IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 12 November 2023, accepted 30 November 2023, date of publication 7 December 2023, date of current version 15 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3340520

RESEARCH ARTICLE

Tasaheel: An Arabic Automative Textual Analysis Tool–All in One

HANEN T. HIMDI [1] **AND FATMAH Y. ASSIRI** [2]

[1]Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia
[2]Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia

Corresponding author: Hanen T. Himdi (hthimdi@uj.edu.sa)

ABSTRACT This paper demonstrates Tasaheel, an automative Arabic textual analysis tool. It offers two types of textual analysis utilities: traditional natural language processing tasks; such as stemming, segmentation, normalization, name entity recognition, and part of speech tagging, by integrating open-sourced Arabic natural language processing packages. The second type of utility offers novel corpus analysis methods, including a detailed word use summary of part of speech, emotion, polarity, linguistics, and domain-specific word labeling. Tasaheel is the first Arabic tool that analyses affixes as a type of comprehensive textual analysis, along with options to search and handle data. We anticipate that Tasaheel will be a conducive tool for Arabic textual analysis.

INDEX TERMS NLP, text analysis, tool, development.

I. INTRODUCTION

There's a great old saying: "Say what you mean, mean what you say." Any text may be read in various ways; the potential of several interpretations inside a text is referred to as "polysemy." Unlike conventional hermeneutic methods of text exegesis, the purpose of textual analysis is to explain the range of possible meanings encoded in the text rather than to discover one "true" interpretation. The Arabic language has a small share of work in the textual analysis domain, though it is regarded as one of the most widely spoken languages, with 330 million speakers worldwide.

In an effort to increase the Arabic textual analysis research effort, work over the past 20 years to produce tools to enable Arabic natural language processing (NLP) has been gained. These resources seek to tackle various NLP tasks simultaneously or specifically, such as tokenization, discretization, or sentiment analysis. The programming languages, interface types, data forms and standards, and degree of public accessibility of these resources all differ. Building new tools that combine multiple resources becomes challenging due to these variations. Furthermore, a number of constraints have impeded current work on Arabic natural language processing. The major causes of these challenges are the multifaceted morphology of Arabic, which generates quite a bit of ambiguity, a scarcity of lexicons, and the broad variety of Arabic dialects, which could make them sustainable with the tools available [1]. As a result, a number of Arabic tools have been developed to deal with the aforementioned challenges. These NLP tools, however, are scattered throughout a number of research efforts and are not always simple to access. Farasa [2] and MADAMIRA [3] are well-known Arabic NLP tools that provide a selection of traditional general NLP tasks. However, other tools, like Tashaphyne, only provide a few NLP tasks, such as normalization. These tools are commonly employed in different areas of study involving textual analysis.

With the significant amount of data available on the Internet, where there are numerous text documents on a daily basis, the majority of these records are held in an unstructured manner, contain valuable information, and, when properly evaluated, can shed light on plenty of topics. Therefore, it is becoming increasingly essential to automatically analyze the text in these documents. Based on these demands, we developed a text analysis tool called Tasaheel, which in Arabic means "making something easier." The tool attempts to give a general Arabic toolkit for text analysis that may be applied to different analytical projects. It provides two different utility aspects for text analysis, which are the main contributions in this paper:

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

- The first utility includes fundamental NLP tasks collected from open sources Arabic NLP packages; which are: stemming, segmentation, normalization, name entity recognition, and part of speech tagging.
- The second utility offers first-of-a-kind functionalities of tag summaries for parts of speech, emotion, polarity, linguistics, and domain-specific words.
- The taggers were created in a manner analogous to part-of-speech(POS) tagging, with the goal of ensuring that the wordlists used for tagging were validated and reliable for further use in future Arabic research.
- Furthermore, in order to provide a full textual analysis, Tasaheel focuses on the significance of affixes in Arabic by presenting an affix extraction option.
- Provide data management and search by encoding a word finder and converting summary results to data presented in Excel spreadsheets.

This work is an extension of the Tasaheel tool presented in the thesis [4], where it included an expansion of domain-specific wordlists.

The remainder of this work is structured as follows: The Arabic language morphology and difficulties in the NLP research are detailed in Section II. An overview of the notion of textual analysis is provided in Section III. The development and utilities of Tasaheel are displayed in Section IV. The limitations of Tasaheel are presented in Section V, and the scope of further research is covered in Section VI, which brings the paper to a close.

II. BACKGROUND

Arabic is the official language of 20 Middle Eastern and African countries, including Saudi Arabia, Qatar, Bahrain, Jordan, Egypt, Lebanon, and Morocco. Because it is the language of Islam's holy book, the Quran, the number of Arabic speakers has increased due to the growing number of Muslim converts worldwide and the Islamic faith's tradition of reading the text in the original language [5]. Since the inception of Islam in the seventh century CE, Arabs have inhabited many traditionally non-Arabic-speaking countries, sometimes adopting non-Arabic loanwords into the Arabic language. For example, the word أُسْتَاذ 'teacher' is a loan from Persian. Furthermore, in the modern era, globalization has introduced many terms to the Arabic language, such as the Internet. Though this is an English term, Arabs have transliterated it phonetically and added it to their terminology with the same meaning and pronunciation. Nevertheless, Modern Standard Arabic (MSA) has retained the syntax, vocabulary, and phraseology of classical Arabic.

A. THE ARABIC ROOT-PATTERN SYSTEM

Arabic morphology is unique. Every word in the Arabic language has a three-letter root representing the base meaning of the word. From each of these roots, dozens of words can be formed. Specific patterns are applied to the roots, changing the root's meaning to form a related word. The logic and cohesion of the Arabic language, which is highly systematic, originates from this advanced root-pattern system. In its most basic elements, Arabic is composed of consonant roots that work in tandem with vowel patterns.

TABLE 1. Example of verb وقى 'protect'.

Root English	Arabic	Pronun-ciation	Number of Phonemes	Part of Speech
He protected	وقى	Waqa	3	Past tense Verb
He protects	يقي	Yaqy	3	Present tense verb
Protect	ق	Qi	2 (one for letter ق + vowel)	Imperative verb

More specifically, a root or جذر is a relatively invariable discontinuous bound morpheme represented by two to five phonemes—typically three consonants in a specific order—that has lexical meaning and interlocks with a pattern to form a stem [6]. An example is shown in Table 1.

The pattern, as mentioned above, is a limited and, in many cases, discontinuous morpheme consisting of one or more vowels and slots for root phonemes (radicals). These interlock with a root to form a stem, either alone or combined with one to three derivational affixes, and generally, the affixes have grammatical meaning [6]. Put simply, patterns are the fixed molds of words into which roots can be inserted. Together, the root letters and the patterns in which they are placed form words. Patterns, like suffixes and prefixes, also carry meanings.

Thus, the root-pattern system consists of roots that have a general meaning, with more specific meanings and functions created by the patterns in which the roots are placed. To better understand how roots and patterns work together, one can consider the common root of كتب or 'wrote', which forms the basis of Arabic words related to writing or inscriptions. The root similarities between all three letters are readily apparent, even to the untrained eye. While the root كتب 'wrote' signifies a word or phrase related to 'writing,' it is clear that three new words are formed when the patterns, such as short vowels, are added. Another example of different patterns with affixes added to the same root is shown in Table 2.

