal Processing (ICSP)*, 2014, pp. 488–493.

[197] I. Shahin, "Text-Independent Emirati-Accented Speaker Identification in Emotional Talking Environment," in *2018 Fifth HCT Information Technology Trends (ITT)*, 2018, pp. 257–262.

[198] F. Mezzoudj, M. Loukam, and F. Z. Belkredim, "Arabic Algerian Oranee Dialectal Language Modelling Oriented Topic," *Int. J. Informatics Appl. Math.*, vol. 2, no. 2, pp. 1–14.

[199] K. A. Kwaik, M. Saad, S. Chatzikyriakidis, and S. Dobnik, "Shami: A corpus of levantine arabic dialects," 2018.

[200] K. Meftouh, S. Harrat, and K. Smaïli, "PADIC: extension and new experiments," 2018.

[201] M. Saad and B. O. Alijla, "Wikidocsaligner: An off-the-shelf Wikipedia documents alignment tool," in *2017 Palestinian International Conference on Information and Communication Technology (PICICT)*, 2017, pp. 34–39.

[202] A. Erdmann, N. Zalmout, and N. Habash, "Addressing noise in multidialectal word embeddings," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 558–565.

[203] K. Abidi, M. A. Menacer, and K. Smaili, "CALYOU: A comparable spoken Algerian corpus harvested from youtube," 2017.

[204] R. Suwaileh, M. Kutlu, N. Fathima, T. Elsayed, and M. Lease, "ArabicWeb16: A New Crawl for Today's Arabic Web," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 673–676.

[205] S. Khalifa, N. Habash, D. Abdulrahim, and S. Hassan, "A large scale corpus of Gulf Arabic," *arXiv Prepr. arXiv1609.02960*, 2016.

[206] G. Kumar, Y. Cao, R. Cotterell, C. Callison-Burch, D. Povey, and S. Khudanpur, "Translations of the CALLHOME Egyptian Arabic corpus for conversational speech translation," in *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, 2014.

[207] L. Abdel-Hamid, "Egyptian Arabic Speech Emotion Recognition using Prosodic, Spectral and Wavelet Features," *Speech Commun.*, 2020.

[208] S. ElSayed and M. Farouk, "Gender identification for Egyptian Arabic dialect in twitter using deep learning models," *Egypt. Informatics J.*, 2020.

[209] S. Shon, A. Ali, Y. Samih, H. Mubarak, and J. Glass, "ADI17: A Fine-Grained Arabic Dialect Identification Dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8244–8248.

[210] T. Tarmom, W. Teahan, E. Atwell, and M. A. Alsalka, "Compression versus traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study," *Nat. Lang. Eng.*, pp. 1–14.

[211] A. Abdelali, H. Mubarak, Y. Samih, S. Hassan, and K. Darwish, "Arabic Dialect Identification in the Wild," *arXiv Prepr. arXiv2005.06557*, 2020.

[212] F. Husain, "Arabic Offensive Language Detection Using Machine Learning and Ensemble Machine Learning Approaches," *arXiv Prepr. arXiv2005.08946*, 2020.

[213] R. AlYami and R. AlZaidy, "Arabic Dialect Identification in Social Media," in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, 2020, pp. 1–2.

[214] M. Alruily, "Issues of Dialectal Saudi Twitter Corpus."

[215] A. A. Al-Rawafi, T. Pujiati, and D. Sudana, "ON THE TYPOLOGY OF THE NEGATION MARKER MÂ IN MODERN ARABIC DIALECTS: KUWAITI, JORDANIAN, SUDANESE, AND YEMENI," *Arab. J. Pendidik. Bhs. Arab dan Kebahasaaraban*, vol. 7, no. 1, pp. 13–31, 2020.

[216] S. Al-Mulla and W. Zaghouani, "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, 2020, pp. 24–31.

[217] N. Ben Abdallah, S. Kchaou, and F. Bougares, "Text and Speech-based Tunisian Arabic Sub-Dialects Identification," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 6405–6411.

[218] O. Ibrahim, H. Asadi, E. Kassem, and V. Dellwo, "Arabic Speech Rhythm Corpus: Read and Spontaneous Speaking Styles," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 5337–5342.

[219] B. Talafha *et al.*, "Multi-Dialect Arabic BERT for Country-Level Dialect Identification," *arXiv Prepr. arXiv2007.05612*, 2020.

[220] K. Duh and K. Kirchhoff, "POS tagging of dialectal Arabic: a minimally supervised approach," in *Proceedings of the acl workshop on computational approaches to semitic languages*, 2005, pp. 55–62.

[221] N. Habash and O. Rambow, "Morphophonemic and orthographic rules in a multi-dialectal morphological analyzer and generator for arabic verbs," in *International symposium on computer and arabic language (iscal), riyadh, saudi arabia*, 2007.

APPENDIX A: Major Analysis Criteria of the Reviewed Dialect Studies

Article ID	Ref.	Year	Resources type	Research Area	Domain	ML Algorithms	Reported performance	Feature	Dialect Type	Document type	Source title
A1	[118]	2017	Speech Corpus	BR (Speech)	Travel and tourism	Gaussian densities	WERs (24.5, Egyptian dialect)	N/A	Egyptian, Gulf, Levantine + MSA	Article	*ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*
A2	[119]	2019	Text corpus	BR (Textual)	Social Media	Hybrid Approach + Rule-Based Approach + Statistical Approach +	WERs (09.80) + CER (10.47) on held-out test set.	n-gram	Tunisian	Article	*ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*
A3	[79]	2014	Text corpus	BR+DI (Textual)	Social Media	Naive Bayes	98.00% Overall Accuracy (NB) 73% Overall F-measure	*n*-gram + Markov Language Model	Multi +MSA	Poster	In *Proceedings of the first international workshop on Social media retrieval and analysis*
A4	[67]	2011	Arabic Commentary Dataset (AOC)	BR+DI (Textual)	News and Media, Website Comments	Language modeling approach	83.50% Accuracy (LEV vs. GLF vs. EGY) 77.70% Precision (Al-Youm Al-Sabe' MSA vs. EGY) 84.40% Recall (Al-Youm Al-Sabe' MSA vs. EGY)	*n*-gram	Levantine, Gulf, and Egyptian	Conference Review	In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*
A5	[120]	2016	Speech Corpus	BR+DI (Text and	News stories	Decision Tree	76.40.%	Frequency-	Palestinian	Conference	In *Proceedings of*

