International Journal of Information Technology and Language Studies (IJITLS)

Research Trends in the Fields of Arabic Natural Language Processing Tasks and Arabic Information Extraction Applications: A Survey Study

Abduladem Aljamel, Hussein Khalil, and Yousef Aburawi
Faculty of Information Technology, Misurata University, Libya
a.aljamel@it.misuratau.edu.ly, hussein.khalil@misuratau.edu.ly, yaburawi@it.misuratau.edu.ly

Article Information

Article type: Review

Article history:

Received: August 30, 2021
Revised: November 17, 2021
Accepted: November 17, 2021

Keywords:

Natural Language Processing Tasks,
Information Extraction Applications,
Arabic Language,
Linguistic Levels,
Arabic Linguistic Resources,
Arabic Technical Resources.

Abstract

This survey has explored the literature on the fields of Arabic NLP tasks and Arabic IE applications to analyze the state-of-the-art trends, identify the research gaps in these research fields, and recommend solutions to fulfill these gaps. This study is set out to gather appropriate research articles in the targeted fields from Academic Search Engines and Academic Databases. Subsequently, these articles were surveyed to obtain information about research trends aspects. That is, the contributions achieved, the methodologies applied, and the technical and linguistic resources utilized. This review study has followed systematic review procedure steps to meet the requirements of high-quality survey studies. The collected and reviewed articles cover different research contributions. For instance, the Morphological resolution in the field of Arabic NLP tasks and the Sentiment Analysis (SA) applications in the field of Arabic IE applications. The findings of this study can be summarized into that most of the researchers in the field of Arabic NLP tasks prefer to contribute to NER and then to the Morphological resolution tasks; however, in the field of Arabic IE they prefer to contribute to SA applications and then to the Question and Answering applications. Secondly, most of the reviewed articles applied methodologies, tools, techniques, and algorithms, not for specific languages such as Machine Learning, Artificial Neural Networks, and Deep Learning Algorithms. Lastly, this study provides the first comprehensive assessment which examines associations between the dataset sources domain types and dataset sources ownership types in addition to the relation between articles' contribution fields and the datasets ownership types. It confirms that the highest-reviewed articles numbers in the field of Arabic NLP tasks are for those that utilize existing and available dataset sources; specifically, in Linguistic domain dataset sources. Nonetheless, the highest reviewed articles numbers in the field of Arabic IE applications are for those whose authors are collecting and creating the dataset sources by themselves; also, in Linguistic domain dataset sources.

1. INTRODUCTION

An increasing amount of data is being made available online. It covers a diversity of domains such as entertainment, financial and economy, education, politics, sports, and others. There is an opportunity in extracting beneficial information from this data to be exploited to inform a variety of applications and services, such as recommender systems to advise financial investors about potential business risk or sentiment analysis to inform the services providers companies about an

emerging consumer trend. However, this online data is diverse in terms of volume and complexity, largely unstructured and constructed in natural human languages. This makes the manual exploitation of this data by end-users very difficult. Therefore, linguistically pre-processing that unstructured data by machines are needed in order to understand and extract useful information for the end-users (Aljamel et al., 2019).

Natural Language Processing (NLP) is an area of research that concerns how machines can computationally analyze and process a natural language text to extract useful information. Research in this area aims to understand the mechanism of processing natural languages by humans to assist in developing appropriate techniques for machines to understand and manipulate Natural Languages. These techniques will perform the desired NLP tasks by using Linguistics and computational disciplines and principles (Chowdhury, 2003; Collobert et al., 2011; Schubert, 2019; Wang et al., 2018). On the other hand, IE is the application of NLP techniques to automatically extract beneficial information from unstructured data. The unstructured data is constructed in natural human languages in different domain knowledge, and the extracted information is for a variety of applications and services. IE applications can be implemented to extract information from a specific domain knowledge or generic domain knowledge. In fact, most IE applications focus on the selected features of the targeted domain-specific knowledge where text variations are tightly constrained, whereas the approaches of extracting information from generic-domain knowledge are considered as a complicated task even for most specialist researchers. Usually, IE approaches are based on seeking the language patterns by applying NLP tasks such as linguistic lexical, syntactic, and semantic constraints (Alam & Awan, 2018; Fasha et al., 2017; Mannai et al., 2018; Wang et al., 2018).

Compared to other languages such as the European languages, NLP tasks and IE applications, which are aiming for Arabic text, have additional challenges. For instance, the Arabic language has rich morphology, has short unwritten vowels, is highly inflectional and derivational and lacks capitalization convention. These characteristics of the Arabic language generate various ambiguities when computational models are applied to the Arabic text, such as lexical, structural, semantic, and anaphoric ambiguities. In automating the process of extracting useful information from unstructured Arabic data, there are computational problems related to the overlapping of the Arabic language characteristics levels. These levels are morphological, syntactic, and semantic levels. The success of accurately analyzing these levels will assist in making sense and meaning of words and disambiguating the extracted information (Fasha et al., 2017; Khalil et al., 2020; Shaalan et al., 2018; Zakria et al., 2019).

Due to the fact that the Arabic language has a complex and ambiguous structure, the computational models have to deal with all aforementioned linguistic levels. Accordingly, Arabic NLP tasks and Arabic IE Applications' techniques, algorithms, and tools require more effort to overcome the complexity and ambiguity of the Arabic linguistic structure. There are many NLP and IE approaches and methods that have not been investigated in extracting meaningful information from unstructured Arabic data. Instead, these approaches and methods are extensively investigated and applied to extract useful information from non-Arabic unstructured data. It could be because those approaches and methods cannot be applied to the linguistic structures of Arabic texts, which require more investigation in areas and issues that the researchers of other natural human languages have not covered (Shaalan et al., 2018). There is a continued research effort in the fields of NLP and IE; nonetheless, the authors of this research believe that there is still room for improvement in the Arabic NLP tasks and Arabic IE methodologies and techniques.

Survey studies or literature review studies represent powerful information sources for practitioners who are looking for the state-of-the-art evidence to guide their investigations and work practices. The results of these studies would assist the researchers in providing a general overview of a specific field of study. Furthermore, they have an important role in assisting the researchers in finding a research gap in a specific area. In the survey studies, the relevant studies in specific field literature are collected then reviewed. The main objective of the literature survey studies is critically analysis the approaches in the reviewed studies to identify the research trends and also to find the research gaps which need to be fulfilled in specific fields (Miswar et al., 2018). In this survey study, the research aspects of contributions achieved, methodologies applied, and the technical and linguistics resources utilized in the fields of the Arabic NLP tasks and Arabic IE applications will be surveyed to obtain information about the current research trends in the targeted fields. Moreover, this study will provide a detailed analysis of the acquired results to identify the research gap which needed to be fulfilled in the research fields of Arabic NLP and Arabic IE.

The remaining parts of this survey study are organized as follows. It begins by reviewing the related works in the literature in section two. Section three introduces the challenges of the linguistic structure of the Arabic language. Section four presents the concepts and the importance of Arabic NLP tasks and Arabic IE applications. The fifth section is concerned with the methodology used for this survey study besides the research questions and research objectives. The stages of this survey study are presented in section six, seven, and eight, including the details of the results' analysis and discussion. Section nine concludes this research and summarizes the main outcomes of this work and outlines suggested further work.

2. LITERATURE REVIEW

Several survey studies have been crafted in the literature that were related to specific tasks in the Arabic NLP field or specific applications in the Arabic IE application for different domains and objectives. For example, Sarhan, et al. (2016) have

produced a survey study to discuss the major techniques used in Arabic Relation Extraction and investigated their strengths and weaknesses to guide future research towards creating an enhanced convenient Relation Extraction algorithm. They concluded that there is still room for improvement in the Arabic Relation Extraction task.

Another survey study was conducted by Salloum, et al. (2018) to provide a broad review of various studies related to the Arabic text mining with more focus on the Holy Quran, Sentiment Analysis (SA), and Web Documents. According to them, the synthesis of the research problems and methodologies of the surveyed studies will help the text mining scholars in pursuing their future studies; for example, researchers could develop clustering techniques or novel classification methods that will be helpful for the analysis of text in large-scale systems in the Arabic context.

In a different survey study, Alian, et al. (2017) presented a literature review survey to examine what researches have been done to solve the problem of Arabic Word Sense disambiguation. The Word Sense Disambiguation is the capability to identify what a word means concerning a context in which it comes into view. The authors of this survey study argued that the researchers in the Arabic language use different datasets with limited sizes. These datasets are not available, which makes considering their results problematic. Also, they observed that the use of Arabic WordNet has several problems such as noise, precision, and limited coverage compared to English WordNet.

To better understand the research directions in the field of Arabic SA, Abo et al. (2019) performed a literature survey study to analyze the current Arabic SA techniques to find the research trends in terms of the techniques applied in addition to comprehensively investigate the articles demographics, productivity, and directions of the Arabic SA research field. Their results showed that not only there is an increase in the research efforts in the field of Arabic SA, but also there is an increase in the number of publications per year since 2015 in the same field.

In the same Arabic SA field, Al-Ayyoub, et al., (2019) present a survey study to review the works done so far on Arabic SA to address and identify the gaps in the current literature to be a foundation for future studies in this field. They surveyed many articles that target Arabic SA that cover the methods, tools, and other resources which may help the Arabic SA researchers in their works. This survey covers both corpus-based and lexicon-based SA approaches. Furthermore, it covered most Arabic dialects and Modern Standard Arabic languages and covered many knowledge domains such as news, tweets, posts, reviews, and others. The reviewers in this survey study have grouped the reviewed articles based on the SA-related problems and showed most of the articles that presented solutions for these problems that face Arabic SA researchers. Finally, they introduce a full summarization of the articles mentioned in this study according to problems related to the Arabic SA research field.

In different domain knowledge, which is the hadiths or Prophetic Mohammed traditions, Azmi, Al-Qabbany and Hussain (2019) have presented literature survey study. This study surveys all major works that have addressed the subject of hadith through various computational and NLP approaches and techniques, grouping them under three categories: hadith content-based articles, narration-based articles, and overall articles. Besides, the authors deeply reviewed the pioneering works with many details appearing for the first time. Lastly, they suggested the application of emerging natural language concept-based sentiment and emotion mining techniques as a future research direction in the domain of Arabic hadith literature.

A comprehensive survey study was carried out by Marie-Sainte, et al. (2019) to survey the Arabic NLP systems based on Machine Learning (ML) Algorithms. The authors presented several systems that utilize Arabic NLP and ML-based systems techniques, their methodologies, challenges, and solutions. According to them, this survey can serve as an important theoretical foundation for researchers who are interested in this field.

Two important themes emerge from the related works discussed so far. Overall, these survey studies highlight the need for the literature survey studies; particularly, in assisting the researchers to summarize the current research trends to identify the gap which needs to be fulfilled in the target field. Secondly, it should be noted from the aforementioned related works, and to the best of the authors' knowledge, that the survey studies to date have not considered all research aspects which are related to Arabic NLP tasks or Arabic IE applications, which are the contributions achieved, methodologies applied, and the technical and dataset resources used. Overall, much of survey studies of the current literature indicate that they pay particular attention to a specific IE application such as Arabic SA. This was the motivation behind presenting this literature survey study in the fields of Arabic NLP tasks and Arabic IE applications instead of Arabic IE application-specific survey study. Specifically, research aspects are the contributions achieved, methodologies applied, and the technical and dataset resources used in the fields of Arabic NLP tasks or Arabic IE applications.

3. THE CHALLENGES OF ARABIC LANGUAGE

Arabic is the mother tongue of millions of speakers in about 20 countries and is used by more than a billion Muslims around the world who perform their daily Islamic activities using the Arabic language. The Arabic language appears in three categories, Standard Arabic, Arabic Dialect, and Modern Standard Arabic. It is essential for Arabic computational linguistics tasks to be able to distinguish between these three types (Abo et al., 2019).

Arabic language is written from right to left, its letters have no capitalization, and its letters change shape according to their position in the word. Furthermore, Arabic language has a set of orthographic symbols, called diacritics or inner diacritics

or short vowels, that carry the intended pronunciation of words. These diacritics support clarifying the sense and meaning of the word. The absence of short vowels leads to several types of ambiguity in Arabic text (Farghaly & Shaalan, 2009; Shaalan et al., 2018).

Arabic language has linguistic features which escalate many more challenges that have influenced the development of language processing tools. These challenges are related to understanding the linguistic levels' characteristics of Arabic language to find computational solutions to process Arabic texts for extracting useful information. The main linguistic levels of Arabic Language are morphological, syntactic and semantic levels (Farghaly & Shaalan, 2009; Khalil & Osman, 2014; Shaalan et al., 2018). The following sub-sections provide a brief report on these linguistic levels.

3.1 Morphological level challenges

Arabic is a highly structured and derivational language where morphology plays a very important role, it has a greatly rich morphology depicted by a mix of templatic and affixational morphemes, complex morphological norms, and a rich part system. The Arabic language contains several rules and symbols of many meanings that differ with changes of tone, character, or punctuation marks. The Morphological level studies have two approaches, form-based and functional morphologies. Form-based morphology is about the form of units making up a word, their interactions with each other, and how they relate to the word's overall form. By contrast, functional morphology is about the function of units inside a word and how they affect its overall behavior syntactically and semantically (Khalil & Osman, 2014; Shaalan et al., 2018).

In the rich part system of Arabic language, the same Arabic word can be joined to various parts of appends and clitics to generate new vocabularies that make the Arabic words synonyms are widespread. As a result, Arabic language is considered as a highly inflectional and derivational language, which make the problem of ambiguity is one of the biggest challenges in Arabic NLP compared to many other languages (Khalatia & Al-Romanyb, 2020; Khalil & Osman, 2014; Shaalan et al., 2018; Zerrouki, 2020).

3.2 Syntactic level challenges

The syntax is a sentence structure study as an independent unit. This includes the word order, the dependency relationships between these words, and the agreement relationships. Syntax analyzers produce a formal description of how words come together to make phrases and sentences. Arabic is syntactically flexible; that is, it has a relatively free word order due to the morphology richness, which can express some syntactic relations. The order of syntactic relations could be Subject-Predicate-Object or Predicate-Subject-Object or Predicate-Object-Subject. These syntactic relation orders are all acceptable sentence structures (Khalatia & Al-Romanyb, 2020; Khalil & Osman, 2014; Maloney & Niv, 1998; Shaalan et al., 2018; Zerrouki, 2020).