TABLE 2. Influence of affixes on the word كتب 'wrote'.

English	Arabic	Pronunciation	Type	Example
Writing	كتب	ktb	Verb	He wrote
Writer	كاتب	katb	Noun	Writer
Book	كتاب	ketab	Noun (singular)	Book
Writers	كتبه	katabah	Noun (plural)	Writers

In short, affixes are clitics added to a word for precision and contextual purposes. Their types depend on their position as prefixes, suffixes, or infixes. Table 3 shows an example of patterns, as affixes, added to the root علم which refers to 'knowledge.' Thus, each generated word shares the meaning of the root 'knowledge.'Moreover, the combination of roots and patterns is highly distinctive and may produce the equivalent of a complete sentence in one word. An example of this design can be seen in the same figure, which shows an Arabic word تعلموها that is the equivalent of an entire three-word sentence in English: 'You learn it.' The prefixes and suffixes that serve as patterns are connected to the root, constructing a logical sentence.

For example, English relies on the relationship between consonants and vowels. Similarly, Arabic relies on the relationship between roots and patterns to form words. Roots also allow Arabic speakers to piece together the meaning of new words based on general concepts. In the examples above, readers could identify the general meaning using the root كتب 'wrote' or علم 'knowledge', while using the patterned consonants and vowels to extract the precise definition and its meaning. With the importance of the Arabic root-pattern system in mind, the difference in content or function types of words in Arabic is now examined.

TABLE 3. Full sentence in a word.

English	Arabic	Root	Prefix	Suffix	Post Suffix
You learn it	تعلّموها	علّم	ت	و	ها

B. CONTENT WORDS
As mentioned earlier, content words have individual meanings, and they can include nouns, verbs, adjectives, and adverbs.

1) NOUNS
Derived from lexical roots, Arabic nouns are formed by placing certain patterns into the root to create different nouns. As in English, Arabic nouns can be common or proper nouns. Compound nouns are formed in Arabic by combining two independent words to form a syntactic unit.

2) ADJECTIVES
Adjectives, in Arabic, are words that describe a noun. Depending on their role, they are divided into two groups: attributive and predicative. Attributive adjectives describe characteristics or an attribute of the noun or pronoun they modify. They are usually positioned before a noun to describe it further. In this case, the adjective must agree with the gender and number of the noun. A predicative adjectives modify or describe the subject of a sentence or clause and are linked to the subject by a linking verb. It provides information about the sentence's subject, thus completing the clause. It acts as a predicate in a nominal sentence and agrees with the noun's gender and number. Arabic adjectives can also have a comparative or superlative degree. Comparative adjectives, compare two nouns, however, superlative adjectives are used to indicate the highest degree of comparison [7].

3) VERBS
As in all languages, verbs in Arabic indicate the action in a sentence. Arabic verbs are composed of a combination of two to five consonants as roots that form the base meaning of the verb. Verbs are categorized, according to their tense, into past, present, and future. There are also imperative and future tense verbs. Though not as commonly used as the other verbs, the latter express actions in the future [7].

4) ADVERBS
Arabic adverbs are mainly derived from nouns or adjectives. Their main function is to modify any part of speech aside from nouns. The adverb can modify verbs, adjectives, other adverbs, and clauses. It also gives extra information about the word in terms of manner, time, and the frequency of performing a specific action.

C. GENDER, PERSON, AND NUMBER
Gender, person, and number are also important components in Arabic morphology. There are three persons—first, second, and third—with the first person having no gender distinction. In the second person, depending on number and gender, there are five forms: masculine singular, feminine singular, dual (two persons), masculine plural, and feminine plural. Finally, in the third person, there are six verbal distinctions and five pronoun distinctions: singular masculine = هو 'he', singular feminine= هي 'she', dual masculine= هما 'they', dual feminine= هما 'they', plural masculine = هم 'they', and plural feminine = هن 'they'. As a result, there are 13 Arabic person categories, whereas English has only seven [8]. Arabic has three numbers: singular, dual, and plural. Thus, there are distinct pronouns for pairs of people or animals. In English, however, any number more than one is treated as a plural. Arabic does not consider quantities to be plural until they are three or more. Patterns, such as affixes (prefixes/suffixes) and vowels, can be attached to a verb or noun to specify gender, person, and number. Table 4 shows an example of affixes attached to a verb.

D. FUNCTION WORDS
Function words are expressed by 'particle' حرف, in the Arabic POS basic structure. Function words do not generally carry meaning by themselves, but are a supportive structure that helps to produce organized and detailed meaning in the text. There are a limited number of particles —less than 100 — in Arabic. Each particle holds a peculiar meaning and functions according to that meaning when added to a word or sentence. Two particles can be combined to express a more definitive meaning for the context; for example, لا سيما which means 'especially,' contains two particles 'la' and 'siyama' and precisely means 'for that' Particle types differ based on their function, such as exception and negation

TABLE 4. Verb affiliation. Note: red indicates the prefix, blue indicates the suffix. Fem.=feminine, Masc.=masculine.

Description	Base Form	Fem. Singular	Masc. Singular	Fem. Dual	Masc. Dual	Fem. Plural	Masc. plural
English	Eat						
Arabic	أكل a'kal	تأكل Ta 'kl	يأكل Ya 'kl	تأكلان Ta ' kulan	يأكلان Ya ' kulan	تأكلن Ta' kulna	يأكلون Ya' kulun
Type	Verb						

particles. Exception particles are used to express an object as separate from a particular group. Usually, these are followed by the expectant, which is a noun. Negation particles are used to negate a statement. Further examples of function words include prepositions, conjunctions, and pronouns.

1) PREPOSITIONS

Though they are limited in number, comprising only 17, Arabic prepositions play a pivotal role in signifying the relationship between one word and another. A preposition may consist of only one letter attached to a noun or a separate word composed of several letters. Each preposition has a linguistic meaning that appears when added before a noun, signifying a location or direction. Prepositions also include derivative prepositions that are a form of a temporal or locational adverb. Some examples include في 'in' and على 'on'.

2) CONJUNCTIONS

Conjunctions are particles that primarily function to connect words or sentences to show a link, such as cause and effect, contradiction, or sequence. There are two types of conjunctions: coordinating and subordinating. Coordinating conjunctions are the type used most in Arabic, as they connect two related words, thoughts, or sentences. Subordinating conjunctions, on the other hand, connect two unequal clauses. When one clause contains a verb, the other clause needs an object. If an object is not present, then the statement becomes unequal. Hence, subordinating conjunctions are used to link the clauses. Conjunctions can be attached to or detached from a word. Since they are function words, each conjunction has a unique meaning and performs a linguistic function [9].

3) PRONOUNS

Pronouns are words used to replace a noun. Like conjunctions, Arabic pronouns can be attached or detached. If attached, they are linked to a word in place of the person/thing and agree with the word's number and gender. For example, the pronoun ي 'ya' is assigned when an imperative verb is directed to a feminine subject. However, the pronoun أ 'aa' is attached when the imperative verb is directed to a masculine subject. Detached pronouns are concrete words used in place of persons and things in a sentence. Similar to attached pronouns, they also agree with the person, number, and gender specifications of the subject and object [9]. Table 5 lists some examples. As pronouns perform a grammatical function when added to a sentence, they are also categorized as particles in Arabic.