			with transcription	acoustic)	summing	(J48) + Naive Bayes + Random forest + Support Vector Machine	Overall Accuracy (Random forest) 98.70% Accuracy Precision Recall F-measure	inverse document frequency (TF-IDF)		Review	*the 2016 8th International Conference on Information Management and Engineering*
A6	[84]	2009	Speech Corpus	DI (Speech)	Corpora recorded under similar acoustic conditions	Parallel PRLM approach	81.60 % Overall accuracy (Parallel PRLM approach)	N/A	Gulf, Iraqi, Levantine, Egyptian + MSA	Conference Review	In *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*
A7	[121]	2018	Arabic Commentary Dataset (AOC)	DI (Textual)	News and Media, Website Comments	Bidirectional LSTM (BLSTM) + Convolutional neural networks (CNN) + Convolutional LSTM (CLSTM) + Long-short term memory (LSTM) +	97.10% Accuracy (LSTM, 3 pairs of dialects) 84.50% Accuracy (LSTM, 3 dialects together)	N/A	Egyptian Gulf, Iraqi, and Levantine	Conference Review	The 4th International Conference on Arabic Computational Linguistics (ACLing 2018)
A8	[122]	2015	Text corpus	DI (Textual)	News, Technology, Entertainment, Money	A hybrid of PSO and fuzzy logic	83.60% % F-measure (A hybrid of PSO and fuzzy logic)	*n*-gram	Egyptian	Article	*Egyptian Computer Science Journal*
A9	[123]	2015	Speech Corpus	BR (Speech)	News and Media	N/A	N/A	N/A	Egyptian, Levantine, Gulf, and North African.	Conference Review	In *Sixteenth Annual Conference of the International Speech Communication Association*

A10	[46]	2010	TuDiCoI (Tunisian Dialect Corpus Interlocutor)	BR (Speech)	Business	N/A	N/A	N/A	Tunisian	Conference Review	In *The international arab conference on information technology (acit), benghazi–libya*.
A11	[124]	2014	Multidialectal Arabic parallel corpus	BR (Textual)	Different topics	Cepstral GMM System	N/A	N/A	Egyptian, Tunisian, Jordanian, Palestinian, Syrian + MSA	Conference Review	In *LREC*
A12	[85]	2011	Speech Corpus	DI (Speech)	Business	N/A	N/A	N/A	Levantine, Iraqi, Gulf, and Egyptian	Conference Review	In *Twelfth Annual Conference of the International Speech Communication Association*.
A13	[72]	2015	Speech Corpus	DI (Speech)	News and Media	Support Vector Machine	60.20% Accuracy (system combination) 100% Accuracy (two binary classification tasks) Accuracy Precision Recall F-measure	Bottleneck Features (BN) & universal background model (UBM)	Multi +MSA	Article	*arXiv preprint arXiv*
A14	[125]	2017	Speech Corpus	DI (Text and acoustic)	News and Media	Gaussian Back-end + GANs Back-end + Support Vector Machine	76.94% Accuracy (ADI subtask) 79.76% Accuracy 80.27% Precision	Frequency-inverse document frequency (TF-IDF) + MFCC-based feature vector + n-	Multi +MSA	Conference Review	In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*

							79.87% Recall (MGB-3 test)	gram			
A15	[73]	2018	Arabic Commentary Dataset (AOC)	DI (Textual)	News and Media, Website Comments	Decision Tree (J48) + KNN + Naive Bayes + Support Vector Machine	76.29% Accuracy 79.00% Precision 76.00% Recall 78.00% F-measure (SVM)	n-gram	Multi +MSA	Conference Review	In *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.* European Language Resources Association.
A16	[126]	2017	Text corpus	DI (Textual)	Social Media	Bidirectional LSTM (BLSTM) + Conditional Random Fields (CRF) + Long-short term memory (LSTM)	92.65% Accuracy & F-measure (BiLSTM-CRF)	N/A	Egyptian + MSA	Conference Review	In *Proceedings of the Third Arabic Natural Language Processing Workshop*
A17	[127]	2010	Speech Corpus	DI (Speech)	Different topics	Gaussian Mixture Models (GMM)	The novel PMVDR-SDC based system combination (+26.4% relative improvement in average error rate)	MFCC	Emirati, Egypt and Iraqi	Conference Review	In *2010 18th European Signal Processing Conference*. IEEE.
A18	[128]	2017	Text corpus	DI (Textual)	Social Media	C5.0 + KNN + Naive Bayes + NNET + Support Vector Machine +	90.07% Accuracy 95.90% Precision 91.00% Recall 93.30% F-measure (SVM)	TF-IDF weighting	Multi +MSA	Conference Review	In *2017 International Conference On Social Media, Wearable And Web Analytics (Social Media). IEEE.*