An additional challenge related to the syntactic level is that Arabic has two types of sentences, which are verbal sentences and nominal sentences. Verbal sentences start with a verb. On the other hand, nominal sentences begin with the subject. This subject could be a definite noun, proper noun, or pronoun in the nominative case, and the predicate is an indefinite nominative noun, proper noun, or adjective that agrees with the subject in number and gender. The predicate can be a prepositional phrase (Zerrouki, 2020).

3.3 Semantic level challenges

Semantic is the study of linguistic expressions meaning. While syntax focuses on the linguistic form of the sentence without giving importance to the meaning, semantics focuses on the relationships between the components of the sentience on the whole meaning of the sentence. However, the boundaries of these two domains are not clear, and they can be considered as they complete each other. In other words, it is challenging to understand the meaning of the sentence without understanding its syntactic, especially within the case of syntactical or semantic ambiguities. There are three basic concepts related to the semantic level, Homonymy, Synonymy, and Semantic Roles. Homonymy is the state of two words having identical forms. They have the same spelling and pronunciation but different meanings. Synonymy is the state of two words having identical meanings but different forms. Semantic Role is the semantic relationship between a predicate and its arguments and satellite terms. Semantic Roles are also called thematic roles or theta roles (Zerrouki, 2020).

However, it is worth mentioning that there is a correlation between the morphological, syntactic, and semantic levels, as they all together help in making sense and meaning of words and in disambiguating the sentence. In fact, the NLP tasks for automating the process of analyzing Arabic sentences must deal with several complex problems which are related to the linguistic levels and the features nature of the Arabic language. If any Arabic IE application does not take the linguistic features of the Arabic language into account, it will certainly be inadequate for extracting beneficial information. Not only the lack of understanding of the linguistic levels' characteristics of Arabic language will make finding computational solutions to process Arabic texts for extracting a useful information a major problem, but also the lack of technical and linguistic resources for Arabic language computational solutions. (Farghaly & Shaalan, 2009; Shaalan et al., 2018).

4. ARABIC NLP TASKS AND ARABIC IE APPLICATIONS

It is necessary here to clarify exactly what is meant by NLP tasks and IE applications. In this survey study, Arabic NLP and Arabic IE are considered as a linguistic data process pipeline by using machines. This pipeline starts with linguistically pre-processing the Natural Language texts to be ready to extract useful information from those texts for a specific application by applying the computational linguistics principles. As Salloum et al. (2018) said, "The basic objective of National Language Processing is concerned with collecting information about the way computers evaluate and derive information from human languages to develop sophisticated software."

In addition, IE applications depend on Natural Language Pre-Processing tasks with semantic processing modules for extracting predefined types of information from a natural language text; in other words, IE could be seen as the activity of NLP. The authors of this survey study believe that the Natural Language Pre-Processing tasks can be categorized into two levels, low-level NLP tasks and high-level NLP tasks. These tasks should be processed on the Natural Language texts to be ready for developing IE applications, such as SA and Domain-Specific Relation Extraction (Nadkarni et al., 2011). In fact, the low-level NLP tasks are independent of the domain knowledge features and it is related to dealing with natural languages and their characteristics and levels; nevertheless, high-level NLP tasks are mostly dependent on the domain knowledge features. The Low-level NLP tasks can be including:

1. Tokenization or identifying individual words and punctuations.
2. Sentence splitting or Sentence boundary detection.
3. Part-Of-Speech or tagging Part-Of-Speech assignment to individual words.
4. Morphological or lemmatization or stemming analysis or decomposition of compound words into roots.
5. Shallow parsing or chunking.

The High-level NLP tasks can be including:

1. Coreference or Orthmatching resolution.
2. Words Normalization.
3. Named Entity Recognition
3. Anaphoric Resolution
5. Semantic Role Labelling
6. Word Sense Disambiguation (Named Entity Disambiguation)
7. Syntax parsing includes Dependency parsing and Constituency parsing
8. Semantic dependency parsing

On the other hand, IE applications can be developed to extract information from a diversity of domain knowledge. IE applications include:

1. Specific Application Terminology Extraction
2. Specific Application Relation Extraction
3. Specific Application Event Extraction
4. Sentiment Analysis and Opinion Mining
5. Machine Translation.
6. Text Summarization
7. Decision Support Systems
8. Question and Answering

Figure 1 depicts the processing flow to extract information from natural language texts.

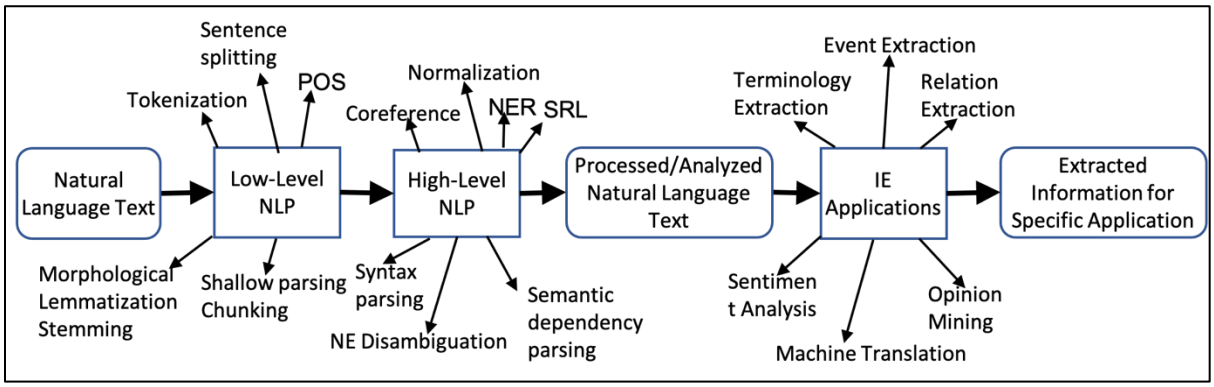

Figure 1. The Processing Flow of Extracting Information from Natural Language Texts

To clarify more, the NLP tasks are the process that deal with linguistic levels, which are, morphological, syntactic, and semantic levels of the language itself; however, the IE applications are employing the products of the NLP tasks to produce useful information for a specific application. This clarification agrees with Marie-sainte et al. (2019) that NLP is an area of investigation that aims to enable machines to analyze the natural human languages texts to perform valuable IE applications. They added that NLP research supports developing techniques and tools that allow software applications to handle natural languages and perform necessary tasks.

Despite that Arabic is a largely used language across the world, there is a lack of linguistic tools and resources, which make it still an under-resourced language. This lack of linguistic and technical resources has influenced negatively on investigating, researching, developing, and implementing the Arabic IE applications. Many studies and researches that have been done so far to extract beneficial information from Arabic texts are for academic and experimental purposes, and they do not adapt for the development process and end-user usage and cannot be integrated with existing systems (Zerrouki, 2020).

Recently, the tools for Arabic NLP tasks have gained increasing importance, and several state-of-the-art systems have been developed for a wide range of applications. However, there are a few of them are dedicated to supporting the Arabic language and they require more improvement to present more dedicated functions and features for handling Arabic texts. In fact, most Arabic NLP tools were originally developed for western languages. As a result, they are not easy to be adapted to the features nature of Arabic language. Improving these tools could be by adding Arabic modules to multilingual tools or adapt existing modules and libraries to be integrated with Arabic requirements (Farghaly and Shaalan, 2009; Shaalan et al., 2018, Zerrouki, 2020).

The authors of the survey study believe that improving the state-of-the-art Arabic NLP tools need more than adding Arabic modules or adapting exist modules in these tools. Improving the exist Arabic NLP tools need more formal and precise grammar of Arabic than the traditional grammar so widely employed today. That can be achieved by investigating the application of contemporary Linguistic Theories such as the Functional Grammar Theories or by Innovating and updating the heritage of traditional Arabic linguistics by preserving the value of their principle (Farghaly & Shaalan, 2009; Shaalan et al., 2018; Zerrouki, 2020).

There are two main approaches for most NLP tasks and IE applications, Rule-based and ML approaches, including the Deep Learning (DL) approaches. While Rule-based approaches rely on transforming the linguistic features space into lexical and syntactic patterns to be applied on natural language texts in order to extract information, ML approaches do not require deep linguistic skills and it is fast and inexpensive. Developers of such techniques, algorithms and tools had to deal with difficult issues and use trained classifiers to extract information from unstructured text. In the absence of complete computationally viable grammars of Arabic, statistical approaches that rely primarily on training data, such as ML systems usually give good results when the training set and the testing data are similar; however, there is no guarantee that more training data will make a significant improvement. Moreover, there may be some structures or entities that are scattered which will make the ML component does not have enough data to make the right generalization (Aljamel et al., 2019; Farghaly & Shaalan, 2009).

The availability of Arabic linguistic datasets resources is another problem facing designing and developing Arabic NLP tasks and Arabic IE applications. The lack of a sizable corpus of Arabic data corpora will make extracting information by applying both Rule-Based and Statistical approaches very difficult. Linguistic Corpora are very important in the advancement of different Arabic NLP tasks, such as Part of Speech (POS) parsing, Named Entities Recognition (NER), and Morphological Analysis. In fact, there are not enough Arabic Linguistic Corpora as compared to the existing English linguistic Corpora. However, efforts are being made to solve this problem. For example, the Multi Arabic Dialect Applications and Resources (MADAR) corpus. This large-scale collection of parallel sentences was built to cover the Dialect Arabic sources from several cities in the Arabic World in addition to Modern Standard Arabic sources. It contains more than 10000 training sentences and covers all 25 city dialects and MSA, and the other consists of 9,000 training sentences. This corpus can be used to train ML language models (Mansour, 2013; Obeid et al., 2020).

In this survey paper, besides the contributions, the recent advancement in methodologies, technical and linguistics resources in the fields of Arabic NLP and Arabic IE will be explored. The authors of this survey study will seek for the studies to provide a complete synopsis (outline) of the existing Arabic NLP techniques, which can be used for extracting useful information from the Arabic context for the target beneficiary groups or end-users. The approaches, techniques, algorithms and tools in the reviewed studies could be proceeded further to improve or develop techniques and methods that will be helpful for the analysis of Arabic text in large-scale systems.

5. METHOD

There are several review methodologies that can be applied in the survey studies. However, in this review article, a systematic review is conducted to investigate the existing literature on the Arabic NLP and Arabic IE research fields. The Systematic review ensures the transparency and coverage of all appropriate research by following a systematic protocol to guarantee the coverage of state-of-the-art research. That protocol follows a systematic and fair selection and evaluation

method. The use of explicit methods allows systematic reviews to aggregate a large body of research evidence in addition to assessing whether the effects or relationships are in the same direction and of the same general magnitude. Moreover, the Systematic review can explain the possible inconsistencies between study results and determine the strength of the overall evidence for every outcome of interest-based on the quality of the included studies and the general consistency among them (Abo et al., 2019; Pare et al., 2015; Paré & Kitsiou, 2016).

A comprehensive review of Paré and Kitsiou (2016) concluded that the main systematic review procedures have involved in the following steps:

Stage 1: Determining the inclusion and exclusion criteria for the identification of eligible reviewed articles.

Stage 2: Employing the appropriate academic search engines and academic databases to find eligible articles for the targeted research fields in survey study.

Stage 3: Analyzing the data extracted from the collected and reviewed articles, then discussing the results exploited from the analyzed data to present a summary of the findings of this survey study.

Before applying the stages of this methodology, the research questions and the objectives of this survey study will be presented in the next two sections.

5.1 The research questions

Following the steps of the systematic review procedures, the research questions are formulated. The main Research Question of this review article is:

"What are the state-of-the-art research trends in the fields of Arabic NLP Tasks and Arabic IE Applications?"

This Research Question is broken into four Research Questions, as pointed up in Table 1 below, along with the motivation of each various Research Question.

RQ#	Research Question	Motivation
RQ1	What are the main contributions in the fields of Arabic NLP tasks and Arabic IE applications?	To identify the main contributions in Arabic language studies for the fields of NLP tasks and IE applications.
RQ2	What are the Methodologies applied in Arabic NLP and Arabic IE studies to achieve the planned Contributions?	To identify the commonly applied Methodologies to achieve the Arabic NLP and IE contributions.
RQ3	Is there a shortage in the availability of the relevant tools, techniques, and algorithms for the research in the fields of Arabic NLP tasks and Arabic IE applications?	To find an aggregated list of relevant tools, techniques, and algorithms for Arabic NLP and Arabic IE research.
RQ4	Is there a shortage in the availability of the relevant dataset sources in the fields of Arabic NLP tasks and Arabic IE applications?	To present the availability of relevant datasets sources for Arabic NLP and Arabic IE research.

Table 1. The Research Questions

5.2 The research objectives

Depending on the previous Research Questions, the objectives of this review article can be summarized as follows:

1. Gathering the largest possible number of state-of-the-art research articles in the field of Arabic NLP tasks and Arabic IE applications.

2. Determining the methods and approaches, then defining the techniques, algorithms, and tools used to implement each method or approach. Then, finding an aggregated list of relevant datasets platforms for Arabic NLP and IE research.

3. Comparing the applied methodologies applied in Arabic NLP tasks and Arabic IE applications studies with the-state-of-the-art methodologies applied on other languages studies to recommend them for the researchers.

4. Determining the linguistic levels which are commonly applied or the linguistic problems which are commonly solved to extract information from Arabic language texts in each article.

5. Exploring the contribution facets which are achieved in the reviewed articles of the fields of Arabic NLP tasks and Arabic IE applications to identify the most successful contribution to be recommended to researchers who are working in solving similar research problems.

The application of the proposed methodology of this survey study will presented in next sections.

6. THE INCLUSION AND EXCLUSION CRITERIA OF SELECTING THE ARTICLES

According to the systematic review procedure aforementioned in this survey study, a search strategy based on explicit inclusion criteria for the identification of eligible studies should be developed. The Inclusion and exclusion criteria are used to select potentially relevant research papers from the dataset sources to answer the relevant research questions in a given systematic review study. In this survey study, Inclusion and exclusion criteria are developed and the conditions in

that Inclusion and exclusion criteria were applied to all retrieved articles to eliminate articles that are not in-line with the objectives of this survey study. The inclusion and exclusion criteria employed in this study are shown in Table 2 below:

Inclusion criteria	
1	The reviewed articles must be about NLP or IE that targeted Arabic Language texts in any knowledge domain.
2	If the reviewed article is about the Arabic IE application, it must be built by applying the Arabic NLP tasks (in Low or High levels or any linguistic level).
3	The reviewed articles must be published in English Language peer-reviewed Journals only.
4	The reviewed articles must be published within the last five years (between January 2017 and February 2021).
5	The reviewed articles must be completely available to the reviewers at any time.
Exclusion criteria	
1	The reviewed articles must not be review or survey study papers.
2	The reviewed articles must not be non-text form (audio or video classification).
3	The reviewed articles must not be books or sections or chapters of a book.
4	The reviewed articles must not a conference proceeding articles.
5	The reviewed articles must not target non-Arabic texts.