4) DETERMINERS

In addition to the function words above-mentioned, Arabic also uses determiners, which are classified as definite and indefinite. The prefix al- is definite and used at the beginning of nouns and adjectives. The indefinite determiner is the diacritic mark $\hat{o}$ attached to the end of case-marking vowels in nouns and adjectives [6]. For example, 'the dog' is expressed as 'al kalb' الكلب, while 'a dog' is 'klbaan' كلباً.

E. AFFIXES AS FUNCTION WORDS

Generally, attached pronouns, prepositions, or conjunctions to a content word are called affixes. Affixes are linguistic elements added to a word to produce an inflected or derived form. When attached to a word, some Arabic affixes convey a grammatical meaning. For example, in English, the prefix 'un-' has the same grammatical role as the function word 'not.' Though not concrete, these affixes are considered, in their role, as function words because they give a functional meaning when attached to a word. Table 6 shows an example of affixes as function words.

F. ARABIC NLP CHALLENGES

The unique morphology of Arabic creates challenges for the Arabic language research community. These challenges have a direct impact on NLP tool processing and, thus, on textual analysis works. Some of these challenges are explained below.

1) ORTHOGRAPHIC VARIATIONS

Some Arabic letters share the same letter shape but have different pronunciations, especially when marks such as single dots or double dots—hamza (ء), or mada (~)—are placed above or below the letter [6]. Thus, NLP tools must distinguish between the letters based on the position of these marks. However, some MSA texts are lax about adding these marks, and the proper marks are sometimes omitted. It is usually up to the reader to determine which word is intended, depending on their familiarity with this practice. For example, the word في meaning 'in' is sometimes written without the two dots beneath it as فى.

TABLE 5. Pronoun types and examples.

Pronoun	Specification	Type	Example
أنتِ anti	Feminine singular	Detached	أنتِ التي نجحت. You are the one who passed.
أنتَ anta	Masculine singular	Detached	أنتَ الذي نجح You are the one who passed.
ي ya	Feminine singular	Attached	ادرسي درسك You study your lesson.
نا na	Feminine and masculine dual	Attached	درسنا الدرس We studied the lesson.

TABLE 6. Affixes as function words: affixes are colored in red.

Prefix/ Suffix	Meaning, Role	Example	English
ل li +verb	purpose, justification	ذهب أحمد ليلعب	Ahmed went to play.
ك ka +Noun	as, similarity	وجهك كالقمر	Your face is as the moon.
س sa +verb	will, future action	سيذهب أحمد إلى المدرسة	Ahmad will go to school.

2) LACK OF CAPITALISATION AND PUNCTUATION

The absence of capitalization and clear punctuation rules in Arabic make pre-processing difficult. During the automatization process, the machine cannot distinguish between one clause and another, as some Arabic sentences may run the length of an entire paragraph without commas, with coordinators linking the statements together and, with the whole section having only one final punctuation mark. Additionally, as proper names in Arabic are not capitalized, their shape is not identifiable. In some cases, a proper noun may be mistaken for a common noun. For example, أيقظتني أحلام could mean 'I was awakened by dreams or 'I was awakened by Ahlam' (a personal name), as أحلام Ahlam means 'dreams' in Arabic and is also a common female name.

3) HOMOGRAPHS

The current habit of readily discarding the written diacritics of words in MSA text creates homographs. As mentioned in [10], diacritics are essential and considered short vowels used to identify the pronunciation of letters. Inevitably, ambiguity arises when diacritics are misplaced or misused, leaving the reader to identify the word according to the overall context and, making it harder for NLP tools to identify the word accurately. As with any language, when there is a misuse of a single diacritic, such as شدة 'shaddah', which doubles the consonant, it can cause confusion in multilingual contexts and will mean failure to identify words correctly. For example, the word مثل 'mathal', when written without a shaddah on the middle letter might imply the meaning, 'similar.' However, when a shaddah is added to the middle letter مثّل,' maththal', it means 'acting.'

4) LACK OF ARABIC LEXICONS

Standard Arabic lexicons include Lisan Al-Arab[1] and Al-Mujam Al-Ghani,[2] which have entries for over 300,000 words. These lexicons are widely used in text analysis projects, such as sentiment, subjectivity, and author analyses, as well as identifying the author's gender [11], [12], [13]. Although these two lexicons are useful, they do not provide easy access to specific lexical categories. As in dictionaries, all the words are arranged in alphabetical order, with each word defined and given its grammatical use, if provided. The researcher needs to search for their desired words and combine similar words that have similar purposes to form a specific lexicon, which could be burdensome. In fact, some researchers have manually compiled specific Arabic lexicons such as Arabic particle lexicons [14], verb lexicons [15], and sentiment lexicons [13]. A recent lexicon is ArDep: An Arabic Lexicon for Detecting Depression [16], which was compiled to recognize the Arabic words and phrases used by people suffering from depression.

On the other hand, researchers have translated available non-Arabic-specific lexicons in other languages such as English and French into Arabic for research use. For example, [17] translated an intensifier lexicon in French to Arabic [18].To enhance the use of lexicons, some authors of sentiment lexicons have assigned each word a score to test a system's ability to predict the sentiment intensity score for a given text. The Multi-Perspective Question Answering (MPQA) subjectivity lexicon, for example, contains 2,718 positive, 4,911 negative, and 570 neutral words. Each word was assigned a score between 0 and 1 indicating the intensity, with 1 indicating the maximum score for a positive sentiment

[1] http://arabiclexicon.hawramani.com/ibn-manzur-lisan-al-arab/
[2] https://nujoomapps.com/product/mojam-al-ghani/

and 0 for a negative one. Another example is the emotion lexicon by [19]. This lexicon included the six basic human emotions, according to [20], as its emotion categories. It has 748 words for expressing anger, 155 for disgust, 425 for fear, 1,156 for joy, 522 for sadness, and 201 for surprise. It gives fine-grained scores to each word, using a scale from 0 to 100 to indicate intensity in the specific emotion category.

5) LACK OF ARABIC NLP TOOLS

Unfortunately, with most research focused on building sentiment lexicons, other domain lexicons have been neglected. The lack of multiple-domain lexicons has caused a lack of NLP tools that support the Arabic language, dampening interest in Arabic research projects. Although there are a number of open source NLP libraries in English that support Arabic language — such as NLTK,[3] TextBlob,[4] Genism,[5] and SpaCy,[6] there are fewer developed with Arabic language specifications such as Farasa [21], Camel [22], and Arabic Linguistic Pipeline [23].

III. TEXTUAL ANALYSIS

Textual analysis is a set of methods used to describe and interpret characteristics of a text by extracting information from textual sources [24]. It requires the researcher to closely analyze the textual content of an item rather than the structure of that item. The concept of textual analysis is largely conducted by alluring NLP capabilities. Natural language processing refers to the techniques that enable the researcher to extract information from textual sources to perform data analysis [25]. For this reason, the field of NLP has been explored by many researchers who aim to automate the extraction process of useful textual items by designing NLP tools for this service [21], [22], [26].