A19	[129]	2014	Text corpus	BR+ DI (Textual)	Social Media	Latent Direchlet Allocation (LDA)	Recall 93.20%	n-gram	Moroccan (Darija)	Article	In *LREC*
A20	[130]	2009	Speech Corpus	DI (Speech)	Different topics	HMM	86.33% Accuracy (Combining phonotactic & prosodic classifiers) 94.90% F-measure (Egyptian)	N/A	Gulf, Iraqi, Levantine, Egyptian,	Conference Review	In *Tenth Annual Conference of the International Speech Communication Association*.
A21	[131]	2010	Speech Corpus	DI (Speech)	Different topics	PRLM and GMM-UBM Approaches + Support Vector Machine	The overall EER (6%)	Feature vector	Gulf, Iraqi, Levantine, Egyptian,	Article	*Odyssey 2010, The Speaker and Language Recognition Workshop*
A22	[132]	2014	Text corpus	BR (Textual)	Media, Politics, Sport.	N/A	N/A	N/A	Tunisian	Conference Review	In *5th International Conference on Arabic Language Processing*
A23	[133]	2013	Speech corpora + Qatari Corpus	BR (Speech)	News and Media	GMM-HMM	WER (28%)	Feature-space MLLR (fMLLR) + MFCC	Qatari (Gulf)	Conference Review	In *Workshop on Less Resourced Languages, new technologies, new challenges and opportunities (LTC 2013)*
A24	[134]	2014	Arabic Online Commentary (AOC)	DI (Textual)	News and Media, Website Comments	Latent Direchlet Allocation (LDA)	97.8% Accuracy 83.60% Accuracy (AOC dataset)	n-gram	Multi + MSA	Article	ARMY RESEARCH LAB ADELPHI MD
A25	[135]	2017	Speech Corpus	DI (Speech)	Different topics	DTW model + Latent	67.90 % Accuracy	MFCC	Egyptian, Iraq,	Article	*International Journal of*

						Direchlet Allocation (LDA) +	(DTW)		Levantine, Kuwait (Gulf)		*Computer Applications*
A26	[136]	2016	Arabic dataset SARA	DI (Speech)	Media shows, episodes and films	GMM-UBM + Identity Vector (Ivector) +	66.70 % Accuracy (GMM-UMB classifier)	i-vector framework + MFCC	Egyptian, Gulf, and Levantine	Conference Review	In *29th International Conference on Computers Applications in Industry and Engineering (CAINE 2016), Denver, USA*.
A27	[59]	2014	Text corpus	DI (Textual)	Different topics	Naive Bayes + Support Vector Machine	92.00% Accuracy (NB Uni, Mag., Extended AOC) 92.00% Accuracy (NB, Egy + Lev., Twitter)	n-gram	Egyptian, Gulf, Levantine, Maghrebi and Iraqi + MSA	Article	In *LREC*
A28	[137]	2018	Text corpus	BR (Textual)	Different topics	N/A	N/A	N/A	Multi + MSA	Article	In *OSACT 3 Proceedings*. LREC.
A29	[138]	2019	Text corpus	BR (Textual)	Social Media	N/A	N/A	N/A	Multi + MSA	Article	*Computer-Mediated Communication and Social Media Corpora. Clermont-Ferrand: Presses Universitaires Blaise Pascal*
A30	[139]	2018	Text corpus	BR (Textual)	Social Media	N/A	N/A	N/A	Multi + MSA	Conference Review	*In 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*. IEEE.
A31	[140]	2019	ArPod Speech	BR + DI	Different	Convolutional	98.00% F-	Chroma	Saudi,	Conference	*In Proceedings of*

			Corpus	(Speech)	topics	neural networks (CNN) + Extratrees + KNN + Multi Layer Perceptron (MLP) + Support Vector Machine +	measure (SVM)	Vector + MFCC + Mel spectrogram + Spectral contrast + Tonnetz	Egyptian, Lebanese and Syrian + MSA	Review	*the 3rd International Conference on Natural Language and Speech Processing,*
A32	[141]	2019	Arabic Dialect Corpus (SMADC)	DI (Textual)	Social Media	Dialectal Terms Method, Voting Methods, Frequent Terms Methods	90.00% Accuracy (the weighted voting method and SMADC)	N/A	Multi	Conference Review	In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*
A33	[142]	2019	MADAR-Corpus	DI (Textual)	Basic Traveling Expression	MNB classifier	85.96% Accuracy	Frequency-inverse document frequency (TF-IDF) + n-gram	Multi + MSA	Conference Review	In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*
A34	[143]	2019	MADAR-Corpus	DI (Textual)	Basic Traveling Expression	Long-short term memory (LSTM) + Naive Bayes + Word embedding (Word2vec)	67.73% Accuracy (Multinomial Naive Bayes) F1-score of 67.31%.	n-gram	Multi + MSA	Conference Review	*The Fourth Arabic Natural Language Processing Workshop co-located with ACL*
A35	[144]	2016	Speech corpus	DI (Speech)	Different topics	Long-short term memory (LSTM) + Support Vector Machine	42.79% Accuracy (SVM) F1-score of 42.64%.	n-gram	Multi + MSA	Conference Review	In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*
A36	[145]	2017	Speech corpus	BR (Speech)	Social Media	N/A	N/A	N/A	Algerian	Conference Review	*Proceedings of the Third Arabic*

											Natural Language Processing Workshop
A37	[146]	2017	Text corpus	BR+DI (Textual)	Politics, health, social issues, religious issues.	Ensemble method	79.00% Accuracy (ensemble method)	n-gram	Multi	Article	*International Journal of Computational Linguistics (IJCL)*
A38	[90]	2014	Speech corpus	BR (Speech)	News and Media	N/A	N/A	N/A	Egyptian	Conference Review	*In International Workshop on Spoken Language Translation (IWSLT 2014)*
A39	[147]	2015	Text corpus	DI (Textual)	Different topics	Decision Tree (J48)	90.80% Accuracy (SVM) F1-score of 90.60%	LM-features + MADAMIRA-features + Modality-features + Meta-features + NER-features	Egyptian + MSA	Conference Review	In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*
A40	[148]	2018	MADAR-Corpus	BR (Textual)	Basic Traveling Expression	N/A	N/A	N/A	Multi + MSA	Conference Review	In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*
A41	[149]	2018	Text corpus	BR (Textual)	Twitter's public message	N/A	N/A	N/A	Oman, Egypt, Iraq, Jordan, Kuwait, Palestine, Qatar, KSA, UAE, Yemen.	Conference Review	In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*