Table 2. The Inclusion and Exclusion Criteria

To assess the risk of bias of selecting studies and extracting data in a duplicate way, in this study, two independent reviewers have engaged in the review process of selecting studies and extracting data to avoid random or systematic errors in the systematic review procedure.

The next section will present the stage of employing the appropriate academic search engines and academic databases to find eligible articles for the targeted research fields in survey study.

7. SEARCHING FOR ELIGIBLE STUDIES

This section delivers details about the sources of the reviewed articles. Reviewers should carefully follow the steps of the systematic review in order to meet quality requirements. The aforementioned steps of the systematic review procedures state the searching for eligible articles should be using multiple databases and information sources, including grey literature sources, without any language restrictions. Academic search engines and Academic databases are the standard sources to access up-to-date scientific publications. The choice of one or more adequate Academic search engines and databases to meet the quality requirements is necessary because these sources determine the number of relevant articles that will be identified. They should be selected to provide the best coverage of the chosen search topic. However, coverage of Academic search engine or Academic database is denoted relative to a specific criterion. The systematic review guidance advises the use of suitable specialized databases that provide high coverage of a specific topic as well as generic resources that have broad coverage. Reviewers should thus consider their specific review topic when deciding which Academic search engines and Academic databases might prove suitable for a systematic search. The quality of search results of these sources depends on the keywords provided to them. The identification of these keywords is vital to find the targeted studies. Hence, extraordinary measures should be considered when selecting these keywords in the search terms (Gusenbauer, 2019 and Gusenbauer and Haddaway, 2020).

Following the steps of the systematic review procedures to find eligible articles, this survey article used academic search engines and academic databases as sources to access up to date scientific publications. The reviewers of this survey have considered the targeted research fields to decide which Academic search engines and Academic databases might prove suitable for a systematic search. Because the targeted research fields of this survey study are related to topics of applying NLP tasks and implementing the IE applications on the Arabic language texts in any domain knowledge. The keywords used to find the targeted studies from the Academic Search Engines and Academic Databases are (“Arabic Nature Language Processing”) OR ("Arabic Information Extraction") besides the keywords which can be derived from them, such as “Arabic NLP” and “Arabic IE”. Examples of these Academic search engines and Databases sources which are used to retrieve reviewed articles in this survey study are Google Scholar and Microsoft Academic search engines, and Emerald Insight, IEEE Access, and Springer Link Academic Databases.

The collected articles have been filtered by applying the inclusion and exclusion criteria to remove the undesirable articles, such as the conference articles, survey articles and other exclusion criteria (see Table 2). In addition, the process step of the systematic review procedures of selecting the articles and assessing the bias risk of the selection and extracting data from the selected and reviewed articles are equipped in a multiplicate manner using three independent reviewers to avoid random or systematic errors in the process. As forementioned, the reviewers of this survey study have selected articles to cover and be relevant to the field of Arabic NLP and Arabic IE. The number of final articles is 44 articles from all Academic Search Engines and Academic Databases. This final set of articles is targeted for analysis.

The reason for the number of collected and reviewed articles is only 44 articles is that the authors of this survey study have applied a very restricted inclusion and exclusion criteria. For example, the reviewed articles must be published in English Language peer-reviewed Journals only. It is common that the journals' procedure of accepting, and publishing articles takes long time. As a result, the researchers are reluctant to publish in peer-reviewed Journals and they prefer to publish in another form of publications such conferences proceedings. Figure 2 below depicts the number of articles published on Arabic NLP tasks and Arabic IE applications field per year.

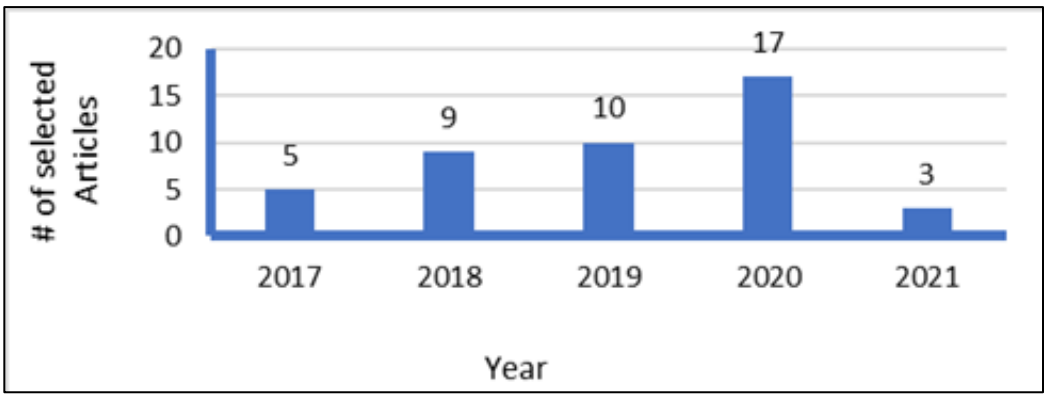

Figure 2. The Number of Articles Published on Arabic NLP tasks Fields and Arabic IE Applications Field per Year

The next section will introduce the last stage of this survey study's methodology, which is about presenting the results and analyzing the retrieved data, then discussing these results.

8. RESULTS AND ANALYSIS

This stage is about analyzing the data which are extracted from the collected and reviewed articles, then discussing the results exploited from the analyzed data to present a summary of the findings of this survey study. This section will be commenced embarked by presenting the classification scheme of those articles then discussion of the results; after that, a summary of this study findings is presented.

8.1 The Classification Scheme of the Reviewed Articles

The previous section addressed several research studies in the field of Arabic NLP tasks and Arabic IE applications. These studies cover different aspects presented in the literature, such as the NLP task POS, Tokenization, Morphological, and NER also in IE applications such as SA, Question and Answering (QA), and Text Classification. In this section, a characterization plan for Classification schemes to classify the reviewed articles is presented, as showed in Figure 3. The characterization plan comprises of five essential dimensions (D#), which are:

D1) The Achieved Contributions Dimension.
D2) The Applied Methodologies Dimension.
D3) The Used Tools, Techniques or Algorithms Dimension.
D4) The Dataset Source Domain Types Dimension.
D5) The Dataset Source Ownership Types Dimension.

The Figure 3 below is presenting the topics and categories which are applied to classify the reviewed articles:

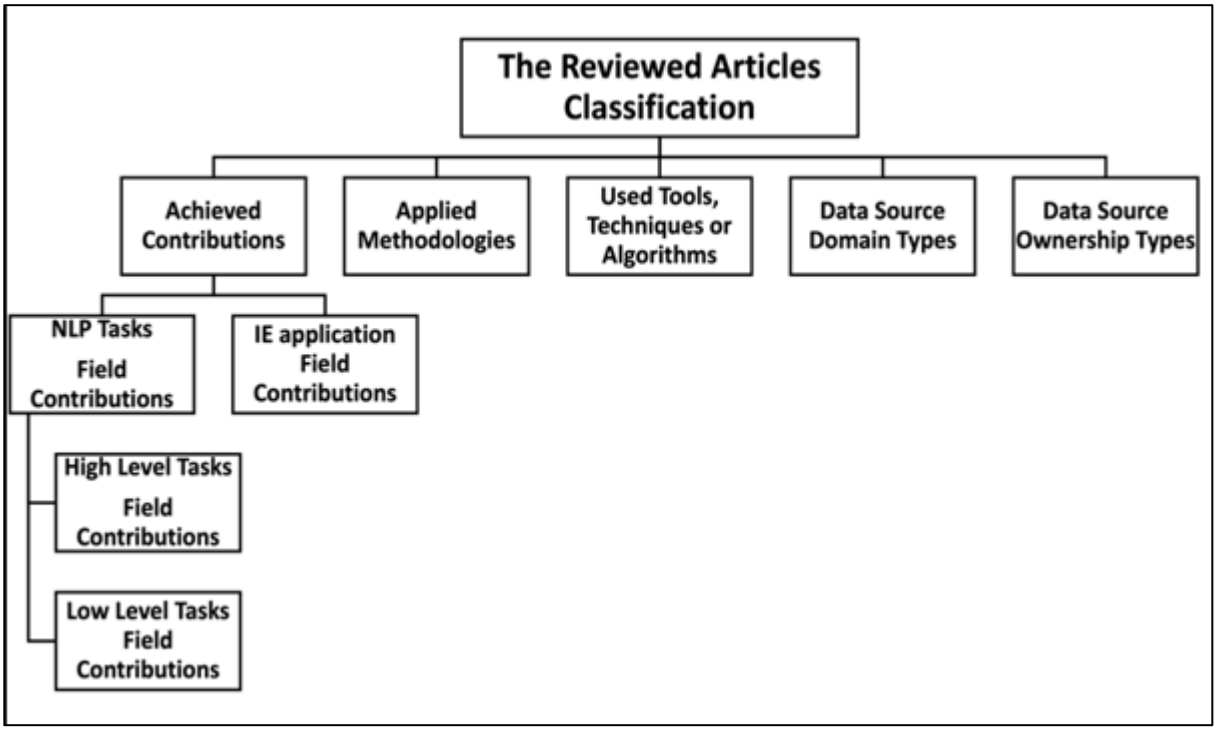

Figure 3. The Classification Categories of the selected Articles

In the next sections, the details of applying the above classification scheme on the collected articles will be presented.

D1) The Achieved Contributions Dimension

The reviewed articles are categorized in terms of the research fields, Arabic NLP tasks and Arabic IE applications. Then, the articles in these fields are classified according to their contribution areas. The Next two subsections present the details of these two contribution classifications.

NLP tasks are about processing and analysing the natural languages texts in many linguistic levels, such as, morphological, syntactic, and semantic levels. However, to better understand the linguistic levels, as aforementioned, these tasks are classified into two levels of NLP tasks, Low-Level NLP tasks and High-Level NLP tasks (see section 4.0 above). In this classification, the reviewed articles are classified in terms of the type of Arabic NLP tasks level contribution. Table 3 below shows the classified articles according to their NLP tasks contribution. They are seven contribution area's classes. Three classes in the Low-level NLP field and four contribution area's classes in the field of High-Level NLP. The description of those classes is as follows:

1. The first class is called "Tokenization". This class includes the articles that are contributed to the area of Tokenization.
2. The second class is called "POS". This class includes the articles that are contributed to the area of Part-Of-Speech parsing.
3. The third class is called "Morphological". This class includes the articles that are contributed to the area of Morphological resolution.
4. The fourth class is called "Coreference". This class includes the articles that are contributed to the areas of Coreference and Anaphora Resolving.
5. The fifth class is called "NER". This class includes the articles that are contributed to the area of Named Entity Recognition.
6. The sixth class is called "Dependency". This class includes the articles that are contributed to the areas of Dependency Parsing.
7. The seventh class is called "Nominal". This class includes the articles that are contributed to the area of Arabic Nominal sentences Parsing.

On the other hand, IE Applications depend on the outcomes of the Low and High levels of NLP tasks to produce useful information for a specific application (see section 4 above). In this class, the reviewed articles are classified in terms of the type of Arabic IE applications contribution such as SA applications. Table 3 below shows the classified articles according to their achieved contributions. They are five contribution area's classes. The description of those classes is as follows:

8. The first class is called "RE". It is for the articles which are contributed to the Specific domain Relation Extraction Application.
9. The second class is called "Sentiment". It is for the articles which are contributed to the SA applications that includes Opinion Mining, Emotions Extraction and Hate or offensive speech detection applications.
10. The third class is called "QA". It is for the articles which are contributed to Question and Answering applications.
11. The fourth class is called "TC". It is for the articles which are contributed to the Text Classification applications.
12. The fifth class is called "KM". It is for the articles which are contributed to the Knowledge Modelling applications.

No.	Research Fields	Contribution Areas Classes	Reviewed Articles	Articles No.
1	Low-Level NLP Tasks	Tokenization	(Almuhareb et al., 2019) [19]	1
2		POS	(Alqrainy & Alawairdhi, 2021) [22] (Abumalloh et al., 2018) [30]	2
3		Morphological	(Soudani et al., 2019) [5] (Alnaied et al., 2020) [6] (Thalji et al., 2018) [12] (Azman, 2019) [13] (Ben-Othman et al., 2020) [18] (Boudchiche & Mazroui, 2019) [25] (Ghembaza et al., 2018) [29]	7
4	High-Level NLP Tasks	Coreference	(Abolohom & Omar, 2017) [16]	1
5		NER	(Saadi & Belhadef, 2020) [1] (Khalil et al., 2020) [2] (Omar & Al-Tashi, 2018) [9] (Al-Smadi et al., 2020) [10] (Muhammad et al., 2020) [14] (Alshammari & Alanazi, 2020) [24] (Karaa & Slimani, 2017) [33] (Mohammed Nadher Abdo Ali et al., 2019) [35] (El Bazi & Laachfoubi, 2018) [36] (Mohammed N.A. Ali et al., 2018) [37] (ASBAYOU, 2020) [38] (Najeeb, 2020) [39] (Almarimi & Enbiah, 2020) [40]	13
6		Dependency	(S. Mohamed et al., 2021) [15]	1
7		Nominal	(Ababou et al., 2017) [27]	1
8	IE Applications	RE	(Zakria et al., 2019) [3] (Taghizadeh et al., 2018) [41]	2
9		Sentiment	(Abdullah et al., 2018) [4] (Alsafari et al., 2020) [7] (Ombabi et al., 2020) [8] (Guellil et al., 2020) [17] (Alswaidan & Menai, 2020) [21] (Eldin et al., 2020) [26] (Aljameel et al., 2021) [32]	7
10		QA	(Al-Smadi et al., 2019) [20] (AL-Shenak et al., 2019) [23] (Albarghothi et al., 2017) [42] (Bakari & Neji, 2020) [43] (Hamza et al., 2021) [44]	5
11		TC	(Alalyani & Marie-Sainte, 2018) [28] (Daoud & El-Seoud, 2017) [34]	2
12		KM	(Ghoniem et al., 2019) [11] (E. H. Mohamed & Shokry, 2020) [31]	2
			The Total Reviewed Articles	44

Table 3. The Classified Articles According to the Contribution Dimension

Figure 4 below, depicts and compares between classified contributions of the reviewed articles. This figure presents the number of articles in each class of the achieved contribution in the fields of the Arabic NLP tasks and the Arabic IE applications which are shown in Table 3 above.