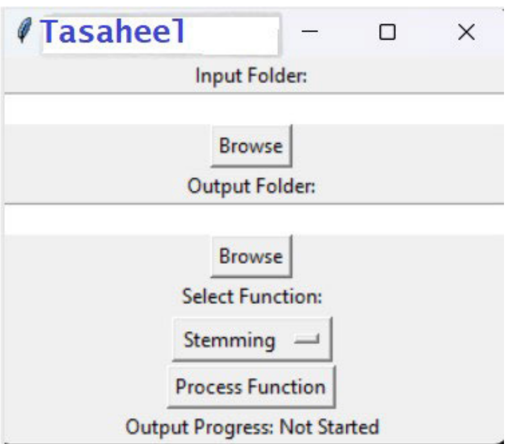

FIGURE 1. **Tasaheel GUI.**

[3]https://www.nltk.org/
[4]https://textblob.readthedocs.io/en/dev/
[5]https://radimrehurek.com/gensim/
[6]https://spacy.io/

Textual analysis dates back to 1969, when [27] investigated its usefulness for extracting rich information and proposed a computer system for performing computational linguistic analysis. More recently, with the rise of various NLP tools and ML classifiers, research on textual analysis has been adopted in accounting [28], stock investments [29], and identifying Arabic conspiracy theories on Twitter [30]. Textual analysis has also been used to minimize the manual efforts needed to analyze qualitative data. A framework by [31] was designed to focus on specific linguistic and artistic elements which is based on textual analysis conceptual theory. Textual analysis is not only used in the above-mentioned research domains, but also in cultural studies, mass communication, media studies, philosophy, and sociology, to name a few. Specifically, it is relevant to these fields because they are related to human behaviors and ways of communicating and coexisting [32]. Since textual analysis provides the writer's perspective by analyzing their written text, several studies have applied textual analysis to the author's writings in search of distinctive insights [33], [34]. With regard to current research, for example, it is more convenient to focus on certain deceptive markers in a text to predict the content's veracity than to attempt to analyze the whole text to try to find nuanced differences between it and a non-fake text. In fact, textual analysis with the aid of NLP and ML has been employed in research to detect fake news [35], [36], [37].

IV. TASAHEEL

A. DEVELOPMENT

Tasaheel was developed using Python programming language, version 3.11.5. It addresses two aspects of textual analysis, fundamental NLP tasks and novel approaches to analysis. Its interface is shown in Figure 1. Tasaheel is available upon request.

B. UTILITIES

Tasaheel was designed to conduct several NLP functions in Arabic. Some of these functions were provided by packages coded in Python that were available online but scattered in several Arabic research platforms. Our aim was to collect and join several NLP packages that supported Arabic, which provided a comprehensive research utilities tool. Some tools that provide such packages are Tashaphyne and the Information Science Research Institute's stemmer (ISRI). On the other hand, there are packages that provide multiple tasks in the form of a unified toolkit, such as Farasa and Stanford CoreNLP. A description of the packages integrated into Tasasheel are described below:

- **Tashaphyne**: An Arabic light stemmer and segmental tool. It provides light stemming, such as removing prefixes/suffixes, and generates segmentation. It relies on using its own built-in customized prefix and suffixes list, which offers more precise stemming. Besides stemming and segmentation, it offers normalization and root extraction.

TABLE 7. Examples of the existing NLP tasks output.

Original	كان الجو جميلا في الربيع . حيث أزهرت شجرة التوت
Segmentation (Farasa)	كان ال+جو جميل+ا في ال+ ربيع . حيث أزهر+ت شجر+ة ال+توت
Original	يذاع برنامج خاص بعنوان موسم الحج ١٥ دقيقة باللغة الإنجليزية لتعريف المستمعين بمناسك الحج. كما تذيع الشبكة مع بداية شهر ذو الحجة آيات تتلى يوميا ١٠ دقيقة
Stemming (ISRI)	أذاع برنامج خاص عنوان موسم حج ١٥ دقيقة لغة إنجليزية تعريف مستمع منسك حج . كما أذاع شبكة مع بداية شهر ذو حجة آية تلا يومي ١٠ دقيقة
Normalize (Tashaphyne)	يذاع برنامج خاص بعنوان موسم الحج ١٥ دقيقه باللغه الانجليزيه لتعريف المستمعين بمناسك الحج. كما تذيع الشبكه مع بدايه شهر ذو الحجه ايات تتلي يوميا ١٠ دقيقه
Original	احب هذا! مصنوع بشكل جيد ومتين ومريح للغاية
POS Tagging (Farasa)	+ ب /PREP + مصنوع/NOUN-MS + !/PUNC +هذا/PRON +أحب/v +متين/ADJ-MS +و/CONJ +جيد/ADJ-MS +شكل /NOUN-MS NOUN +ال+غاي+ة /DET +ل/PREP + مريح/ADJ-MS +و /CONJ

- **ISRI** [38]: It is a light-stemming approach for the Arabic language that does not necessitate the use of any pattern or word dictionaries. The suggested stemmer is regarded as innovative because of its implementation of intelligent rules and the removal of affixes in the form of prefixes and suffixes from the word. Additionally, it endeavors to reduce the level of ambiguity associated with the primary letters to the greatest extent feasible.
- **Stanford CoreNLP** [39]: A multi-language NLP tool that can be used for many languages. For Arabic, it provides parsing, tokenization, sentence splitting, name entity recognition, and POS tagging. It offers utilities through a Python package.
- **Farasa**: An Arabic-specific tool that provides NLP utilities through a collection of Java libraries. The utilities include discretization, segmentation, POS tagging, NER, and parsing.

1) PART 1: TRADITIONAL NLP TASKS

This part describes fundamental NLP tasks that were integrated from their available packages online. Examples are shown in table 7 and they are as following:

- **Stemming**: Stemming refers to the procedure of transforming several inflected word forms into a standardized canonical form [41]. It offers the following packages: ISRI, Farasa, and Tashaphyne.
- **Segmentation**: The process of splitting apart text into comprehensible units such as words, phrases, or topics [42]. Due to Arabic's unique morphology, it is necessary to segment text into morphemes to decrease the ambiguity in Arabic text that is created from the attached affixes. Here, the libraries available are Tashaphyne and Farasa.
- **Normalisation**: Normalization reduces word ambiguity and removes unnecessary randomness associated with the text to unify the word variation [43]. This option is provided as a full normalization utility, offered by Tshaphyne, or a single normalization method, we compiled. The utilities offered are: removing numbers, non-Arabic letters, characters, stop words (the user provides the list of stop words), or diacritic marks.
- **Name Entity Recognition (NER)**: NER capabilities were provided by Farasa and Stanford CoreNLP libraries [44]. It entails identifying significant information in the text and categorizing it into a set of predetermined categories.
- **POS Tagging**: It is the process of assigning a word in a text to a certain part of speech based on both the definition it provides and the context in which it occurs [45]. To assign a POS to each word in a sentence, POS taggers were used. Further, the user is given two POS tagger types: Farasa or Stanford CoreNLP. The POS tags presented in Farasa and Stanford CoreNLP are shown in Table 8 and Table 9, respectively.