A42	[150]	2018	DART corpus	BR (Textual)	Twitter's public message	N/A	N/A	N/A	Multi	Conference Review	In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*
A43	[151]	2018	Text corpus	DI (Textual)	Twitter's public message	Support Vector Machine +	65.00% Accuracy (the n-gram based SVM trained)	n-gram	Levantine	Conference Review	In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*
A44	[152]	2016	Multi-Genre Broadcast (MGB-2)	DI (Text and acoustic)	News and Media	Bidirectional LSTM (BLSTM) + FDNN + Long-short term memory (LSTM) + TDNN +	21.5% WER (TDNN)	MFCC + *n*-gram	Egyptian, Gulf, Levantine, and North African + MSA	Conference Review	In *2016 IEEE Spoken Language Technology Workshop (SLT)*
A45	[153]	2019	MADAR-Corpus	DI (Textual)	Basic Traveling Expression	Multinomial Naive Bayes (MNB) classifier	67.90% Accuracy)	N/A	Multi + MSA	Conference Review	In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*
A46	[154]	2018	Speech corpus	DI (Speech)	Different topics	Feed-Forward Back Propagation Neural Networks (FFBPNN)	98.50% Accuracy (feature extraction phase the Perceptual Linear Prediction technique	MFCC + Perceptual Linear Prediction Coefficients + Vector Quantization (VQ)	Tunisian dialect (Darija)	Article	International Arab Journal of Information Technology

							(PLP))				
A47	[155]	2017	Pan-Arabic corpus + Tira	DI (Textual)	Twitter's public message	Support Vector Machine, SMO classifier	75.20% Accuracy (SMO classifier, n-gram, variety)	n-gram	Levantine, Gulf, Egypt and Maghrebi.	Article	In *CLEF (Working Notes)*
A48	[60]	2014	Text corpus	DI + BR (Textual)	Twitter's public message	N/A	Manual evaluation > 93.00% Accuracy	n-gram	Multi	Conference Review	In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*
A49	[156]	2014	Arabic Online Commentary (AOC)	DI (Textual)	News and Media, Website Comments	Support Vector Machine	89.10% Accuracy	n-gram	Egyptian + MSA	Conference Review	In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*
A50	[157]	2014	Text corpus	DI + BR (Textual)	Basic Traveling Expression	Naive Bayes classifier	98.00% Accuracy	n-gram	Multi + MSA	Conference Review	In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*
A51	[158]	2019	MADAR-Corpus	DI (Textual)	Basic Traveling Expression	Bidirectional LSTM (BLSTM) + Bidirectional Gated Recurrent Units (BiGRU) + Naive Bayes + Random forest +	F1 score of 63.02% (combination of Naive Bayes and Random Forest)	Frequency-inverse document frequency (TF-IDF)	Multi + MSA	Conference Review	In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*
A52	[159]	2019	MADAR-Corpus	DI (Textual)	Basic Traveling Expression	Ensemble Model + Logistic Regression	69:30 % Accuracy	n-gram	Multi + MSA	Conference Review	In *Proceedings of the Fourth Arabic Natural Language Processing*

											Workshop
A53	[160]	2019	MADAR-Corpus	DI (Textual)	Basic Traveling Expression	Logistic Regression + Support Vector Machine, mNB, logreg, MLP, Baseline	67.40% Accuracy (Baseline)	n-gram	Multi + MSA	Conference Review	In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*
A54	[161]	2019	MADAR-Corpus	DI (Textual)	Basic Traveling Expression	Naive Bayes	62.17% F1-score (MNB with character 4-grams)	N/A	Multi + MSA	Conference Review	In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*
A55	[162]	2019	MADAR-Corpus	DI (Textual)	Basic Traveling Expression	Baseline Ensemble + CharCNN + FastText embeddings + Long-short term memory (LSTM) +	66.60% F1-score (Char TFIDF + WordTFIDF + NN, Baseline)	n-gram	Multi + MSA	Conference Review	In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*
A56	[163]	2019	MADAR-Corpus	DI (Textual)	Basic Traveling Expression	RNN	65.80% F1-score (RNN)	Language Model features	Multi + MSA	Conference Review	In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*
A57	[164]	2018	Arabic Online Commentary (AOC)	DI (Textual)	News and Media, Website Comments	Attention-BiLSTM + Bidirectional LSTM (BLSTM) + Bidirectional Gated Recurrent Units (BiGRU) + Convolutional neural networks (CNN) + Convolutional LSTM	87.65 % Accuracy (BiGRU)	*n*-gram	Multi + MSA	Conference Review	In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*,

						(CLSTM) + Logistic Regression + Long-short term memory (LSTM) + Naive Bayes + Support Vector Machine					
A58	[165]	2018	The Multi-Genre Broadcast (MGB-3)	DI (Speech)	Different topics	Convolutional neural networks (CNN) + CNN-biLSTM + Support Vector Machine +	52.89% F1-score (Multi-Input CNN)	MFCC + Mel spectrogram	Egyptian, Gulf, Levantine, and North African + MSA	Conference Review	In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*
A59	[166]	2018	The ADI Shared Task data set	DI (Textual)	Different topics	Convolutional neural networks (CNN) + GRU recurrent layer	57.60% F1-score (CNN with a GRU recurrent layer)	Frequency-inverse document frequency (TF-IDF) + n-gram	Egyptian, Gulf, Levantine, and North-African +MSA	Conference Review	In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*
A60	[167]	2018	The ADI Shared Task data set	DI (Textual)	Different topics	Kernel Discriminant Analysis (KDA) + Kernel Ridge Regression (KRR)	62.28% F1-score (audio embeddings)	n-gram	Egyptian, Gulf, Levantine, and North-African +MSA	Article	*arXiv preprint arXiv*
A61	[168]	2017	KALAM'DZ	BR (Speech)	YouTube, other Social Media, Online Radio and TV.	N/A	N/A	N/A	Algerian	Conference Review	*Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP)*
A62	[169]	2017	The ADI Shared Task data set	DI (Textual)	Different topics	The meta-classifier	71.70% Accuracy	*n*-gram	Egyptian, Gulf, Levantine, and North-	Conference Review	*Proceedings of the Fourth Workshop on NLP for Similar Languages,*