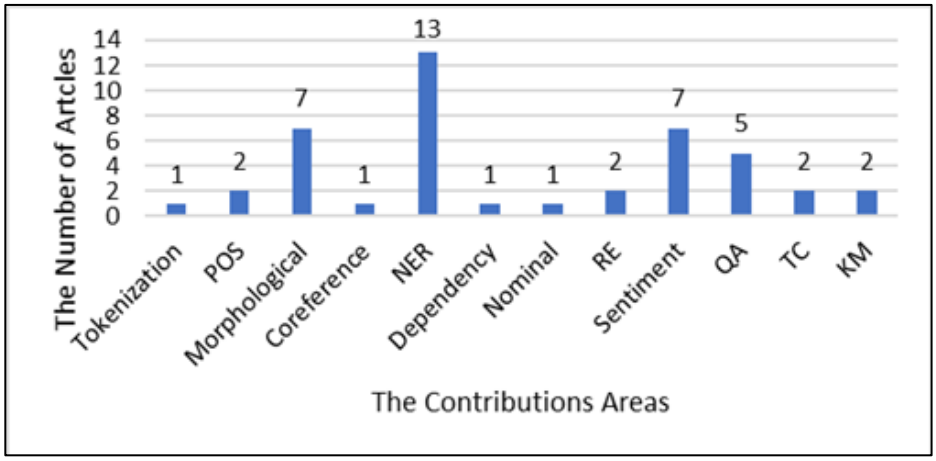

Figure 4. The Arabic NLP tasks and Arabic IE applications Contributions Classification Dimension of the reviewed Articles (map the contribution numbers to the numbers in Table 4)

D2) Applied Methodologies Classification Dimension:

These are several methodologies that can be applied to NLP tasks and IE applications. These methodologies are applied to achieve the planned contributions in the reviewed articles. The authors of this survey article extracted their contributions facets from the reviewed articles, then they are classified according to the methodology applied dimension as shown in Table 4 below. The description of those classes is as follows:

1. The first class is called "Rule-Based". This class includes the articles that apply Rule-based, Linguistic Pattern, Pattern Techniques and Linguistic-Based Formal Definition approaches.
2. The second class is called "Hybrid". This class includes the hybrid between linguistic approaches and statistical approaches.
3. The third class is called "ML, ANN, DL". This class includes the articles that apply ML Based, ANN Modelling or DL Based approaches.
4. The fourth class is called "Corpus-Based". This class includes the articles that apply Corpus-Based, Lexicon Based or Linked Open Data Based approaches.
5. The fifth class is called "SWT". This class includes articles that apply Ontology Learning or Semantic Web Technology-based approaches.
6. The sixth class is called "Statistical". This class includes articles that apply approaches based on Special Statistical Models.
7. The seventh class is called "Optimisation". This class includes articles that apply approaches based on Optimisation Problems or Evolutionary Algorithms.

No.	Contribution Areas Classes	Reviewed Articles	Articles No.
1	Rule-Based	(Khalil et al., 2020) [2] (Alnaied et al., 2020) [6] (Omar & Al-Tashi, 2018) [9] (Thalji et al., 2018) [12] (Abolohom & Omar, 2017) [16] (Ben-Othman et al., 2020) [18] (Al-Smadi et al., 2019) [20] (Alqrainy & Alawairdhi, 2021) [22] (Ababou et al., 2017) [27] (Ghembaza et al., 2018) [29] (ASBAYOU, 2020) [38] (Almarimi & Enbiah, 2020) [40]	12
2	Hybrid	(Soudani et al., 2019) [5] (Karaa & Slimani, 2017) [33]	2
3	ML, NN, DL	(Saadi & Belhadef, 2020) [1] (Zakria et al., 2019) [3] (Abdullah et al., 2018) [4] (Alsafari et al., 2020) [7] (Ombabi et al., 2020) [8] (Al-Smadi et al., 2020) [10] (Muhammad et al., 2020) [14] (Guellil et al., 2020) [17] (Almuhareb et al., 2019) [19] (Alswaidan & Menai, 2020) [21] (AL-Shenak et al., 2019) [23] (Abumalloh et al., 2018) [30] (E. H. Mohamed & Shokry, 2020) [31] (Aljameel et al., 2021) [32] (Mohammed Nadher Abdo Ali et al., 2019) [35] (El Bazi & Laachfoubi, 2018) [36] (Mohammed N.A. Ali et al., 2018) [37] (Taghizadeh et al., 2018) [41] (Hamza et al., 2021) [44]	19
4	Corpus Based	(Azman, 2019) [13] (S. Mohamed et al., 2021) [15] (Alshammari & Alanazi, 2020) [24] (Alalyani & Marie-Sainte, 2018) [28] (Daoud & El-Seoud, 2017) [34]	5
5	SWT	(Ghoniem et al., 2019) [11] (Albarghothi et al., 2017) [42] (Bakari & Neji, 2020) [43]	3
6	Statistical	(Boudchiche & Mazroui, 2019) [25] (Eldin et al., 2020) [26]	2
7	Optimisation	(Najeeb, 2020) [39]	1
		The Total Reviewed Articles	44

Table 4. The Classified Articles According to The Applied Methodology and Approaches

Figure 5 below presents and compares between the classes of the methodologies which were applied in the reviewed articles which are shown in Table 4 above. These methodology classes are Rule-based methodologies class, ML Based, ANN Modelling and DL Based methodologies class, Hybrid approaches class, Corpus-Based methodologies class, Semantic Web Technology based methodologies class, Special Statistical Models methodologies class; lastly, Evolutionary Algorithms based methodologies class. As depicted in Table 4

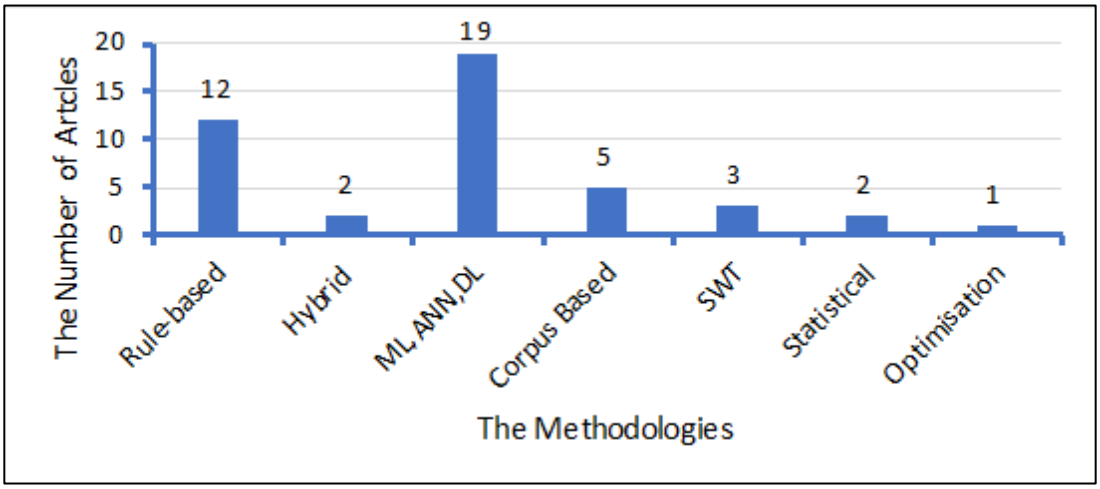

Figure 5. Applied Methodologies Classification Dimension of the Reviewed Articles

D3) The Used Tools, Techniques or Algorithms Classification Dimension

Arabic NLP Tools, Techniques, and Algorithms are very necessary resources for Arabic NLP tasks and Arabic IE applications. They often employ diversity of linguistic and statistical features in the language Knowledge. However, there is a shortage in the availability of the relevant tools, techniques, and algorithms because of the lack of research participants to build them for the Arabic Language Field. In this classification dimension, the reviewed articles are classified in terms of the used Tools, Techniques or Algorithms in the reviewed articles, whether these tools, techniques, or algorithms are specifically developed for analysing Arabic Language texts or for any Language texts as depicted in Table 5 below. The description of those classes is as follows:

1. The first class is called "Arabic Language Specific". This class includes the articles that utilise Arabic Language Specific Tools, Techniques or Algorithms.
2. The second class is called "Generic Languages". This class includes the articles that utilise Generic Languages Tools, Techniques or Algorithms applied.
3. The third class is called "Developed". This class includes the articles that utilise developed tools, techniques, or algorithms as a contribution.

No.	Technical Type Classes	Reviewed Articles	Articles No.
1	Arabic Language Specific	(Khalil et al., 2020) [2] GATE, (Soudani et al., 2019) [5] MorphToolKit & MADAMIRA, (Omar & Al-Tashi, 2018) [9] Stanford Arabic Parser, (S. Mohamed et al., 2021) [15] Arabic PADT & FARASA, (Abolohom & Omar, 2017) [16] Arabic Statistical POS Tagger (ASPOST), (Al-Smadi et al., 2019) [20] AraNLP & FARASA, (Alswaidan & Menai, 2020) [21] Natural Language ToolKit (NLTK), (Alshammari & Alanazi, 2020) [24] AMIRA tool, (Boudchiche & Mazroui, 2019) [25] The Morphosyntactic Analyser Alkhalil Morpho System, (Eldin et al., 2020) [26] ATKS tools, FARASA library, Stanford Arabic Parser, (Almarimi & Enbiah, 2020) [40] NooJ, (Albarghothi et al., 2017) [42] Python API scripts and Stanford Java API to maintain Arabic for Arabic NLP, (Bakari & Neji, 2020) [43] Stanford Parser, Arabic NER (ArNER), Khoja Stemmer	13
2	Generic Languages	(Saadi & Belhadef, 2020) [1] Deep Neural Network-based Arabic NER, (Zakria et al., 2019) [3] Naive Bayes Classifier Algorithm, (Abdullah et al., 2018) [4] WEKA, (Alsafari et al., 2020) [7] ML and DL Algorithms, (Ombabi et al., 2020) [8] CNN and LSTM Algorithms, (Al-Smadi et al., 2020) [10] DL Algorithm, (Ghoniem et al., 2019) [11] Genetic-Whale Optimization Model, (Muhammad et al., 2020) [14] Conditional Random Field (CRF) and Structured Support Vector Machine (SSVM) algorithms for Arabic NER, (Guellil et al., 2020) [17] ML and DL Algorithms, (Almuhareb et al., 2019) [19] DL Algorithm, (AL-Shenak et al., 2019) [23] ML Algorithms, (Alalyani & Marie-Sainte, 2018) [28] WEKA, (Abumalloh et al., 2018) [30] ANN Algorithm, (Aljameel et al., 2021) [32] ML Algorithms, (Karaa & Slimani, 2017) [33] ML Algorithms, (Mohammed Nadher Abdo Ali et al., 2019) [35] ANN Algorithm, (Mohammed N.A. Ali et al., 2018) [37] Recurrent Neural Network (RNN), (Najeeb, 2020) [39] The Genetic algorithm (GA), (Taghizadeh et al., 2018) [41] ML Algorithm, (Hamza et al., 2021) [44] ML Algorithms, ANN Modelling.	20
3	Developed	(Alnaied et al., 2020) [6] AMIR algorithm, (Thalji et al., 2018) [12] Rule-Based Root Extraction Algorithm for Arabic Language, (Azman, 2019) [13] The system of RootIT for verb root identification, (Ben-Othman et al., 2020) [18] Arabic Verbs Roots and Conjugation Automation Tool, (Alqrainy & Alawairdhi, 2021) [22] Arabic Morph Syntactic Tagger (AMT), (Ababou et al., 2017) [27] Parsing Arabic Nominal Sentences	11

		Tool, (Ghembaza et al., 2018) [29] Linguistic-Based Morphological Analysis Approach, (E. H. Mohamed & Shokry, 2020) [31] A Quranic Semantic Search Tool (QSST), (Daoud & El-Seoud, 2017) [34] The Classified (Arabic) Ads through SMS (CATS) system, (El Bazi & Laachfoubi, 2018) [36] Topic models as features for Arabic NER system, (ASBAYOU, 2020) [38] Rule-Based Arabic NER and classification system,	
The Total Reviewed Articles			**44**

Table 5. The Classified Articles According to the Tools, Techniques and Algorithms used

Figure 6 below, graphically presents and compares the data in Table 5. It shows the number of articles in each class, which are Arabic Language specific class, Generic Language Class and Developed Class of tools, techniques or algorithms classifications dimension.

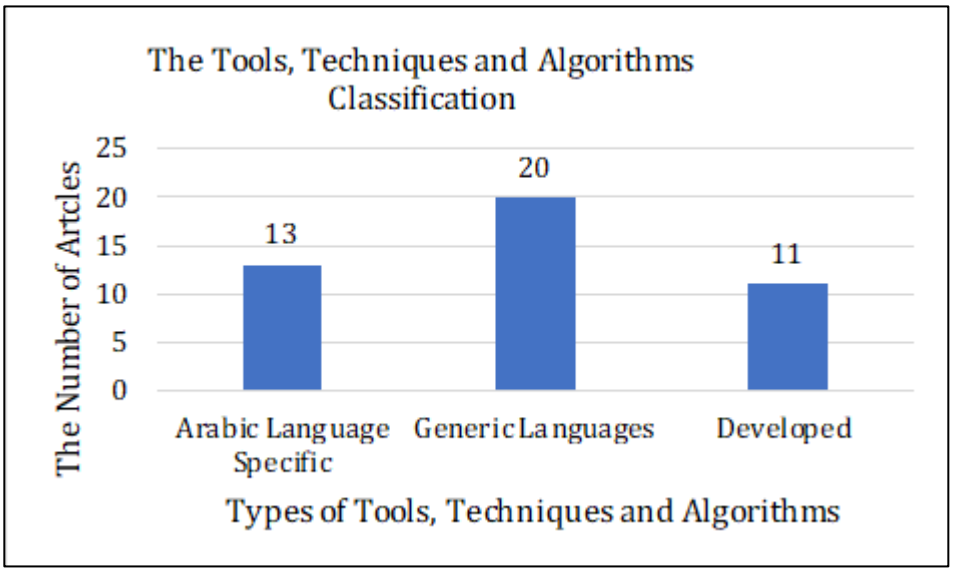

Figure 6. The Used Tools, Techniques, and Algorithms in the Reviewed Articles Classification Dimension

D4) Data Sources Domain Type Classification Dimension

NLP tasks and IE applications often employ many of knowledge resources such as data or datasets sources which are very important NLP and IE resources. These datasets; usually, support the NLP tasks and IE applications by providing linguistic and statistical features. For example, lexical concept hierarchies, verb roots, POS tags, Question sentences and others. It is believed that there is a shortage of relevant dataset sources for Arabic NLP and Arabic IE because of the lack of research participants to create Arabic dataset sources. In this classification dimension, the reviewed articles are classified in terms of the dataset sources domain type. Table 6 below shows the classified articles according to the dataset sources domain types dimension. The description of those classes is as follows:

1. The first class is called "Online". This class includes the articles that utilise Dataset Sources Domain Types of Online News, Broadcast News, and Newswire, Online Newspapers, Wikipedia articles, World Wide Web as a Large Documents Corpus or social media Text.
2. The second class is called "Linguistic". This class includes the articles that utilise Dataset Sources Domain Types of Arabic roots corpus, Arabic verbs, Lexical Markup Framework based dictionaries and corpora, Corpora of Linguistic Annotated, Tagged sentences (or/and) words (or/and) stop-words (or/and) Tokens (or/and) NEs (or/and) nominal sentences or annotated Quran (or Hadith) Corpus.
3. The third class is called "Review". This class includes the articles that utilise Dataset Sources Domain Types of Annotated Arabic Opinions, reviews or hate speech datasets.
4. The fourth class is called "Classification". This class includes the articles that utilise Dataset Sources Domain Types of documents dataset (classified into classes), Text Classification Dataset, SMS message or Arabic Questions Dataset.