The input folder is provided by the user (which may include an unlimited number of text files), and the output files are (i)Tagged text files as shown in Figure 2. (ii) A summary file for each tagged document as shown in Figure 3 which displays the number of occurrences of each tag and its lexical density was calculated as follows: (1), as shown at the bottom of the next page.

2) PART 2: NOVEL ANALYSIS APPROACHES

In this section, we detail the compilation of novel analysis approaches in support of a thorough textual analysis. We develop taggers that target analyzing text from different perspectives. In order to develop the taggers, word resources, and matching approaches were first put in place.

3) CREATION OF WORDLISTS

Not only do languages with unique morphologies like Arabic lack the relevant corpora available for a high-resource language such as English, but they also lack the basic lexical resources. While this lack of readily available lexical resources created a challenge for this study, it also produced

TABLE 8. Farasa POS tags.

Farasa	
Content and Function words tags	
NOUN: noun	V: verb
DET: determiner	CONJ: conjunction
PREP:preposition	PRON: pronoun
ADJ: adjectives	ADV: adverb
ABBREV: abbreviation	FOREIGN: foreign character
PUNC: punctuation	PART: particle
Affixes tags	
FS: female/singular	MS: male/singular
FD: female/dual	MD: male/dual
FP: female/plural	MP: male/plural

TABLE 9. Stanford CoreNLP POS tags [40].

Stanford Arabic POS	Tag Set	Abbreviation
Noun	Noun, singular or mass with the determiner "Al" (ال)	DTNN
	Proper noun, singular with the determiner "Al" (ال)	DTNNP
	Proper noun, plural with the determiner "Al" (ال)	DTNNPS
	Noun, plural or mass with the determiner "Al" (ال)	DTNNS
	Noun, singular or mass	NN
	Proper noun, singular	NNP
	Proper noun, plural or mass	NNPS
	Noun, plural	NNS
	Noun	NOUN
Verb	Verb, base form	VB
	Verb, past tense	VBD
	Verb gerund or present participle	VBG
	Verb, past participle	VBN
	Verb, non-3rd person singular present	VBP
	Verb, past participle	VN
Adjective	Adjective with the determiner "Al" (ال)	DTJJ
	Adjective, comparative with the determiner "Al" (ال)	DTJJR
	Adjective	JJ
	Adjective, comparative	JJR
	Adj	ADJ
Adverb	particle	RB
	Wh-adverb	WRB
Conjunction	Coordinating conjunction	CC
	Preposition or subordinating conjunction	IN
Preposition	Preposition or subordinating conjunction	IN
Pronoun	Personal pronoun	PRP
	Possessive pronoun	PRPS

opportunities. Since these lexical resources had to be created from the ground up, they could be crafted to meet the specifications and goals of this research. As previously stated, most Arabic lexicons were either created and annotated manually [13] or translated from non-Arabic lexicons [17], [46]. The first phase of lexicon creation is quite intense, and the second phase involves the somewhat tedious work of translating words and removing any duplicates that might be produced by translation. Henceforth, the term 'wordlist' is used for convenience and to distinguish it from lexicons, which may be associated with scores. In other words, all the words have the same purpose within the feature category.

4) EMOTION AND POLARITY WORDLISTS

The emotion and polarity wordlists were compiled from the words included in previous lexicons. Specifically, to create

$$\text{Lexical Density (L)} = \frac{\text{(Total number of occurrences of each feature in a class)} \times 100}{\text{Total number of words in the whole class}} \tag{1}$$

و/VBD دعا/VBD أمير/NNP منطقة/NN مكة/NNP المكرمة/DTJJ من/IN سبق/VBD لهم/NN الحج/DTNN
بفتح/VBD المجال/DTNN لمن/NNP لم/RP يؤدوا/VBP الفريضة/DTNN من/IN قبل/NN
بدوره/NNP قال/VBD وكيل/NN إمارة/NN منطقة/NN مكة/NNP المكرمة/DTJJ رئيس/NN اللجنة/
DTNN الإشرافية/DTJJ للحملة/NN الوطنيه/DTNN الإعلامية/DTJJ "الحج/NNP عباده/NNP
وسلوك/NNP حضاري"/NNP

FIGURE 2. Sample of stanford CoreNLP tagged text.

```
CONJ: 12411 Time
lexical density: 3.7918 %
_____________________________
V: 12540 Time
lexical density: 3.8313 %
_____________________________
NOUN+NSUFF-FS: 8440 Time
lexical density: 2.5786 %
_____________________________
DET+NOUN+NSUFF-FD: 990 Time
lexical density: 0.3025 %
_____________________________
NOUN-MS: 43001 Time
lexical density: 13.1378 %
_____________________________
NUM-MP: 17031 Time
lexical density: 5.2034 %
```

FIGURE 3. Summary of POS tags using Farasa tagger.

the emotion wordlist, we extracted the words from Bing Liu's English emotion lexicon [47]. Fortunately, the words had previously been translated into Arabic in [46]. The emotion wordlist categories contained the following six emotions:

- 748 words denoting anger
- 155 words denoting disgust
- 425 words denoting fear
- 1,156 words denoting joy
- 522 words denoting sadness
- 201 words denoting surprise

Similarly, words from the Arabic sentiment lexicon created by [13] were extracted to form the polarity wordlists. The lexicon included positive and negative words. All the words included in the positive lexicon were grouped to form the polarity wordlist, comprised of 2,006 words. On the other hand, all the words included in the negative lexicon were grouped to form the negative wordlist, composed of 4,783 negative words.

It is important to note that certain words in the polarity wordlists are unavoidably repeated in the emotion wordlist. This is because emotions involve a broader and larger analysis than sentiment to cover the specific details of the desires, goals, and intentions linked to a person's facial expressions. Examples of the polarity and emotion wordlists are provided in Table 10.

TABLE 10. Emotion and polarity wordlists.

Content Feature	Example	Translation
Anger	امتعاض, نقمة, غضب, سخط, حنق, غيظ	anger, exasperation, indignation, resentment
Sadness	يدمع, دمع, تدمع, دموع, الدموع, بكاء, هم	worry, crying, tears, tearing
Fear	لعين, مفزع, مرعب, مروع, مخيف, رهيب	terrible, scary, horrific, horrific, terrifying, damned
Joy	لذيذ, كريم, صالح, طيب, فائدة	good, generous, delicious, beneficial
Surprise	حيرة, تحير, مذهول, مندهش, حيران	surprise, amazement, confusion
Disgust	نفور, تقزز, تنافر, نفور, مقت, اشمئزاز	disgust, repulsion, loathing
Positive	فاخر, مجاني, حكيم, راقي	wise, free, luxurious, classy
Negative	بخيل, لئيم, وسخ, ثورة	dirty, stingy, revolution, mean

In total, we have organized six emotion wordlists - anger, fear, sad, surprise, disgust, and joy - that are part of the emotion wordlists category, and two polarity wordlists - positive and negative - that are part of the polarity wordlists category.