									African +MSA		*Varieties and Dialects*
A63	[170]	2016	The ADI Shared Task data set	DI (Textual)	Different topics	Logistic Regression + Naive Bayes + Neural Networks + Support Vector Machine	51.36% Accuracy (SVM with a linear kernel)	*n*-gram	Egyptian, Gulf, Levantine, and North-African +MSA	Conference Review	*Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*
A64	[171]	2016	The ADI Shared Task data set	DI (Textual)	Different topics	Kernel Discriminant Analysis (KDA) + Kernel Ridge Regression (KRR) +	Accuracy of 50.91% and a weighted F1 score of 51.31%.	n-gram	Egyptian, Gulf, Levantine, and North-African +MSA	Conference Review	*Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*
A65	[172]	2016	Transcribed speech corpus	DI + BR (Text and acoustic)	Different topics	Median Ensemble + Mean Probability Ensemble system + Voting Ensemble	51.00% F1-score	N/A	Egyptian, Gulf, Levantine, and North-African +MSA	Conference Review	*Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*
A66	[173]	2016	Linguistic resources (dataset and lexicons)	DI + BR (Textual)	Different topics	Support Vector Machine	92.94% F1-score (SVM, Character 5-6-grams + DV)	n-gram	Algerian, Egyptian, Gulf, Levantine ,Mesopotamian, Moroccan, Tunisian + MSA	Conference Review	*Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*
A67	[83]	2009	Transcribed speech corpus	DI (Text and acoustic)	Different topics	Gaussian Mixture Models (GMM)	ERR (9.37%)	MFCC + Shifted-delta-cepstra (SDC)	Emirati (Gulf), Egypt, Iraq, Palestine, Syria.	Conference Review	In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*
A68	[174]	2019	Speech Corpus	DI (Speech)	Different topics	Long-short term memory	Accuracy of 94.00%	MFCC	Yemen, Saudi, Iraq,	Conference Review	*2019 IEEE International*

						(LSTM) + RNN			Egypt, and Sudan		*Conference on Automatic Control and Intelligent Systems (I2CACIS 2019)*
A69	[175]	2018	Speech Corpus	DI (Speech)	Different topics	Gaussian Mixture Model-Universal Background Model (GMMUBM)	Accuracy of 63.50%	MFCC	Egyptian and Levantine	Conference Review	In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*
A70	[176]	2017	The Multi-Genre Broadcast (MGB-3)	DI (Textual)	Different topics	Support Vector Machine	Accuracy of 75.00%	Lexical features + n-gram	Egyptian, Gulf, Levantine, and North-African +MSA	Conference Review	In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*
A71	[177]	2018	KALAM'DZ	DI (Speech)	YouTube, other Social Media, Online Radio and TV.	DNN Modeling	47% Precision 47:8 % Recall	MFCC + Shifted Delta Cepstral (SDC) coefficients	Algerian	Conference Review	In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*
A72	[178]	2019	The Multi-Genre Broadcast (MGB-3)	DI (Speech)	Different topics	DNN Modeling + Gaussian Mixture Models (GMM) + Gaussian Mixture Model-Universal Background Model (GMMUBM)	Accuracy of 73.80%	MFCC + UBNF/i-vector	Egyptian, North African, Gulf, Levantine + MSA	Conference Review	In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*
A73	[116]	2009	Speech Corpus	DI + BR (Speech)	Different topics	Gaussian densities	Accuracy of 99.34%	MFCC	Egyptian	Conference Review	*2009 Eighth International Symposium on Natural Language*

											Processing
A74	[179]	2018	Speech Corpus	DI + BR (Speech)	News and Media	Multi-lingual phone recognizers, Convolutional neural networks (CNN) + Support Vector Machine	Accuracy of 73.27%	n-gram	Egyptian, Gulf, Levantine, North African + MSA	Conference Review	In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*
A75	[180]	2014	Speech Corpus	DI (Speech)	Different topics	AdaBoostM1 + BayesNet + Bagging + Decision Tree (J48) + Decision Table (DT) + IBk + JRip + Multi Layer Perceptron (MLP) + Naive Bayes + OneR + Random forest + Sequential Minimal Optimization (SMO) +	Accuracy of 80.00% (AdaBoostM)	MFCC + Short-time Fourier transform (STFT) + Wavelet transform	Egyptian, Jordanian	Conference Review	2014 5th International Conference on Information and Communication Systems (ICICS)
A76	[48]	2013	Speech Corpus	BR (Speech)	Different topics	N/A	N/A	N/A	Gulf, Egypt, Levantine + MSA	Conference Review	In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*
A77	[47]	2013	Speech Corpus	BR (Speech)	Different topics	N/A	N/A	*n*-gram	Gulf, Egyptian, North Africa, Levantine	Conference Review	In *2013 1st International Conference on Communications, Signal Processing,*