No.	Data Sources Domain Type Classes	Reviewed Articles	Articles No.
1	Online	(Saadi & Belhadef, 2020) [1] (Khalil et al., 2020) [2] (Zakria et al., 2019) [3] (Abdullah et al., 2018) [4] (Alnaied et al., 2020) [6] (Alsafari et al., 2020) [7] (Omar & Al-Tashi, 2018) [9] (Ghoniem et al., 2019) [11] (Alswaidan & Menai, 2020) [21] (Aljameel et al., 2021) [32] (Karaa & Slimani, 2017) [33] (Almarimi & Enbiah, 2020) [40] (Taghizadeh et al., 2018) [41]	13
2	Linguistic	(Soudani et al., 2019) [5] (Al-Smadi et al., 2020) [10] (Thalji et al., 2018) [12] (Azman, 2019) [13] (Muhammad et al., 2020) [14] (S. Mohamed et al., 2021) [15] (Abolohom & Omar, 2017) [16] (Ben-Othman et al., 2020) [18] (Almuhareb et al., 2019) [19] (Alqrainy & Alawairdhi, 2021) [22] (Alshammari & Alanazi, 2020) [24] (Boudchiche & Mazroui, 2019) [25] (Ababou et al., 2017) [27] (Ghembaza et al., 2018) [29] (Abumalloh et al., 2018) [30] (E. H. Mohamed & Shokry, 2020) [31] (Mohammed Nadher Abdo Ali et al., 2019) [35] (El Bazi & Laachfoubi, 2018) [36] (Mohammed N.A. Ali et al., 2018) [37] (ASBAYOU, 2020) [38] (Najeeb, 2020) [39]	21
3	Review	(Ombabi et al., 2020) [8] (Guellil et al., 2020) [17] (Eldin et al., 2020) [26]	3
4	Classification	(Al-Smadi et al., 2019) [20] (AL-Shenak et al., 2019) [23] (Alalyani & Marie-Sainte, 2018) [28] (Daoud & El-Seoud, 2017) [34] (Albarghothi et al., 2017) [42] (Bakari & Neji, 2020) [43] (Hamza et al., 2021) [44]	7
The Total Reviewed Articles			44

Table 6. Dataset Sources Domain Type Classification

Figure 7 Figure 7. The Dataset Sources Domain and Type Classification Dimension in the Reviewed Articlesbelow shows what domain types of dataset sources have been used by the authors of the reviewed articles, which are presented in Table 6. Figure 7 below shows graphically the number of the reviewed articles in each class of dataset source domain types classification dimension. Which are Online, Linguistics, Reviews and Text Classification classes of types and domains of dataset sources.

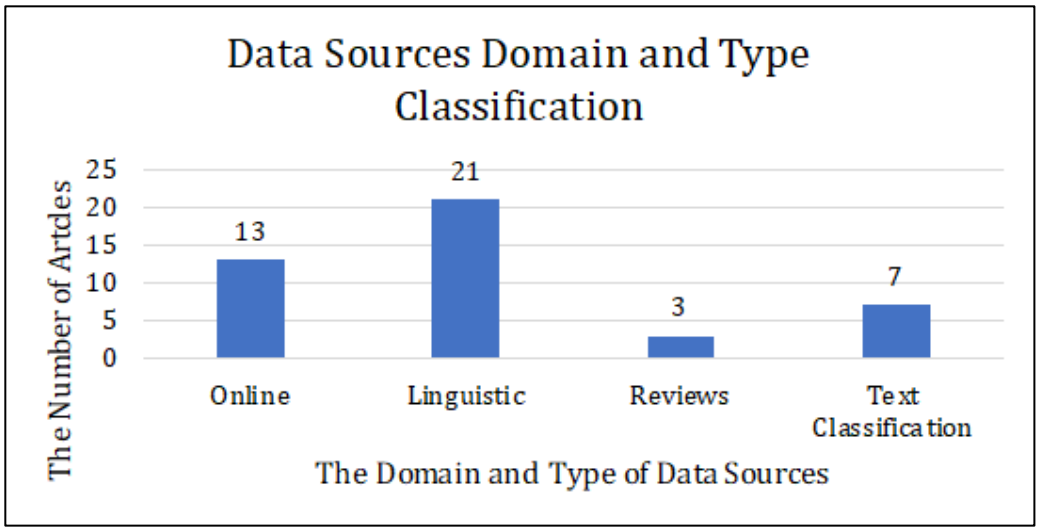

Figure 7. The Dataset Sources Domain and Type Classification Dimension in the Reviewed Articles

D5) Dataset Sources Ownership Types Classification Dimension

In this classification dimension, the reviewed articles are classified in terms of the ownership type of dataset sources. Table 7 below shows the classified reviewed articles according to three categories of the dataset's ownership types. The description of those classes is as follows:

1. The first class is called "Authors". It is for the articles which used datasets and these datasets are collected by the Authors of these articles.
2. The second class is called "Other Parties". It is for the articles which used datasets and these datasets are collected and created by another party.
3. The third class is called "Contributions". It is for the articles which their authors are collected the datasets as a contribution for these articles.

No.	Data Sources Ownership Type Classes	Reviewed Articles	Articles No.
1	Authors	(Khalil et al., 2020) [2] (Zakria et al., 2019) [3] (Alsafari et al., 2020) [7] (Guellil et al., 2020) [17] (Ben-Othman et al., 2020) [18] (Al-Smadi et al., 2019) [20] (Alqrainy & Alawairdhi, 2021) [22] (Eldin et al., 2020) [26] (Ababou et al., 2017) [27] (Ghembaza et al., 2018) [29] (Abumalloh et al., 2018) [30] (Karaa & Slimani, 2017) [33] (Daoud & El-Seoud, 2017) [34] (Najeeb, 2020) [39] (Almarimi & Enbiah, 2020) [40] (Albarghothi et al., 2017) [42] (Bakari & Neji, 2020) [43] (Hamza et al., 2021) [44]	18
2	Other Parties	(Saadi & Belhadef, 2020) [1] KALIMAT, (Abdullah et al., 2018) [4] CSIT-2016, (Soudani et al., 2019) [5] Morpho-semantic graphs of Tashkeela, ZAD and Al-muâajam dictionary, (Alnaied et al., 2020) [6] EveTAR (2016) dataset, (Ombabi et al., 2020) [8] ElSahar and El-Beltagy dataset and Arabic NLTK dataset, (Omar & Al-Tashi, 2018) [9] Saif & Aziz dataset, [10] WikiFANE$_{GOLD}$ (Ghoniem et al., 2019) [11] (ACE) corpora, ANERcorp dataset, (Thalji et al., 2018) [12], Thalji, (Azman, 2019) [13] Al-ShAh dataset, (Muhammad et al., 2020) [14] ANERCrop Corpus, (Abolohom & Omar, 2017) [16] Qurapro Corpus, (Almuhareb et al., 2019) [19] Arabic Treebank (ATB) data and an ATB clitics segmentation schema, (Alswaidan & Menai, 2020) [21] the AETD dataset, IAEDS dataset and the SemEval-2018 dataset, (AL-Shenak et al., 2019) [23] Arabic Document Dataset, (Boudchiche & Mazroui, 2019) [25] The Nemlar corpus, (E. H. Mohamed & Shokry, 2020) [31] Arabic Original Qur'an" dataset, (Aljameel et al., 2021) [32] Twitter dataset, (Mohammed Nadher Abdo Ali et al., 2019) [35] ANERCorpus, (El Bazi & Laachfoubi, 2018) [36] AQMAR, (Mohammed N.A. Ali et al., 2018) [37] ANERcorp, (ASBAYOU, 2020) [38] ANERCorp corpus and French Press Agency (FPA) Corpus, (Taghizadeh et al., 2018) [41] ACE-2004 dataset	23
3	Contributions	(S. Mohamed et al., 2021) [15] (Alshammari & Alanazi, 2020) [24] (Alalyani & Marie-Sainte, 2018) [28]	3
The Total Reviewed Articles			44

Table 7. The Classified Articles According to the dataset's ownership type used to Achieve or Evaluate Their Contribution.

Figure 8 below graphically illustrates the data in Table 7. The figure below and the table above present the number of reviewed articles in each class of the ownership types of dataset sources dimension. They are authors ownership type class, other party ownership type class and authors contribution dataset type class.

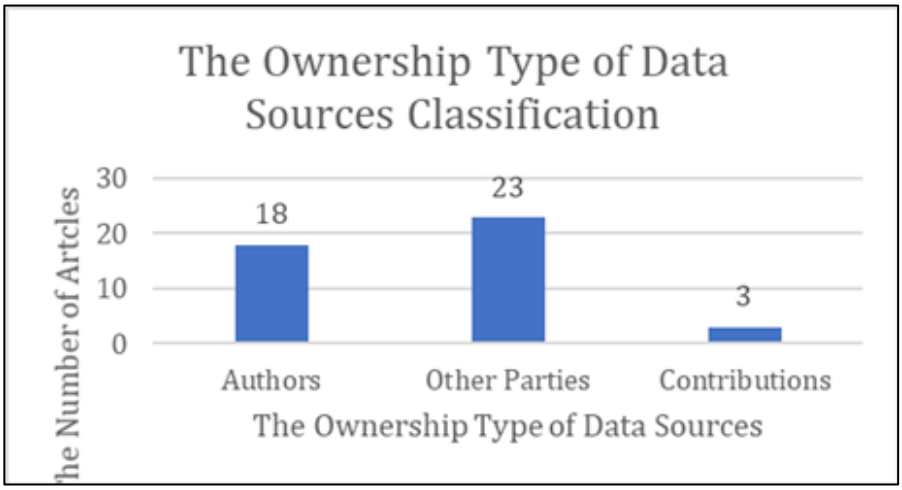

Figure 8. The Ownership Type of Dataset Sources Classification Dimension in the Reviewed Articles

8.2 The Analysis and discussion

As explained earlier, the reviewed articles in this survey study are classified according to a classification scheme, which are: the achieved contributions dimension, the applied methodologies dimension, the utilised techniques, algorithms or tools dimension, and the used dataset sources domain types dimension, and the used dataset sources ownership types dimension. Then the classified articles are organized and visualized in tables and figures as shown in the previous section.

Returning to the questions posed at the beginning of this study, the central question of this survey study asks about the state-of-the-art research trends in the fields of Arabic NLP tasks and Arabic IE applications. This research question has been broken into four research questions, as shown in Table 1 above. It is now possible in the sections that follow to answer those research questions by discussing, analyzing, and evaluating the results which are summarized in the abovementioned tables and figures.

RQ1) What are the main contributions in the fields of Arabic NLP tasks and Arabic IE applications?

The first question in this study sought to determine the contribution trends in the fields of low and high levels of Arabic NLP and Arabic IE. In the low level of Arabic NLP field and as shown in Table 3 and Figure 4, the highest number of articles' contribution is in the Morphological Resolution task of the Arabic words' category. It represents around 15.9% of all reviewed articles (44 Articles). Then, the articles contribution in POS parsing with 4.5%; lastly, the articles' contribution in tokenization tasks with 2.3% from all reviewed articles. This result may be explained by the fact that studying the Morphological Resolution of the Arabic words is very important in Arabic language. Morphology plays a very important role in making the machines understand the linguistics features of Arabic Language texts because Arabic language is a highly structured and derivational language.

As shown in Table 3 and Figure 4, the reviewed articles' contribution trend in the high-level Arabic NLP tasks is in the NER tasks. It represents around 29.7% of all reviewed articles. Then comes the articles which have contributions in the tasks of coreference resolution, anaphora resolving, dependency parsing, and the Arabic nominal sentences parsing contributions with of 2.3% each of the reviewed articles. This result may partly be explained by the fact that NER is very important task for IE applications such as a specific domain terminology extraction application, a specific domain relations extraction application and SA application. Researchers in analyzing Arabic language texts prefer to investigate the Arabic NER in spite that it is a difficult problem because of the various characteristics of the Arabic language.

For the reviewed articles which have a contribution in the field of Arabic IE applications, it is clear from Table 3 and Figure 4 that the contribution trend is related to Sentient Analysis, Opinion Mining, Emotions Extraction, Detecting Hate Speech, Hate and offensive speech detection applications. They represent 15.9% from all reviewed articles. After that, the contributions in the applications question and answering with 11.3%. In the meanwhile, the contributions in the specific domain relation extraction applications, text classification applications and domain knowledge applications represent 4.5% of all reviewed articles for each application category. The reason for this could be because this kind of applications are based, on most articles, on some linguistic features and named entities in the texts, which are part of the low and high levels of NLP tasks of Arabic texts. In particular, the NER task which is a research trend for the high-level NLP tasks; consequently, there will be an abundant of Arabic NER tools that will be used to create Arabic SA application.