5) LINGUISTIC WORDLISTS

Inspired by the previous work of [17] and [46], we heuristically created a wordlist for each linguistic category to be further embedded into Tasaheel. We followed two methods to organize the words to form the linguistic wordlist. First, we formed the intensifier and hedges wordlists by translating the English lexicons available for that category. Intensifiers were translated from the English intensifier lexicon [48], and hedges were translated from the English hedge lexicon [49]. The words were translated using Google Translate, and duplicate words produced by the translation were removed. Second, as not all the linguistic categories were available in other languages, we created further wordlists by referring to a range of reliable, well-known Arabic lexical resources. The latter was beneficial, as these resources provided words denoting the linguistic categories needed in our work. We organized the words for each linguistic category, relying mainly on the Arabic lexical resources, as shown in Table 11.

6) DOMAIN-SPECIFIC WORDLISTS

Linguistic Inquiry and Word Count (LIWC) is a text-analysis program built by James Pennebaker and his collaborators [50]. A dictionary that categorizes terms serves as the foundation of LIWC. This dictionary applies to English, with some efforts to translate it to other languages such as Deutch [51]. Researchers who wish to use LIWC on non-English texts have traditionally relied on translations of the dictionary

TABLE 11. Lexicon resources.

Lexicon Name	Author	Publisher Country	First Publishing Date
لسان العرب, Lisan Al Arab	ابن منظور, Ibn Manthur	Tunisia	1290
المعجم الوسيط, Al-Mu'jam Al-Waseet	مجمع اللغه العربيه بالقاهره, Arabic Language Association in Cairo	Egypt	1960
المعجم الغني, Al-Mujam Al-Ghani	عبد الغني أبو العزم, Abdul Ghani Abu Al-Azem	Morocco	2016

into the language of the texts [52]. We chose this platform as it provides various lexicons that were uniquely compiled in domain-specific categories. We make use of these lexicons to organize the Domain-Specific(DS) wordlists. To achieve that, we follow the translation approach conducted by [53]. We translated the words included in 9 LIWC lexicons using the Python Google Translate API [54]. It employs Google's neural machine translation technique to instantaneously translate texts. The words translated were extracted from the following lexicons:

- 53 words denoting social
- 23 words denoting polite
- 29 words denoting family
- 17 words denoting friend
- 43 words denoting culture
- 57 words denoting tech
- 24 words denoting home
- 42 words denoting health
- 39 words denoting mental

For simplicity, we use the term "wordlists" to associate the translated lexicons with our tool's development.

7) WORDLIST REVISION

Assuming that the emotion and polarity wordlists were credible for use because they had already been used in Arabic research studies [11], [13], [17], [46], [55], we assigned three female Arabic linguistics scholars from Umm Al-Qura University in Makkah, Saudi Arabia, to revise the linguistic and Arabic translated DS wordlists. They classified each word as 'approved' or 'not approved' based on its fit with its featured category. Aggregation was based on voting; at least two scholars had to agree on the word's compatibility with its featured category. We followed Fleiss's Kappa metric to measure the inter-annotator agreement, reaching 0.72 [56].

As a result, 10 linguistic and 9 DS wordlist categories were organized, containing concordant words within the functionality of each category. Table 12 describes the linguistic wordlist, along with the number of concordant words it contains, with examples of the words translated into English. The DS wordlists can be found in the LIWC dictionary repository.[7]

[7]https://www.liwc.app/dictionaries

عند بداية الشوط الثاني , تمكن الفريق من الاستحواذ علي الكرة , لكنهم [opposite], مع ذلك
, لم[negators] يكونوا فعالين كما ينبغي امام المرمي , كما ان المدافعين الانجليز كانوا يقظين[positive]
ومتموضعين بشكل جيد[joy] .

FIGURE 4. EPL tagging.

```
time: 2899 Time
lexical density : 0.3237 %
_______________________________
place: 8565 Time
lexical density : 0.9563 %
_______________________________
positive: 15983 Time
lexical density : 1.7846 %
_______________________________
joy: 8865 Time
lexical density : 0.9898 %
_______________________________
negators: 6501 Time
lexical density : 0.7259 %
_______________________________
negative: 10695 Time
lexical density : 1.1941 %
```

FIGURE 5. Summary of EPL tagging.

- **Emotion, Polarity, Linguistic (EPL), and DS Tagger** Here, the emotion, polarity, linguistic, and DS wordlists were integrated. We developed a tagger that tags words that fit these wordlists. The tool provides two separate tagger options, either EPL or DS word tagging. Moreover, the output of any of these options is two output files: (i) each input file displaying the matching words tagged with its category, as shown in Figure 4. (ii) A summary file for each tag matched in all the input files, as shown in Figure 5. This file also displays the number of occurrences of each tag and its lexical density. We would like to note that the words in wordlist matching are based on the approach that uses the exact string matching method recommended by [57], where each word in the list is compared to each word in the

TABLE 12. Linguistic wordlists with the number of concordant words each contains.

Lexical Wordlist	Meaning	Word Example (Translated into English)
Assurance [7]	transitions used to indicate assurance	نف ,عين ,أن A'an, a'in, nafs for sure, surely, certainly
Negations [7]	used to dispute the truth of a statement	ل ,لن ,لا la, lan, lam no, not, never
Intensifiers [14]	to strengthen the meaning of a word	جميع ,تماما ,جدا jidan, tamaman, jame very, too, not at all
Hedges [7]	to soften / express hesitation / unassurance	احتمال ,يجب ,من الممكن min almomkin, yaji'b, ehtimal maybe/ should/ could/ may
Justification [9]	to show cause / justification	من أجل ,لذلك ,بسبب bisabb, lithalik, min ajel because/ to/ for that
Temporal [8]	to show time	غدا ,البارحه albariha, ghadan yesterday, tomorrow
Spatial [10]	to show space	عند ,فوق ,تحت taht, foug, e'nd under, over
Illustration [6]	used to portray	مثل ,مثال mithal, mathal for example
Exceptions [6]	used to indicate omission	سوى ,عدى ,إلا e'la, a'da, siwa except
Opposition [4]	to indicate adversity	إنما ,لكن lakn, e' nama but /although

text files and a match is displayed in the output files. End users are free to handle and update the words in the customized wordlists separately.

- **Affix Analyser:** [58] stated that affixes play a significant role in changing words' meaning and, thus, the grammatical function. As extracting the affixes might be helpful for NLP projects, especially those focusing on textual analysis, and this option was added to the tool. To the best of our knowledge, no work has previously been done to extract affixes in Arabic, although some related work was performed to remove affixes to obtain the roots of the words, and some Arabic NLP tools have employed specific tools for this purpose [59].
  In this context, by identifying the word POS in a sentence, one may create a query based on a syntactic rule that may help identify any affixes attached to it. Making use of the affixes produced may provide a more precise analysis of the text. Some affixes are prepositions, pronouns, or conjunctions that perform similar grammatical roles to detached prepositions, pronouns, or conjunctions. To extract affixes, tagging files under Farasa were used, as it provides specific Arabic tags with consideration to Arabic affixes. The user is given two options to extract affixes, i.e., to extract either prefixes or suffixes.
  Moreover, a unique approach was followed that is similar to the information retrieval method when searching for a query to extract affixes. The tool performs string matching from right to left when searching for a prefix as the desired affix, as shown in algorithm 1. However, the string matching is approached from left to right when looking for a suffix as the targeted affix, Algorithm 2.
  For example, some affixes tagged as prepositions may have the same grammatical role as some of the linguistic categories listed above. In the justification category, nine words were constants; however, one was an affix that held the same grammatical role as the other constant words. In particular, the affix ل 'li' means 'for that cause' when attached to a present verb, which produces a justification. To search for this affix, we search for this proposition in the affix analyzer and input the Farasa-tagged files. In Farasa-tagged text files, the words and their connected affixes are assigned to detailed POS tags showing the affix type. ل 'li' is tagged as a preposition in