											and their Applications (ICCSPA)
A78	[181]	2014	The BBN/AUB corpus	DI (Speech)	Different topics	Artificial neural networks (ANNs) + Rhythm metrics	Accuracy of 70.79%	N/A	Levantine	Conference Review	*2014 9th International Symposium on Communication Systems, Networks & Digital Sign (CSNDSP)*
A79	[182]	2017	Text corpus	DI + BR (Textual)	Twitter's public message	Decision Tree (J48) + Naive Bayes + Rule-Based Approach	Accuracy of 71.18% (Naive Bayes)	Feature vector	Gulf, Iraqi, Levantine, Moroccan, Sudanese, Egyptian	Conference Review	*2017 International Conference on Computer and Applications (ICCA)*
A80	[183]	2011	Speech Corpus	DI+BR (Speech)	Different topics	HMM	Accuracy of 100% (MSA) Accuracy of 95.00% (Egyptian)	Delta-Delta + MFCC	Egyptian + MSA	Conference Review	In *International Conference on Informatics Engineering and Information Science*, Springer, Berlin, Heidelberg
A81	[75]	2017	MCA corpus	DI + BR (Textual)	Social Media	Logistic Regression + Naive Bayes + Rule-Based Approach + Support Vector Machine	Accuracy of 83.84% (Rule-based classifier)	N/A	Moroccan	Conference Review	In *International Conference on Arabic Language Processing*. Springer, Cham.
A82	[184]	2015	TAD corpus and TLD corpus	BR (Textual)	Social Media	N/A	73% Accuracy 94% Precision 60% Recall 73% F-measure	N/A	Tunisian	Conference Review	In *International Conference on Web Engineering*, Springer, Cham.
A83	[185]	2016	Speech Corpus	DI (Speech)	Different topics	Feed forward back-propagation neural network (FFBPNN) + Principal	Accuracy of 98.30 % (FFBPNN)	MFCC	Tunisian, Moroccan.	Article	*International Journal of Speech Technology*

						component analysis (PCA) + Support vector machine (SVM)					
A84	[61]	2017	Curras PAL corpus morphologically annotated corpus of the Palestinian Arabic dialect	BR (Textual)	Raw text, annotations, experiment data	N/A	N/A	N/A	Palestinian	Article	*Language Resources and Evaluation*
A85	[186]	2018	SUAR (SaUdi corpus for NLP Applications and Resources)	BR (Textual)	Social media	N/A	N/A	N/A	Saudi	Conference Review	*The 4th International Conference on Arabic Computational Linguistics (ACLing 2018)*
A86	[187]	2019	Egyptian Dialect Gender Annotated Dataset (EDGAD)	DI (Textual)	Twitter's public message	Logistic Regression + Random forest	Accuracy of 87.60% (weighted ensemble for the RF classifier)	Feature vector + n-gram	Egyptian	Article	*Egyptian Informatics Journal*
A87	[89]	2008	The Saudi Accented Arabic Voice Bank (SAAVB)	BR (Text and acoustic)	Different topics	N/A	N/A	N/A	Saudi	Article	*Journal of King Saud University-Computer and Information Sciences*
A88	[188]	2015	Speech Corpus	DI (Speech)	Different topics	DBN + HMM + MLP/HMM + SVM/HMM + SVM/DBN	Accuracy of 87.67% (SVM/DBN)	N/A	Gulf, Egyptian, Levantine + MSA	Conference Review	*19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems*
A89	[189]	2017	Nemlar10 +NetDC11 corpora	DI (Textual)	Different topics	DNN Modeling	WER (89%)	n-gram	Algerian	Conference Review	*The 3rd International Conference on Arabic Computational*

											Linguistics, ACLing 2017
A90	[190]	2018	Speech Corpus	BR (Speech)	Different topics	N/A	N/A	N/A	Egyptian, Gulf, Levantine + MSA	Conference Review	*The 4th International Conference on Arabic Computational Linguistics (ACLing 2018)*
A91	[191]	2018	PADIC corpus	DI (Textual)	Different topics	Naive Bayes + Support Vector Machine	Accuracy of 92.00% (Combined)	Frequency-inverse document frequency (TF-IDF)	Algerian, Tunisian, Moroccan, Syrian and Palestinian	Conference Review	*The 4th International Conference on Arabic Computational Linguistics (ACLing 2018)*
A92	[192]	2019	Speech Corpus	DI (Speech)	Different topics	HMM	Accuracy of 90.00% (HMMSRS)	Delta-Delta + MFCC	Moroccan	Conference Review	*International Symposium on Machine Learning and Big Data Analytics for Cybersecurity and Privacy (MLBDACP)*
A93	[193]	2018	Speech Corpus	DI (Speech)	Emirati sentences	CSPHMM	Accuracy of 95.90% (CSPHMM3s, neutral environment) Accuracy of 59.30% (CSPHMM3s, shouted environment)	MFCC	Emirati (Gulf)	Article	*International Journal of Speech Technology*
A94	[194]	2019	Speech Corpus	DI (Speech)	Emirati sentences	Gaussian Mixture Models (GMM) + Neural	Accuracy of 83.97% (Novel sequential GMM-DNN)	N/A	Emirati (Gulf)	Article	*IEEE Access*

						Networks + SVMs and MLP classifiers					
A95	[195]	2019	Speech Corpus	DI (Speech)	Emirati sentences	HMM	Accuracy of 65.00% (HMM3s)	N/A	Emirati (Gulf)	Conference Review	In *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)* IEEE.
A96	[196]	2014	Speech Corpus	DI (Speech)	Emirati sentences	HMM + Gaussian Mixture Models (GMM) + Vector Quantization (VQ)	Accuracy of 100% (VQ, text-dependent systems) Accuracy of 94.48% (VQ, text-independent systems)	N/A	Emirati (Gulf)	Conference Review	ICSP2014 Proceedings
A97	[197]	2018	Speech Corpus	DI (Speech)	Emirati sentences	HMM	Accuracy of 65.90% (HMM3s)	MFCC	Emirati (Gulf)	Conference Review	In *2018 Fifth HCT Information Technology Trends (ITT)*. IEEE.
A98	[198]	2020	PADIC (Parallel Arabic DIalectal Corpus) + Oranee textual corpus [198]	BR (Textual)	Different topics	N/A	N/A	n-gram	Algiers, Annaba cities, Palestinian, Syrian, Tunisian, Moroccan + MSA	Article	*International Journal of Informatics and Applied Mathematics*
A99	[106]	2020	Habibi (a multi Dialect multi National Arabic Song Lyrics Corpus)	BR (Speech)	Arabic songs	Bidirectional LSTM (BLSTM) + Bidirectional Gated Recurrent Units (BiGRU) +	Accuracy of 93.00% (CNN model)	N/A	Multi	Conference Review	*Twelfth International Conference on Language Resources and Evaluation. European Language*