It is also worth noting that the current study was not designed to evaluate the contributions' achievements of the reviewed articles; nevertheless, this study is limited by evaluating the contributions' trends in the fields of Arabic NLP and Arabic IE. In fact, the reviewers of this survey study rely on peer-reviewed Journals for evolution the contributions

achievements of the selected articles as it is one of conditions in the Inclusion and exclusion criteria were applied to all retrieved articles.

RQ2) What are the applied Methodologies in Arabic NLP and Arabic IE studies to achieve the planned Contributions?

The answer to this research question attempts to find the trends in using the methodologies and approaches to achieve planned contributions in the reviewed articles. It is apparent from Table 4 and Figure 5 that most of reviewed articles were applying methodologies and approaches that are based on ML, Artificial Neural Networks (ANN) and DL Algorithms. The number of articles which apply this kind of methodologies category represent around 43.3% of all reviewed articles. The next methodology class is the rule-based methodologies class. The number of articles which apply the rule-based methodologies represent around 27.3% of the whole reviewed articles. Then the corpus-based methodologies which represent around 11.3%, Then the Semantic Web Technology based methodologies which represent around 6.8%, then the Hybrid based methodologies which represent 4.5% and Evolutionary Algorithms based methodologies which represent 2.3%. A possible explanation for this might be that the methodologies which are based on ML, NN and DL algorithms do not require deep linguistics skills and they can be applied more effectively than other approaches which require hand-crafting rule sets such as Rule-based methodologies. Another possible explanation for this is that these algorithms are not a language specific and can be applied to substitute the shortage of the Arabic language specific algorithms-based methodologies.

RQ3) Is there a shortage in the availability of the relevant tools, techniques, or algorithms for the research in the fields of Arabic NLP tasks and Arabic IE applications?

Referring to the third research question, this study has explored the reviewed articles to find out what kind of tools, techniques and algorithms which are utilized to solve problems or conduct experiments in the fields of Arabic NLP tasks and Arabic IE application. In addition, the study attempts to confirm whether there is a shortage of tools, techniques, or algorithms which are developed specifically to manipulate the linguistic features of Arabic Language texts. As shown in Table 5 and Figure 6, the reviewed articles are classified into three classes of tools, techniques, and algorithms. The first class is for the applied Tools, Techniques or Algorithms which are developed specifically to process the linguistic features of Arabic Language texts; for example, MADAMIRA, Stanford Arabic Parser, Arabic PADT, FARASA library, NooJ, Khoja Stemmer and others (See Table 5). The number of articles in the first class represents 29.7% from all reviewed articles. The second class is for the applied Tools, Techniques or Algorithms which are not developed for processing texts in a specific language; for example, ML and DL algorithms based NLP tools, The Genetic algorithm (GA) based NLP tools, and others. The number of articles in the second class represents 45.3% from all reviewed articles. The third class is for the applied Tools, Techniques, or algorithms which are developed as a contribution for the reviewed articles; for example, AMIR algorithm, Rule-Based Root Extraction Algorithm for Arabic Language, the system of RootIT for verb root identification, Arabic Verbs Roots and Conjugation Automation Tool, Arabic Morph Syntactic Tagger (AMT), Parsing Arabic Nominal Sentences Tool, Linguistic-Based Morphological Analysis Approach, A Quranic Semantic Search Tool (QSST), and others. The number of articles in the third class represents 25% from all reviewed articles.

It can be seen from the classified articles in Table 5 and Figure 6 that the authors of around half of the reviewed articles have applied Tools, Techniques, or Algorithms that are not developed for processing texts of a specific language. These articles are represented in second class by 45.5% of all reviewed articles. In practicable, these Tools, Techniques or Algorithms can be configured to be applied for processing any natural language which require less efforts than developing Tools, Techniques or Algorithms exclusively for processing Arabic texts. For this reason, authors desire using ready, available, configurable Tools, Techniques or Algorithms than developing Tools, Techniques or Algorithms.

RQ4) Is there a shortage in the availability of the relevant dataset sources in the fields of Arabic NLP tasks and Arabic IE applications?

Dataset sources are crucial for research in the fields of NLP tasks and IE applications. They are used to enrich the Arabic research and assess the results of the applied methodologies; besides of evaluating the tools and techniques and training and validating the algorithms. With respect to the fourth research question, this survey study attempts to focus on the availability and variety of dataset sources for the studies of Arabic Language specific NLP tasks and IE applications. In this survey study, not only the reviewed articles have been classified in terms of dataset sources' domain types dimension, as shown in Table 6 and Figure 7 but also they have been classified in terms of dataset sources' ownership types dimension, as shown in Table 7 and Figure 8. From the abovementioned data in tables and figures, it is apparent that there are a diversity of domain types and ownership types of the dataset sources or datasets used in the reviewed articles.

Table 6 and Figure 7 reveal that nearly half of the reviewed articles (47.6%) have used datasets of annotated texts with Linguistic features such as Arabic roots, Arabic verbs, Lexical Markup, tagged sentences or tokens, tagged NEs and others. In addition, a diversity of domains of online dataset sources are used by around 29.7% of the reviewed articles. These online dataset sources include social media Text, online broadcast news, Wikipedia articles and others. In another domain type of dataset sources, complete documents datasets are used by around 15.9% of the reviewed articles. The contents of these datasets include, Text Classification Dataset, SMS message datasets, Arabic Questions Datasets, and others. it is

also noted that around 6.8% of the reviewed articles use annotated datasets in the domain of Arabic opinions or reviews or hate speech datasets.

Table 6 and Figure 7; however, provides an overview about the distribution of data sources' ownership types which are utilized in all reviewed articles. They show that the number of reviewed articles that their authors exploited other researchers' or owners' datasets for their study is more than half of the reviewed articles (52.3%). Examples of these datasets are, KALIMAT, CSIT-2016, ZAD and Al-muâajam dictionary, ElSahar and El-Beltagy dataset and Arabic NLTK dataset, ACE-2004 dataset, ANERcorp dataset, Thalji dataset, Arabic Original Qur'an" dataset, Twitter dataset and others. However, 40.9% of the reviewed articles used datasets that their authors collected and created these datasets for their study by themselves. It is worth to mention that the contribution of some articles is building datasets for deferent purposes; however, the number of these articles is only three which represent around 6.8% of all the reviewed articles.

To better understand the trends of domain and ownership types of the used dataset sources, the authors of this survey study have classified the reviewed articles in the domain type and the contribution type classes according to the ownership types. This classification is developed for the purpose of investigating the relationship between the domain types and ownership types of dataset sources in one side and the relationship between the contribution types and ownership types of dataset sources in another side in the reviewed articles. Table 8 and Figure 9 below provides the breakdown the reviewed articles in each dataset domain type class into their ownership type classes.

	Domain Types	**Authors Ownership type Articles**	**Other parties Ownership type Articles**	**Contributions Ownership type Articles**	
1	Online	(Khalil et al., 2020) [2] (Zakria et al., 2019) [3] (Alsafari et al., 2020) [7] (Karaa & Slimani, 2017) [33] (Almarimi & Enbiah, 2020) [40]	(Saadi & Belhadef, 2020) [1] (Abdullah et al., 2018) [4] (Alnaied et al., 2020) [6] (Omar & Al-Tashi, 2018) [9] (Ghoniem et al., 2019) [11] (Alswaidan & Menai, 2020) [21] (Aljameel et al., 2021) [32] (Taghizadeh et al., 2018) [41]	--	13
		5	8	0	
2	Linguistic	(Ben-Othman et al., 2020) [18] (Alqrainy & Alawairdhi, 2021) [22] (Ababou et al., 2017) [27] (Ghembaza et al., 2018) [29] (Abumalloh et al., 2018) [30] (Najeeb, 2020) [39]	(Soudani et al., 2019) [5] (Al-Smadi et al., 2020) [10] (Thalji et al., 2018) [12] (Azman, 2019) [13] (Muhammad et al., 2020) [14] (Abolohom & Omar, 2017) [16] (Almuhareb et al., 2019) [19] (Boudchiche & Mazroui, 2019) [25] (E. H. Mohamed & Shokry, 2020) [31] (Mohammed Nadher Abdo Ali et al., 2019) [35] (El Bazi & Laachfoubi, 2018) [36] (Mohammed N.A. Ali et al., 2018) [37] (ASBAYOU, 2020) [38]	(S. Mohamed et al., 2021) [15] (Alshammari & Alanazi, 2020) [24]	21
		6	13	2	
3	Reviews	(Guellil et al., 2020) [17] (Eldin et al., 2020) [26]	(Ombabi et al., 2020) [8]	--	3
		2	1	0	
4	Classification	(Al-Smadi et al., 2019) [20] (Daoud & El-Seoud, 2017) [34] (Albarghothi et al., 2017) [42] (Bakari & Neji, 2020) [43] (Hamza et al., 2021) [44]	(AL-Shenak et al., 2019) [23]	(Alalyani & Marie-Sainte, 2018) [28]	7
		5	1	1	
The Total Reviewed Articles					44

Table 8. The Classification of the domain Type Classes of Reviewed Articles According to Ownership Type of the Dataset Sources

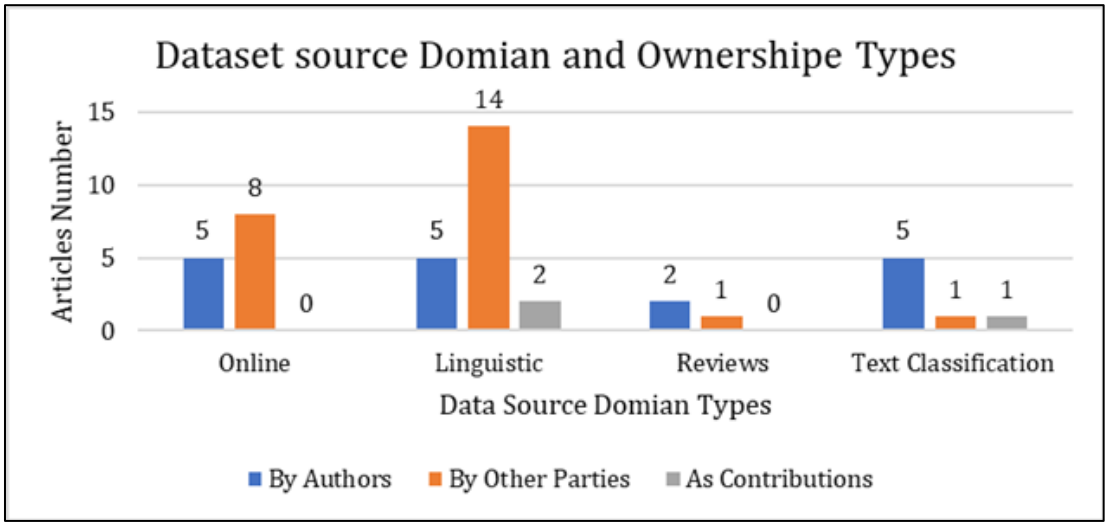

Figure 9. The Classification of the domain Type Classes of Reviewed Articles According to Ownership Type of the Dataset Sources

It is apparent from the data in Table 8 and Figure 9 above that the largest number of the reviewed articles is for the articles which are using other parties ownership dataset sources of different kinds of Linguistic annotations domain type. They are 14 articles that represent around 31.9% of the whole reviewed articles. In addition, from Figure 9 above, it can be seen that the number of reviewed articles which are using other parties ownership dataset sources of online sources domain is greater than the number of reviewed articles which are using authors ownership dataset sources of the online sources domain type. However, it is clear from the data in Table 8 and Figure 9 that the number of reviewed articles which are using authors ownership dataset sources in both reviews and text classification domain types are greater than the other reviewed articles which are using other parties and contributions ownership type dataset sources. Table 9 and Figure 10 below provide the breakdown of the reviewed articles in each contribution type class into ownership type classes.

	Contribution Type	Authors Ownership type Articles	Other parties Ownership type Articles	Contributions Ownership type Articles	
1	**Low-Level NLP Tasks**	(Ben-Othman et al., 2020) [18] (Alqrainy & Alawairdhi, 2021) [22] (Ghembaza et al., 2018) [29] (Abumalloh et al., 2018) [30]	(Soudani et al., 2019) [5] (Alnaied et al., 2020) [6] (Thalji et al., 2018) [12] (Azman, 2019) [13] (Almuhareb et al., 2019) [19] (Boudchiche & Mazroui, 2019) [25]	--	10
		4	6	0	
2	**High-Level NLP Tasks**	(Khalil et al., 2020) [2] (Ababou et al., 2017) [27] (Karaa & Slimani, 2017) [33] (Najeeb, 2020) [39] (Almarimi & Enbiah, 2020) [40]	(Saadi & Belhadef, 2020) [1] (Omar & Al-Tashi, 2018) [9] (Al-Smadi et al., 2020) [10] (Muhammad et al., 2020) [14] (Abolohom & Omar, 2017) [16] (Mohammed Nadher Abdo Ali et al., 2019) [35] (El Bazi & Laachfoubi, 2018) [36] (Mohammed N.A. Ali et al., 2018) [37] (ASBAYOU, 2020) [38]	(S. Mohamed et al., 2021) [15] (Alshammari & Alanazi, 2020) [24]	16
		5	9	2	
3	**IE Applications**	(Zakria et al., 2019) [3] (Alsafari et al., 2020) [7] (Guellil et al., 2020) [17] (Al-Smadi et al., 2019) [20] (Eldin et al., 2020) [26] (Daoud & El-Seoud, 2017) [34] (Albarghothi et al., 2017) [42] (Bakari & Neji, 2020) [43] (Hamza et al., 2021) [44]	(Abdullah et al., 2018) [4] (Ombabi et al., 2020) [8] (Ghoniem et al., 2019) [11] (Alswaidan & Menai, 2020) [21] (AL-Shenak et al., 2019) [23] (E. H. Mohamed & Shokry, 2020)	(Alalyani & Marie-Sainte, 2018) [28]	18

			[31] (Aljameel et al., 2021) [32] (Taghizadeh et al., 2018) [41]		
		9	8	1	
				The Total Reviewed Articles	44

Table 9. The Classification of the domain ownership type Classes of Reviewed Articles According to the Article Contributions

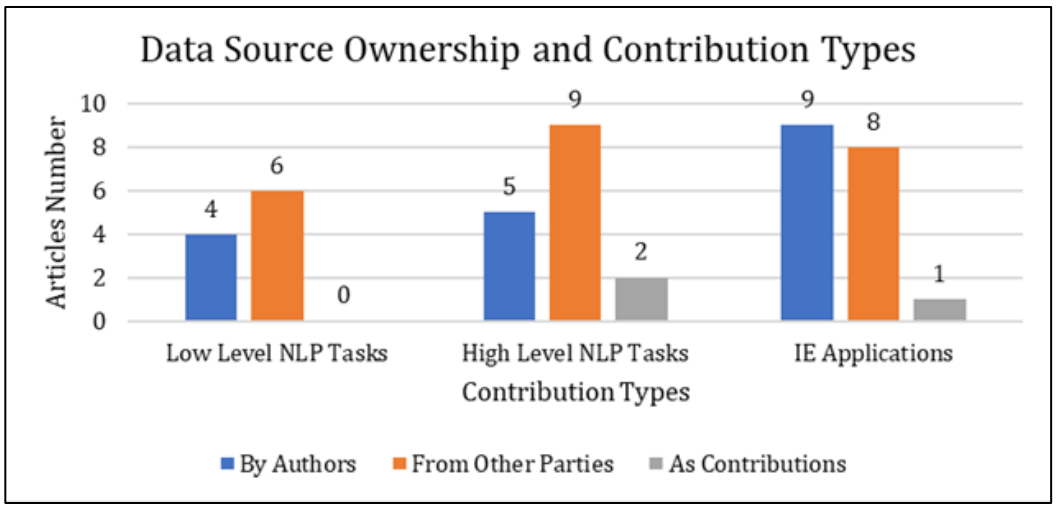

Figure 10. The Classification of Contribution Type Classes of Reviewed Articles According to Ownership Type of the Dataset Sources

Also, it can be seen in Table 9 and Figure 10, that the number of the articles, which are contributed to the fields of both Levels of Arabic NLP tasks and using other parties ownership dataset sources, are more than the articles which are using the other both dataset sources ownership types, dataset sources collected by the authors ownership type and dataset sources created as a contribution ownership type. However, the number of the articles, which are contributed to the fields Arabic IE applications and using dataset sources collected by the authors are more than the articles which are using dataset sources of both ownership types, other parties' ownership dataset sources type and dataset sources created as a contribution ownership type.