Algorithm 1

$W = w1, w2, w3.$ ▷ words
$T = t1, t2, t3.$ ▷ tags
$F(\text{TEXT\_FILE}) = w1t1, w2t2, w3t3 \ldots.$ ▷ word_tag
input (affix + BASE_TAG)
search (right to left string matching)
for $i = 0$ **to** EOF **by** 1 **do**
  if $t_i = \text{BASE\_TAG}\&\&W_i$ (first_letter = affix) **then**
    return W_i
  end if
end for

Algorithm 2

$W = w1, w2, w3.$ ▷ words
$T = t1, t2, t3.$ ▷ tags
$F(\text{TEXT\_FILE}) = w1t1, w2t2, w3t3 \ldots.$ ▷ word_tag
input (affix + BASE_TAG)
search (left to right string matching)
for $i = 0$**to**EOF**by**1 **do**
  if $t_i = \text{BASE\_TAG}\&\&W_i$ (first_letter = affix) **then**
    return W_i
  end if
end for

this case. Given this, the following query 6 was formed. Figure 7 shows the output file with the result of this query for a tested dataset. Where the file indicates that the matching query of this affix is present one time in a file named "test_3" and one time in another file named "test_9". Figure 8 demonstrates the place of query match highlighted in the "test_3" file.

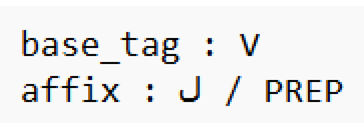

FIGURE 6. Query formation.

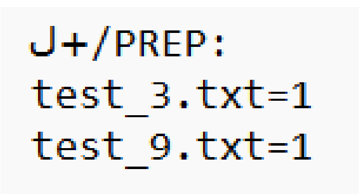

FIGURE 7. Query output sample.

- **Word Finder:** There might be an interest in investigating a certain word used in a group of text files for further implementations. A word-matching function that gives the option for the user to input a word is included. When the match with this word is found in the files, an output file of the word's matching file and the number of occurrences in that file is generated.
- **Excel Output:** For the convenience of researchers to keep track of the generated data, this option is provided in Tasaheel to automatically upload all the generated results from any summary file produced in the previous options into an Excel worksheet. This Excel worksheet contains the tags as columns and file numbers as rows, with the number of occurrences of each tag in each file, as shown in Figure 9.

FIGURE 8. Query place result.

V. LIMITATIONS

Tasaheel adheres to several limitations. First of all, it is unable to handle homograph issues. When Arabic text needs to have its diacritical marks removed, homographs present a problem. Second, the program only handles documents in text format. Third, because our tool matches terms based on exact word matching in EPL and DS tagging, the variants of words may cause a mismatch. To solve this issue, segmenting the input files before EPL and DS tagging is advised. Lastly, EPL and DS tagging is performed without taking into account any word-related metrics, such as polarity ratings. The presence of specific words that may convey meaning is what our tool is aiming for. We were primarily interested in identifying the emotions conveyed by the words rather than their specific polarity. We wanted to understand the overall emotional tone of the text rather than determine if it was positive or negative. Instead of focusing on sentiment analysis, we aimed to capture the general emotional polarity of the text.

VI. FUTURE WORK

We plan to integrate more Arabic NLP suites into our tool. Moreover, we plan to provide textual analysis on several text formats, such as DOCX and PDF. We also intend to overcome the homographs issue by setting syntactic algorithms that may identify words based on their POS tag. Finally, future work will also focus on testing and evaluating the tool on more large Arabic datasets.

VII. CONCLUSION

As displayed in this paper, we present an automated Arabic textual analysis tool that provides several NLP tasks and introduces novel utilities to support textual analysis. Traditional NLP tasks include stemming, segmentation, normalization, NER, and POS tagging with a variety of integrated packages. On the other hand, innovative and novel NLP tasks include a comprehensive detailed POS tags summary; emotion, polarity, linguistics, and domain-specific

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	File	S	PUNC	V	NOUN-MS	DET	ADJ-MS	NOUN	NSUFF-FS	ADJ	NSUFF-FD	PREP	NUM-MP	NOUN-M	NSUFF-FP	CONJ	PRON	E	PART	NSUFF-
2																				
3	ksa11.txt	9	38	27	148	131	10	64	51	27	24	58	3	15	14	45	7	9	14	
4	ksa1010.txt	1	5	1	11	17		4	6	6	1	6	4	1	2	4	3	1	1	
5	ksa100100.txt	7	17	26	157	101	10	62	43	30	15	56	7	10	27	29	3	7	10	
6	ksa101101.txt	3	16	8	62	102	9	43	43	30	5	38	2	16	20	31	10	3	5	
7	ksa102102.txt	50	25	80	215	147	25	59	44	25	12	92	17	27	24	50	32	50	77	
8	ksa103103.txt	2	17	5	67	32	5	19	17	8	5	29	4	6	2	13	4	2	5	
9	ksa104104.txt	6	24	18	103	105	14	47	29	26	19	37	4	4	18	41	10	6	12	
10	ksa105105.txt	4	17	14	51	41	4	35	25	16	13	12	4	3	13	17	3	4	2	
11	ksa106106.txt	1	5	3	30	36	10	14	8	5	3	15	1		6	2	5	1	2	
12	ksa107107.txt	2	14	3	39	39	2	20	16	8	6	14		5	5	10		2		
13	ksa108108.txt	5	14	20	80	97	12	60	52	33	14	47	7	3	22	17	7	5	8	
14	ksa109109.txt	3	12	3	41	33	5	13	14	6	5	11	1	1		6	1	3		
15	ksa1111.txt	5	29	32	102	110	11	66	53	26	10	64		9	27	40	19	5	25	
16	ksa110110.txt	3	10	5	64	59	13	18	14	11	8	16	4	3	6	15	2	3	2	
17	ksa111111.txt	3	15	24	114	117	24	48	44	24	8	62	1	13	17	40	17	3	14	
18	ksa112112.txt	6	35	8	86	101	10	58	46	26	16	32		13	20	26	1	6		
19	ksa113113.txt	2	19	11	48	68	10	46	45	23	7	30	7	1	14	13	7	2	4	
20	ksa114114.txt	3	31	14	101	115	7	41	41	27	10	53	1	15	12	40	11	3	7	
21	ksa115115.txt	1	12	3	55	49	14	11	10	5	2	14	1	2	3	10	2	1	2	

FIGURE 9. Excel output.

word tagging with detailed summaries. This utility is the first offered for the Arabic language. Tasaheel is the first Arabic tool that provides affixes' extractors as a form of an in-depth textual analysis. In order to support users with technical functionality, our tool further provides conversion of text files into Excel data. Finally, the tool at hand involves locating particular words within specified folders. Tasaheel can be provided for researchers upon request.