						Convolutional neural networks (CNN) + Convolutional LSTM (CLSTM) + Logistic Regression + Long-short term memory (LSTM) + Naive Bayes + Support Vector Machine					*Resources Association (ELRA), FRA.*
A100	[74]	2018	Multi-Genre Broadcast 3 (MGB-3)	DI (Speech)	Different topics	Baseline Embedding	Accuracy of 73.00% (a single feature set) Accuracy of 78.00% (multiple featurest)	MFCC	Multi + MSA	Article	*arXiv preprint arXiv*
A101	[63]	2014	Egyptian Treebank	BR (Textual)	Different topics	N/A	N/A	N/A	Egyptian	Article	In *LREC*
A102	[199]	2018	Text corpus	BR (Textual)	Twitter's public message	N/A	N/A	*n*-gram	Levantine	Conference Review	In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
A103	[200]	2018	PADIC	BR (Text and acoustic)	Different topics	N/A	N/A	N/A	Multi	Conference Review	*7th International Conference on Advanced Technologies ICAT*, Apr 2018, Antalya, Turkey.
A104	[201]	2017	Comparable	BR	Different	N/A	N/A	N/A	Egyptian +	Conference	2017 Palestinian

			corpus19 + WikiDocsAligner	(Textual)	topics				MSA	Review	International Conference on Information and Communication Technology
A105	[202]	2018	Text corpus	BR (Textual)	Forums, comments & blogs.	N/A	N/A	N/A	Multi	Conference Review	In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*
A106	[10]	2014	Lexicon	BR (Textual)	Bilingual dictionary	N/A	N/A	N/A	Moroccan + MSA	Conference Review	5th International Conference on Arabic Language Processing CITALA, Oujda, Morocco 11/ 2014
A107	[203]	2017	CALYOU	BR (Text and acoustic)	Different topics	N/A	N/A	N/A	Algerian	Conference Review	18th Annual Conference of the International Communication Association (Interspeech), Aug 2017, Stockholm, Sweden
A108	[204]	2016	ArabicWeb16	BR (Textual)	Different topics	N/A	N/A	N/A	Multi + MSA	Conference Review	In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*
A109	[205]	2016	Gumar Corpus	BR (Textual)	Novels	N/A	N/A	N/A	Gulf + MSA	Article	*arXiv preprint arXiv*
A110	[206]	2014	Transcribed speech corpus	BR (Text and acoustic)	Different topics	N/A	N/A	N/A	Egyptian + English	Conference Review	In *Proceedings of International Workshop on Spoken Language Translation*

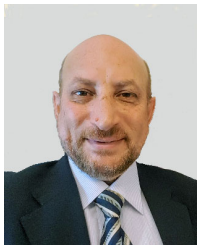
											(IWSLT).
A111	[207]	2020	A semi-natural Egyptian Arabic speech emotion (EYASE) database	BR + DI (Speech)	Egyptian sentences	SVM + kNN	95.00% Accuracy (SVM, Males, Angry) 92.20% Accuracy (Knn, Males, Angry)	Prosodic features + Spectral features (MFCC and LTAS) + Wavelet features	Egyptian	Article	*Speech Communication*
A112	[208]	2020	Dialect Gender Annotated Dataset (EDGAD) + PAN AP'17	DI (Textual)	Twitter's public message	Neural Network (ANN), Convolutional Neural Network (CNN), Long–Short Term Memory (LSTM), Convolutional Bidirectional Long-Short Term Memory (C-Bi-LSTM) and Convolutional Bidirectional Gated Recurrent Units (C-Bi-GRU)	91.37% Accuracy (C-Bi-GRU multichannel model)	N/A	Egyptian	Article	Egyptian Informatics Journal
A113	[209]	2020	Speech Corpus + Multi-Genre Broadcast 3 (MGB-3)	BR + DI (Speech)	YouTube videos	The Gaussian Mixture Model-Universal Background Model (GMM-UBM)	82.00% Accuracy 82.10% Precision 83.30% Recall (Supervised task)	MFCCs and delta and delta-delta	Middle East and Northern Africa.	Conference Review	In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*
A114	[210]	2020	Saudi dialect corpus (SDC);	DI (Textual)	Social Media	Prediction by partial	99.80% Accuracy	n-grams	Egyptian, Saudi,	Article	*Natural Language Engineering*

			Egyptian dialect corpus (EDC).			matching (PPM) compression-based classifier, sequential minimal optimization (SMO)	(PPM)		English + MSA		
A115	[211]	2020	Text corpus + QADI Dataset + MADAR Dataset	BR + DI (Textual)	Twitter's public message	SVM classifier and fine-tuned BERT and AraBERT	91.50% Accuracy	n-grams	Multi + MSA	Article	*arXiv preprint arXiv*
A116	[212]	2020	Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) in Language Resources and Evaluation Conference (LREC) 2020.	DI (Textual)	Twitter's public message	Support Vector Machine (SVM), logistic regression and decision tree, ensemble machine learning model (bagging, random forest, and AdaBoost).	88% using the best ensemble machine learning model.	count and TF-IDF features	Multi + MSA	Article	*arXiv preprint arXiv*
A117	[213]	2020	Text corpus	BR+ DI (Textual)	Twitter's public message	Margin classifiers, decision trees, generative and deterministic probabilistic models	87.00% Accuracy (Egyptian Dialect Binary Classification, SVM)	Inverse document frequency (TF-IDF).	Multi	Conference Review	In *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*
A118	[214]	2020	Text corpus	BR (Textual)	Twitter's public message	N/A	N/A	Shortening, compounded words, misspelled words, abbreviatio	Saudi + MSA	Conference Review	*The International Arab Journal of Information Technology*