As abovementioned, the reviewed articles that using other parties' ownership dataset sources, which are Linguistic annotations domain type have the highest number of articles comparing to the other classes. This may partly be explained by not only the fact that building dataset with tagging many kinds of linguistic features in linguistic texts requires computational efforts, is time consuming and requires linguistic and technical specialists in the field of Arabic NLP and Arabic IE. This can be confirmed from the data in Table 9 and Figure 10. It can be seen that the articles which are contributed to the fields of NLP tasks and used datasets created by other parties are more than the articles which used datasets created by their authors. However, the articles which are contributed to the fields of IE application and used datasets created by their authors are more than the articles which used datasets created by other parties. It is because the required datasets in the fields of NLP tasks are related to Linguistic annotations domain more than in the fields of IE applications.

It is noteworthy to mention that not all authors who are building datasets for their studies make these datasets available for other researchers. For example, all the authors of reviewed articles in the class of authors' dataset sources ownership type have never mentioned the availability of their datasets to other researchers. This could be because making the datasets available for other researchers requires more efforts to be ready for manipulating by other tools, techniques or algorithms including the research copyrights process.

8.3 The findings summary and recommendations

The main goal of the current survey study is to report the latest research trends in the fields of Arabic NLP tasks and Arabic IE applications. This includes the contribution achieved, the methodologies applied and, the technical and linguistic resources utilized. The aim is to find and recommend how to fill the research gaps in these fields. Nevertheless, the evaluability of contributions and achievements of the reviewed articles has not been discussed because the authors have relied on the credibility of the journals in which the reviewed articles were published. For this reason, the authors adopted the inclusion and exclusion criteria of selecting the articles. One of these criteria is that the reviewed articles must be

published in peer-reviewed English Language Journals only. In fact, it is believed that Academic Journals is more trustable than other kind of publishing methods. This study has gone some way towards enhancing our understanding of the research gaps in methodologies applied and, the technical and linguistic resources used in the fields of Arabic NLP tasks and Arabic IE applications. The summary of the extracted findings from the targeted reviewed studies will be presented below, after that, the suggested recommendations will be stated. The summary of those findings is:

1. The most obvious finding to emerge from this study is that most of the researchers in the field of Arabic NLP prefer to contribute to the NER tasks; after that, to contribute to the task of the morphological resolution of the Arabic words' category. However, the contribution trend in the IE applications field is related to Sentient Analysis. However, it can be confirmed that the research in the fields of low-level NLP tasks is very small; for examples, the Tokenization, Sentence's splitter and POS parsing tasks.
2. The second major finding was that most of reviewed articles were applying methodologies and approaches that are based on ML, ANN and DL Algorithms. The next number of articles in applied methodology type classes is for the Rule-based methodologies class.
3. One of the more significant findings to emerge from this study is that authors of around half of the reviewed articles have applied Tools, Techniques or Algorithms that are not for specific language because these Tools, Techniques or Algorithms can be configured to be used for processing any natural language.
4. This is the first study of substantial duration which examines associations between the dataset sources domain types and dataset sources ownership types in addition to the relation between articles' contribution fields and the datasets ownership types. The contribution of this study has been to confirm that the reviewed articles that use other parties' ownership dataset sources of Linguistic domain type have the highest number of articles comparing to the other dataset sources ownership types in the dataset sources domain types. This can be confirmed more by finding that the authors prefer utilizing the ready and available dataset sources when the contribution is related to Arabic NLP tasks, and they prefer collecting and creating their own dataset sources when the contributions are related to Arabic IE applications.

The findings of this study have two important implications for future practice which are related to the required technical and linguistic resources to assist processing the Arabic texts. Firstly, these findings suggest several courses of action for improving the exist Arabic NLP techniques or developing new techniques to increase the availability of the Arabic-Specific tools, techniques, and algorithms resources for processing Arabic Texts. Continued efforts are needed to more formal and precise grammar of Arabic than the traditional grammar so widely employed in the existing techniques. That can be achieved by Innovating and updating the heritage of traditional Arabic linguistics by preserving the valuable of their principles or by investigating the application of the modern Linguistic Theories such as the Arabic Functional Grammar Theories.

The second important practical implication is ensuring to provide an appropriate support for Arabic linguistic dataset sources. Supporting Arabic linguistic dataset sources should be a priority for all researchers interested in investigating the Arabic NLP tasks and Arabic IE applications. The Arabic linguistic dataset sources important in assessing, training, and validating the improved or developed linguistic tools, techniques, and algorithms. As a result, more linguistic dataset sources should be made available to enrich the contributions in the fields of Arabic NLP tasks and Arabic IE applications. This can be achieved not only by encouraging researchers to make their linguistic dataset sources available to other researchers but also by encouraging the cooperation between the researchers to combine their efforts to produce high-quality linguistic dataset sources to support a diversity of research studies areas.

9. CONCLUSIONS AND FUTURE WORKS

This survey study has examined the literature on the fields of Arabic NLP tasks and Arabic IE applications to summarize and analyze the state-of-the-art trends in these fields; hence, identifying the gaps which required to be fulfilled in these fields. This study set out to gather the largest possible number of state-of-the-art research articles in the targeted fields. Subsequently, these articles were surveyed to obtain information about the contribution achieved, the methodologies applied, the technical and linguistic dataset resources utilized in the gathered articles.

This review article study has followed systematic review procedure steps to meet the requirements of high-quality survey studies. The main systematic review procedure steps which are applied in this survey study as follows. Firstly, developing a search strategy based on explicit inclusion criteria for the identification of eligible studies. Secondly, gathering eligible studies using multiple databases and information sources and assessing risk of bias in a duplicate manner using more than one reviewer. Thirdly, analyzing and discussing the collected data and presenting a summary of the results in proper methods. Lastly, interpreting the results into findings conclusions and recommendations.

The source of the selected and reviewed articles in this survey study is Academic Search Engines and Academic Databases. This final set of articles which were targeted for analysis contains 44 articles from all areas of Arabic NLP tasks and Arabic IE applications. The reason for the number of collected and reviewed articles is only 44 articles is that the authors of this survey study have applied a very restricted inclusion and exclusion criteria. For example, the reviewed articles must be published in English Language Journals only and conference proceedings articles were not included. Nevertheless, the

evaluability of contributions achievements of the reviewed articles has not been discussed because the authors have relied on the credibility of the journals in which the reviewed articles were published in.

The collected and reviewed articles cover different aspects which are presented in the literature of the field of Arabic NLP tasks such as the NLP tasks POS, Tokenization, Morphological, and NER; besides, the aspects which are presented in the literature of the field of Arabic IE applications such as SA, QA, and Text Classification. These articles were classified by applying a classification scheme. This scheme comprises of four essential dimensions, which are the achieved contribution dimension, the applied Methodologies dimension, the utilized tools, techniques and algorithm dimension, the type of the used dataset sources domain types dimensions and the type of the used dataset sources ownership dimension. Following the systematic review procedure steps, the results, which were acquired from the classified articles, were presented to be analyzed and discussed to obtain the findings conclusions and recommendations.

This study has shown that the common area of research in the Arabic NLP field is the NER tasks and then to the task of the Arabic words' morphological resolution tasks. Nevertheless, the common area in the field of Arabic IE applications is the Sentient Analysis applications. It is noteworthy to mention that research in the areas of low-level Arabic NLP tasks is not very common between the researchers such as the Tokenization, Sentence's splitter, and POS parsing tasks. Secondly, the most of reviewed articles were applying methodologies and approaches that are based on ML NN and DL Algorithms. Thirdly, the results of this investigation show that around half of the reviewed articles have applied Tools, Techniques or Algorithms that are not specifically designed for processing Arabic language texts. Lastly, this study provides the first comprehensive assessment which examines associations between the dataset sources domain types and dataset sources ownership types in addition to the relation between articles' contribution fields and the datasets ownership types. It confirms that the reviewed articles that using other parties' ownership dataset sources of Linguistic domain type have the highest number of articles comparing to the other dataset sources domain and ownership types. Furthermore, the authors prefer using the ready and available dataset sources when the contribution is related to Arabic NLP tasks, and they prefer collecting their own dataset sources when the contributions are related to Arabic IE applications. As a results of these findings, greater efforts are needed to ensure that the provision of both technical and linguistic resources should be considered when studying Arabic NLP tasks and Arabic IE applications. This could be realized by improving or developing Arabic text processing tools, techniques, and algorithms. Also, it can be realized by encouraging researcher to make their dataset sources available to other researchers and by encouraging researchers to collaborate to produce a high-quality linguistic dataset source.

The project was limited in several ways. The major limitation of this survey study is subjected to the publishers' scope of this survey study. This study has only considered the English peer-reviewed Journal articles. It is unfortunate that the study did not include conference proceedings articles and articles which are written in Arabic Language. Another limitation could be related to that this study has only evaluated the contributions' trends in the fields of Arabic NLP and Arabic IE; in particular, this survey study did not evaluate the contributions' achievements of the reviewed articles. Considerably further work will need to be done to overcome these limitations. The study should be repeated to include more articles' publishers such as the conference proceedings articles and that articles written in Arabic Language. This will increase the number of the selected and reviewed articles and could develop a deeper understanding of the research trends in the fields of Arabic NLP and Arabic IE. Moreover, further studies need to be carried out in order to validate and evaluate the contribution of the selected and reviewed articles, specifically, if the publishers' scope of these articles are augmented by the conference proceedings articles and Arabic Language articles. A greater focus on augmented the publishers' scope of the selected and reviewed articles with evaluating and validating the contributions on those articles could produce interesting findings that account more for the research trends in the fields of Arabic NLP tasks and Arabic IE applications.

References

Ababou, N., Mazroui, A., & Belehbib, R. (2017). Parsing Arabic Nominal sentences using context free grammar and fundamental rules of classical grammar. *International Journal of Intelligent Systems and Applications*, *9*(8), 11–24. https://doi.org/10.5815/ijisa.2017.08.02

Abdullah, M., AlMasawa, M., Makki, I., Alsolmi, M., & Mahrous, S. (2018). Emotions extraction from Arabic tweets. *International Journal of Computers and Applications*, *42*(7), 661–675. https://doi.org/10.1080/1206212X.2018.1482395

Abo, M. E. M., Raj, R. G., Qazi, A., & Zakari, A. (2019). Sentiment Analysis for Arabic in Social Media Network: A Systematic Mapping Study. *ArXiv Preprint, ArXiv ID: 1911.05483*.

Abolohom, A., & Omar, N. (2017). A Computational Model for Resolving Arabic Anaphora using Linguistic Criteria. *Indian Journal of Science and Technology. Publisher: Indian Society for Education and Environment.*, *10*(3), 1–6. https://doi.org/10.17485/ijst/2017/v10i3/110637

Abumalloh, R. A., AlSerhan, H. M., BinIbrahim, O., & AbuUlbeh, W. (2018). Arabic Part-of-Speech Tagger, an Approach Based on Neural Network Modelling. *International Journal of Engineering & Technology. Publisher: Science Publishing Corporation*, *7*(2.29), 742. https://doi.org/10.14419/ijet.v7i2.29.14009

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2019). A comprehensive survey of arabic sentiment analysis. *Information Processing and Management. Pergamon*, *56*(2), 320–342. https://doi.org/https://doi.org/10.1016/j.ipm.2018.07.006

AL-Shenak, M., Nahar, K. M. O., & Halawani, K. M. H. (2019). Aqas: Arabic question answering system based on svm, svd, and lsi. *Journal of Theoretical and Applied Information Technology. Little Lion Scientific*, *97*(2), 681–691. https://doi.org/ISSN: 1992-8645

Al-Smadi, M., Al-Dalabih, I., Jararweh, Y., & Juola, P. (2019). Leveraging Linked Open Data to Automatically Answer Arabic Questions. *IEEE Access*, *7*(March), 177122–177136. https://doi.org/10.1109/ACCESS.2019.2956233

Al-Smadi, M., Al-Zboon, S., Jararweh, Y., & Juola, P. (2020). Transfer Learning for Arabic Named Entity Recognition with Deep Neural Networks. *IEEE Access*, *8*, 37736–37745. https://doi.org/10.1109/ACCESS.2020.2973319

Alalyani, N., & Marie-Sainte, S. L. (2018). NADA: New Arabic dataset for text classification. *International Journal of Advanced Computer Science and Applications. Publisher: The Science and Information (SAI) Organization*, *9*(9), 206–212. https://doi.org/10.14569/ijacsa.2018.090928

Alam, T. M., & Awan, M. J. (2018). Domain Analysis of Information Extraction Techniques. *INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING*, *9*(6), 1–9.