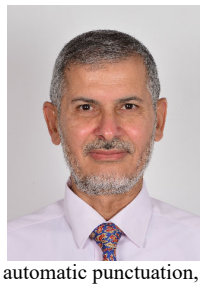
								ns, dialectal words (slang), neologisms, concatenation, word elongation and idiomatic expressions.			
A119	[215]	2020	Text corpus	BR+ DI (Textual)	Published articles and YouTube	N/A	N/A	N/A	Kuwaiti, Jordanian, Sudanese, and Yemeni.	Article	*Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban*
A120	[216]	2020	Text corpus	BR (Textual)	Qatari traditional expressions and idioms	N/A	N/A	N/A	Qatari (Gulf)	Conference Review	In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*
A121	[217]	2020	Speech Corpus	BR+ DI (Speech)	Different topics	Support Vector Machines (SVM), Naive Bayes (NB), and a Multilayer Perceptron (MLP)	93.75% F-1 score (best speech-based identification system),	n-grams + Spectral features Extraction + MFCC	Tunisia	Conference Review	In *Proceedings of The 12th Language Resources and Evaluation Conference*
A122	[218]	2020	Speech Corpus	BR (Speech)	Different topics	N/A	N/A	N/A	Egyptian + MSA	Conference Review	*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*
A123	[219]	2020	The Nuanced	BR+ DI	Twitter's	SVM, Logistic	26.78% F-1	TF–IDF	Multi	Conference	*arXiv preprint*

			Arabic Dialect Identification (NADI)	(Textual)	public message	Regression (LR) and Naive Bayes (NB)	score			Review	*arXiv*
A124	[62]	2006	Arabic Treebank	BR (Textual)	Different topics	N/A	N/A	N/A	Levantine + MSA	Conference Review	In *LREC*
A125	[88]	2005	The LDC CallHome corpus of Egyptian Colloquial Arabic, and the FBIS Modern Standard Arabic corpus.	BR +DI (Text and acoustic)	Different topics	HMMs	N/A	MFCC	Egyptian + MSA	Article	*Speech Communication*
A126	[220]	2005	The ECA corpus	BR +DI (Textual)	Different topics	HMM	68.48 % Accuracy	N/A	Egyptian + MSA	Conference Review	*Proceedings of the acl workshop on computational approaches to semitic languages*
A127	[54]	2006	CTS Treebank	BR +DI (Textual)	Different topics	Transductive SVMs, Spectral Graph Transducers, and a novel Transductive Clustering method.	TSVM (63.54 % Accuracy	N/A	Levantine	Conference Review	*Proceedings of the 2006 conference on empirical methods in natural language processing*
A128	[112]	2006	Text corpus	BR (Textual)	Different topics	N/A	N/A	N/A	Iraqi	Conference Review	In *LREC*
A129	[64]	2006	Penn Arabic Treebank (ATB) + the Levantine Arabic Treebank (LATB)	BR (Textual	Different topics	N/A	N/A	N/A	Levantine + MSA	Conference Review	*Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*
A130	[221]	2007	Text corpus	BR	Different	N/A	N/A	N/A	Levantine	Conference	*International*

				(Textual)	topics				+ MSA	Review	*symposium on computer and arabic language*



Ashraf Elnagar is a Professor of Artificial Intelligence at the Department of Computer Science, University of Sharjah, UAE. He received his BSc (1986) and MSc (1988) in Computer Science from the University of Kuwait, Kuwait, and his and PhD (1993) from the University of Alberta, Canada. During his service at the University of Sharjah, he served as the founding chair of the Dept. of Computer Science, Chair of the MIS Department, and Dean of the Community College. He won a number of teaching, research and community and professional service awards. He is the recipient of the 1999 Shoman's Best Young Researcher Award in the Arab World in the fields of Mathematics, Computer Science and Statistics. His research interests include intelligent systems (robotics), machine learning, natural language processing, pattern analysis and recognition, and IT education.



Sane M Yagi received his education in Jordan, U.S.A., and New Zealand. He is a professor of Linguistics at the Department of Foreign Languages, University of Sharjah. His research is in computational linguistics, CMC, CALL, and TEFL. His research is currently in the broad field of Arabic Computational Linguistics. The primary themes are corpus development, computational lexicography & lexicology, computational morphology, syntactic parsing, automatic punctuation, and machine learning.

Ali Bou Nassif is currently an Assistant Professor at University of Sharjah, UAE, as well as an Adjunct Research Professor at Western University, Canada. He obtained a Master's degree in Computer Science and a Ph.D. in Electrical and Computer Engineering from Western University, Canada in 2009 and 2012, respectively. Ali's research interests include the application of statistical and artificial intelligence models in different areas such as software engineering, electrical engineering, e-learning, security and social media. Ali is a registered professional engineer (P.Eng) in Ontario, as well as a member of IEEE Computer Society.

Ismail Shahin is an Associate Professor of Speech and Speaker Recognition at the Department of Electrical and Computer Engineering, University of Sharjah, United Arab Emirates. He received his B.Sc., M.Sc., and Ph.D. degrees in electrical engineering in 1992, 1994, and 1998, respectively, from Southern Illinois University at Carbondale, U.S.A. He has 70 journal and conference publications. He won a number of teaching, research and community service awards. He has remarkable contribution in organizing conferences, symposiums and workshops. His research interests include speech recognition, speaker recognition under neutral, stressful, and emotional talking conditions, emotion and talking condition recognition, gender recognition using voice, and accent recognition in Arabic and English.

Said A. Salloum graduated with distinction from The British University in Dubai with an MSc in Informatics (Knowledge and Data Management). He got his bachelor's degree in computer science from Yarmouk University. He is currently working at the University of Sharjah "Research Institute of Sciences and Engineering (RISE)" as a researcher. He works in different research areas in Computer Science, including data analysis, machine learning, knowledge management, and Arabic Language Processing. Salloum is an Oracle expert since 2013 with several internationally recognized certificates.