Albarghothi, A., Khater, F., & Shaalan, K. (2017). Arabic Question Answering Using Ontology. *Procedia Computer Science*, *117*, 183–191. https://doi.org/10.1016/j.procs.2017.10.108

Ali, Mohammed N.A., Tan, G., & Hussain, A. (2018). Bidirectional recurrent neural network approach for arabic named entity recognition. *Future Internet*, *10*(12), 1–12. https://doi.org/10.3390/fi10120123

Ali, Mohammed Nadher Abdo, Tan, G., & Hussain, A. (2019). Boosting Arabic Named-Entity Recognition with Multi-Attention Layer. *IEEE Access*, *7*, 46575–46582. https://doi.org/10.1109/ACCESS.2019.2909641

Alian, M., Awajan, A., & Al-kouz, A. (2017). Arabic Word Sense Disambiguation - Survey. *International Conference on New Trends in Computing Sciences (ICTCS), 11-13 October 2017*, *November 2019*. https://doi.org/10.1109/ICTCS.2017.23

Aljameel, S. S., Alabbad, D. A., Alzahrani, N. A., Alqarni, S. M., Alamoudi, F. A., Babili, L. M., Aljaafary, S. K., & Alshamrani, F. M. (2021). A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent covid-19 outbreaks in Saudi Arabia. *International Journal of Environmental Research and Public Health. Publisher: Multidisciplinary Digital Publishing Institute (MDPI)*, *18*(1), 1–12. https://doi.org/10.3390/ijerph18010218

Aljamel, A., Osman, T., Acampora, G., Vitiello, A., & Zhang, Z. (2019). Smart Information Retrieval: Domain Knowledge Centric Optimization Approach. *IEEE Access*, *7*(MI), 4167–4183. https://doi.org/10.1109/ACCESS.2018.2885640

Almarimi, A. A., & Enbiah, E. M. (2020). Recognition System for Libyan Entity Names. *European Journal of Electrical Engineering and Computer Science*, *4*(6), 1–5. https://doi.org/10.24018/ejece.2020.4.6.263

Almuhareb, A., Alsanie, W., & Al-Thubaity, A. (2019). Arabic Word Segmentation With Long Short-Term Memory Neural Networks and Word Embedding. *IEEE Access*, *7*, 12879–12887. https://doi.org/10.1109/ACCESS.2019.2893460

Alnaied, A., Elbendak, M., & Bulbul, A. (2020). An intelligent use of stemmer and morphology analysis for Arabic information retrieval. *Egyptian Informatics Journal*, *21*(4), 209–217. https://doi.org/10.1016/j.eij.2020.02.004

Alqrainy, S., & Alawairdhi, M. (2021). Towards developing a comprehensive tag set for the Arabic language. *Journal of Intelligent Systems*, *30*(1), 287–296. https://doi.org/10.1515/jisys-2019-0256

Alsafari, S., Sadaoui, S., & Mouhoub, M. (2020). Hate and offensive speech detection on Arabic social media. *Online Social Networks and Media*, *19*(September), Article 100096. https://doi.org/10.1016/j.osnem.2020.100096

Alshammari, N., & Alanazi, S. (2020). An Arabic dataset for disease named entity recognition with multi-annotation schemes. *Data. Publisher: Multidisciplinary Digital Publishing Institute (MDPI)*, *5*(3), 1–8. https://doi.org/10.3390/data5030060

Alswaidan, N., & Menai, M. (2020). Hybrid Feature Model for Emotion Recognition in Arabic Text. *IEEE Access*, *8*, 37843–37854. https://doi.org/10.1109/ACCESS.2020.2975906

ASBAYOU, O. (2020). Automatic Arabic Named Entity Extraction and Classification for Information Retrieval. *International Journal on Natural Language Computing*, *9*(6), 1–22. https://doi.org/10.5121/ijnlc.2020.9601

Azman, B. (2019). Root Identification Tool for Arabic Verbs. *IEEE Access*, *7*, 45866–45871. https://doi.org/10.1109/ACCESS.2019.2908177

Azmi, A. M., Al-qabbany, A. O., & Hussain, A. (2019). Computational and natural language processing based studies of hadith literature : a survey. *Artificial Intelligence Review*, *52*(2), 1369–1414. https://doi.org/10.1007/s10462-019-09692-w

Bakari, W., & Neji, M. (2020). A novel semantic and logical - based approach integrating RTE technique in the Arabic question – answering. *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-020-09684-0

Ben-Othman, M. T., Al-Hagery, M. A., & El-Hashemi, Y. M. (2020). Arabic Text Processing Model: Verbs Roots and Conjugation Automation. *IEEE Access*, *8*, 103913–103923. https://doi.org/10.1109/ACCESS.2020.2999259

Boudchiche, M., & Mazroui, A. (2019). A hybrid approach for Arabic lemmatization. *International Journal of Speech Technology*, *22*(3), 563–573. https://doi.org/10.1007/s10772-018-9528-3

Chowdhury, G. (2003). Natural Language Processing. In *The Annual Review of Information Science and Technology* (Vol. 37). https://doi.org/ISSN 0066-4200

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, *12*, 2461–2505.

Daoud, D. M., & El-Seoud, M. S. A. (2017). Employing information extraction for building mobile applications. *International Journal of Interactive Mobile Technologies*, *11*(2), 99–112. https://doi.org/10.3991/ijim.v11i2.6569

El Bazi, I., & Laachfoubi, N. (2018). Arabic Named Entity Recognition using topic modeling. *International Journal of Intelligent Engineering and Systems*, *11*(1), 229–238. https://doi.org/10.22266/ijies2018.0228.24

Eldin, S. S., Mohammed, A., Eldin, A. S., & Hefny, H. (2020). An enhanced opinion retrieval approach via implicit feature identification. *Journal of Intelligent Information Systems*. https://doi.org/10.1007/s10844-020-00622-9

Farghaly, A., & Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and Solutions. *ACM Transactions on Asian Language Information Processing*, *8*(4), 1–19. https://doi.org/10.1145/1644879.1644881

Fasha, M., Obeid, N., & Hammo, B. (2017). A Proposed Model for Extracting Information from Arabic-Based Controlled Text Domains. *Proceedings of the New Trends in Information Technology (NTIT), 25-27 April 2017*, 86–92.

Ghembaza, M. I. E., Aloufi, K. S., & Smai, A. (2018). Arabic Solid-Stems for an Efficient Morphological Analysis. *Arabian Journal for Science and Engineering*, *43*(12), 7373–7383. https://doi.org/10.1007/s13369-017-2938-8

Ghoniem, R. M., Alhelwa, N., & Shaalan, K. (2019). A novel hybrid genetic-whale optimization model for ontology learning from Arabic text. *Algorithms. Publisher: Multidisciplinary Digital Publishing Institute (MDPI)*, *12*(9), 1–32. https://doi.org/10.3390/a12090182

Guellil, I., Adeel, A., Azouaou, F., Chennoufi, S., Maafi, H., & Hamitouche, T. (2020). Detecting hate speech against politicians in Arabic community on social media. *International Journal of Web Information Systems. Emerald Publishing*, *16*(3), 295–313. https://doi.org/10.1108/IJWIS-08-2019-0036

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods. Wiley Online Library*, *11*(2), 181–217. https://doi.org/10.1002/jrsm.1378

Hamza, A., En-Nahnahi, N., Zidani, K. A., & El Alaoui Ouatik, S. (2021). An arabic question classification method based on new taxonomy and continuous distributed representation of words. *Journal of King Saud University - Computer and Information Sciences*, *33*(2), 218–224. https://doi.org/10.1016/j.jksuci.2019.01.001

Karaa, W., & Slimani, T. (2017). A new approach for arabic named entity recognition. *International Arab Journal of Information Technology*, *14*(3), 332–338.

Khalatia, M. M., & Al-Romanyb, T. A. H. (2020). Artificial Intelligence Development and Challenges (Arabic Language as a Model). *International Journal of Innovation, Creativity and Change*, *13*(5), 916–926.

Khalil, H., & Osman, T. (2014). Challenges in information retrieval from unstructured arabic data. *Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014*, 456–461. https://doi.org/10.1109/UKSim.2014.115

Khalil, H., Osman, T., & Miltan, M. (2020). Extracting Arabic Composite Names Using Genitive Principles of Arabic Grammar. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *19*(4), 1–16. https://doi.org/10.1145/3382187

Maloney, J., & Niv, M. (1998). TAGARAB: A Fast, Accurate Arabic Name Recogniser Using High Precision Morphological Analysis. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 8–15.

Mannai, M., Karâa, W. B. A., & Ghezala, H. H. Ben. (2018). Information extraction approaches: A survey. In D. K. Mishra, A. T. Azar, & A. Joshi (Eds.), *Information and Communication Technology. Advances in Intelligent Systems and Computing* (Vol. 625, pp. 289–297). Springer, Singapore. https://doi.org/10.1007/978-981-10-5508-9_28

Mansour, M. A. (2013). The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus. *International Journal of Humanities and Social Science*, *3*(12), 81–90.

Marie-sainte, S. L., Alalyani, N., Alotaibi, S., Ghouzali, S., & Abunadi, I. (2019). Arabic Natural Language Processing and Machine Learning-Based Systems. *IEEE Access*, *7*, 7011–7020. https://doi.org/10.1109/ACCESS.2018.2890076

Miswar, Suhardi, & Kurniawan, N. B. (2018). A Systematic Literature Review on Survey Data Collection System. *International Conference on Information Technology Systems and Innovation (ICITSI), 22-26 Oct. 2018*, 177–181. https://doi.org/10.1109/ICITSI.2018.8696036

Mohamed, E. H., & Shokry, E. M. (2020). QSST: A Quranic Semantic Search Tool based on word embedding. *Journal of King Saud University - Computer and Information Sciences*, *xx*(xx), xx. https://doi.org/10.1016/j.jksuci.2020.01.004

Mohamed, S., Hussien, M., & Mousa, H. M. (2021). ADPBC: Arabic Dependency Parsing Based Corpora for Information Extraction. *International Journal of Modern Education and Computer Science (IJMECS). Publisher: Modern Education and Computer Science (MECS) Press*, *13*(1), 54–61. https://doi.org/10.5815/ijitcs.2021.01.04

Muhammad, M., Rohaim, M., Hamouda, A., & Abdel-Mageid, S. (2020). A comparison between conditional random field and structured support vector machine for Arabic named entity recognition. *Journal of Computer Science*, *16*(1), 117–125. https://doi.org/10.3844/jcssp.2020.117.125

Nadkarni, P. M., Ohno-machado, L., & Chapman, W. W. (2011). Natural language processing : an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Najeeb, M. M. A. (2020). A novel hadith processing approach based on genetic algorithms. *IEEE Access*, *8*, 20233–20244. https://doi.org/10.1109/ACCESS.2020.2968417

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., & Habash, N. (2020). CAMeL tools: An open source python toolkit for arabic natural language processing. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings. 13-14-15 May 2020*, 7022–7032.

Omar, N., & Al-Tashi, Q. (2018). Arabic nested noun compound extraction based on linguistic features and statistical measures. *GEMA Online Journal of Language Studies. Publisher: Universiti Kebangsaan Malaysia Press*, *18*(2), 93–107. https://doi.org/10.17576/gema-2018-1802-07

Ombabi, A. H., Ouarda, W., & Alimi, A. M. (2020). Deep learning CNN – LSTM framework for Arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining*, *10*(Article number: 53), 1–13. https://doi.org/10.1007/s13278-020-00668-1

Paré, G., & Kitsiou, S. (2016). Methods for Literature Reviews. In F. L. and C. Kuziemsky (Ed.), *Handbook of eHealth Evaluation: An Evidence-based Approach* (pp. 157–179). University of Victoria.

Pare, G., Trudel, M., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management. Elsevier*, *52*, 183–199. https://doi.org/http://dx.doi.org/10.1016/j.im.2014.08.008

Saadi, A., & Belhadef, H. (2020). Deep neural networks for Arabic information extraction. *Smart and Sustainable Built Environment, Emerald Publishing*, *9*(4), 467–482. https://doi.org/10.1108/SASBE-03-2019-0031

Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. (2018). A Survey of Arabic Text Mining. In *Studies in Computational Intelligence* (pp. 417–431). Springer International Publishing. https://doi.org/10.1007/978-3-319-67056-0_20

Sarhan, I., El-Sonbaty, Y., & El-Nasr, M. A. (2016). Arabic Relation Extraction : A Survey. *International Journal of Computer and Information Technology*, *05*(05), 430–437.

Schubert, L. (2019). Computational Linguistics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy, SEP* (Spr2019ed.). Stanford University. https://plato.stanford.edu/archives/spr2019/entries/computational-linguistics/

Shaalan, K., Siddiqui, S., Alkhatib, M., & Monem, A. A. (2018). Challenges in Arabic Natural Language Processing. In N. El Gayar & C. Y. Suen (Eds.), *Computational Linguistics, Speech and Image Processing for Arabic Language* (pp. 59–83, Chapter 3). World Scientific Publishing. https://doi.org/10.1142/9789813229396_0003

Soudani, N., Bounhas, I., & Slimani, Y. (2019). MOSSA: a morpho-semantic knowledge extraction system for Arabic information retrieval. *International Journal of Knowledge and Web Intelligence. Inderscience Publisher*, *6*(2), 106–141. https://doi.org/10.1504/ijkwi.2019.103622

Taghizadeh, N., Faili, H., & Maleki, J. (2018). Cross-Language Learning for Arabic Relation Extraction. *Procedia Computer Science*, *142*, 190–197. https://doi.org/10.1016/j.procs.2018.10.475

Thalji, N., Hanin, N. A., Al-Hakeem, S., Hani, W. B., & Thalji, Z. (2018). A novel rule-based root extraction algorithm for Arabic language. *International Journal of Advanced Computer Science and Applications. Publisher: Science and Information Organization*, *9*(10), 120–128. https://doi.org/10.14569/IJACSA.2018.091015

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, *77*(November 2017), 34–49. https://doi.org/10.1016/j.jbi.2017.11.011

Zakria, G., Farouk, M., Fathy, K., & Makar, M. N. (2019). Relation Extraction from Arabic Wikipedia. *Indian Journal of Science and Technology*, *12*(46), 01–06. https://doi.org/10.17485/ijst/2019/v12i46/147512

Zerrouki, T. (2020). *Towards An Open Platform For Arabic Language Processing*. Degree of Doctor of Science, Thesis, National School of Computer Science (ESI), Algiers.