DECLARATION OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

(Hanen T. Himdi and Fatmah Y. Assiri contributed equally to this work.)

REFERENCES

[1] K. M. O. Nahar, A. F. Al Eroud, M. Barahoush, and A. M. Al-Akhras, "SAP: Standard Arabic profiling toolset for textual analysis," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 2, pp. 222–229, Apr. 2019.

[2] K. Darwish and H. Mubarak, "Farasa: A new fast and accurate Arabic word segmenter," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 1070–1074. [Online]. Available: https://aclanthology.org/L16-1170

[3] A. Pasha, M. Al-Badrashiny, M. T. Diab, A. E. Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth, "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2014, pp. 1094–1101. [Online]. Available: https://api.semanticscholar.org/CorpusID:10887722

[4] H. T. Himdi, "Classification of Arabic real and fake news based on Arabic textual analysis," Ph.D. thesis, Univ. Strathclyde, 2022. [Online]. Available: https://stax.strath.ac.uk/concern/theses/nv9353359

[5] C. Holes, *Modern Arabic: Structures, Functions, and Varieties*. Washington, DC, USA: Georgetown Univ. Press, 2004.

[6] K. C. Ryding, *A reference grammar of modern standard Arabic*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[7] F. Abu-Chacra, *Arabic: An Essential Grammar*, 2nd ed. Evanston, IL, USA: Routledge, Jun. 2007. [Online]. Available: https://www.taylorfrancis.com/books/9781134119189

[8] M. A.-H.-A. Shamsan and A.-M. Attayib, "Inflectional morphology in Arabic and English: A contrastive study," *Int. J. English Linguistics*, vol. 5, no. 2, p. 139, Mar. 2015.

[9] A. Bouchentouf, *Arabic for Dummies*. Hoboken, NJ, USA: Wiley, 2013.

[10] S. Elkateb, W. Black, P. Vossen, D. Farwell, H. Rodríguez, A. Pease, M. Alkhalifa, and C. Fellbaum, "Arabic WordNet and the challenges of Arabic," in *Proc. Int. Conf. Challenge Arabic NLP/MT*, London, U.K., Oct. 2006, pp. 15–24. [Online]. Available: https://aclanthology.org/2006.bcs-1.2

[11] H. M. Al-Barhamtoshy, H. T. 2, M. M. Khamis, and T. F. Himdi, "Semantic and sentiment analysis for Arabic texts using intelligent model," *Biosci. Biotechnol. Res. Commun.*, vol. 12, no. 2, pp. 266–274, Jun. 2019.

[12] K. Alsmearat, M. Al-Ayyoub, R. Al-Shalabi, and G. Kanaan, "Author gender identification from Arabic text," *J. Inf. Secur. Appl.*, vol. 35, pp. 85–95, Aug. 2017.

[13] S. Mohammad, M. Salameh, and S. Kiritchenko, "Sentiment lexicons for Arabic social media," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, May 2016, pp. 33–37.

[14] D. Namly, K. Bouzoubaa, Y. Tahir, and H. Khamar, "Development of Arabic particles lexicon using the LMF framework," in *Proc. Colloque pour les Etudiants Chercheurs en Traitement Automatique du Langage Naturel et ses applications (CEC-TAL)*, Sousse, Tunisia, 2015, pp. 1–9.

[15] N. Loukil, K. Haddar, and A. B. Hamadou, "A syntactic lexicon for Arabic verbs," in *Proc. 7th Int. Conf. Lang. Resour. Eval. (LREC)*, 2010, pp. 269–277.

[16] N. S. Alghamdi, H. A. H. Mahmoud, A. Abraham, S. A. Alanazi, and L. García-Hernández, "Predicting depression symptoms in an Arabic psychological forum," *IEEE Access*, vol. 8, pp. 57317–57334, 2020.

[17] J. Karoui, F. B. Zitoune, and V. Moriceau, "SOUKHRIA: Towards an irony detection system for Arabic in social media," *Proc. Comput. Sci.*, vol. 117, pp. 161–168, Jan. 2017.

[18] J. Karoui, B. Farah, V. Moriceau, N. Aussenac-Gilles, and L. Hadrich-Belguith, "Towards a contextual pragmatic model to detect irony in tweets," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, vol. 2, 2015, pp. 644–650.

[19] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proc. ACM Symp. Appl. Comput.*, Mar. 2008, pp. 1556–1560.

[20] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, vol. 98. Hoboken, NJ, USA: Wiley, 1999, pp. 45–60.

[21] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for Arabic," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*, 2016, pp. 11–16.

[22] O. Obeid, N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, G. Inoue, F. Eryani, A. Erdmann, and N. Habash, "Camel tools: An open source Python toolkit for Arabic natural language processing," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 7022–7032.

[23] A. A. Freihat, G. Bella, H. Mubarak, and F. Giunchiglia, "A single-model approach for Arabic segmentation, POS tagging, and named entity recognition," in *Proc. 2nd Int. Conf. Natural Lang. Speech Process. (ICNLSP)*, Apr. 2018, pp. 1–8.

[24] T. Loughran and B. Mcdonald, "Textual analysis in accounting and finance: A survey," *J. Accounting Res.*, vol. 54, no. 4, pp. 1187–1230, Sep. 2016.

[25] S. Pandey and S. K. Pandey, "Applying natural language processing capabilities in computerized textual analysis to measure organizational culture," *Organizational Res. Methods*, vol. 22, no. 3, pp. 765–797, Jul. 2019.

[26] G. R. Weir, "Corpus profiling with the posit tools," in *Proc. 5th Corpus Linguistics Conf.*, 2009.

[27] D. E. Walker, "Computational linguistic techniques in an on-line system for textual analysis," in *Proc. Int. Conf. Comput. Linguistics COLING*, 1969, pp. 1–10.

[28] J. L. Gandía and D. Huguet, "Textual analysis and sentiment analysis in accounting: Análisis textual y del sentimiento en contabilidad," *Revista Contabilidad-Spanish Accounting Rev.*, vol. 24, no. 2, pp. 168–183, 2021.

[29] Z. McGurk, A. Nowak, and J. C. Hall, "Stock returns and investor sentiment: Textual analysis and social media," *J. Econ. Finance*, vol. 44, no. 3, pp. 458–485, Jul. 2020.

[30] A. Al-Hashedi, B. Al-Fuhaidi, A. M. Mohsen, Y. Ali, H. A. G. Al-Kaf, W. Al-Sorori, and N. Maqtary, "Ensemble classifiers for Arabic sentiment analysis of social network (Twitter data) towards COVID-19-related conspiracy theories," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–10, Jan. 2022.

[31] N. Brown and J. Collins, "Systematic visuo-textual analysis: A framework for analysing visual and textual data," *Qualitative Rep.*, vol. 26, no. 4, pp. 1275–1290, Apr. 2021.

[32] A. McKee, *Textual Analysis: A Beginner's Guide*. Newbury Park, CA, USA: SAGE, 2003.

[33] C. L. M. Jeronimo, L. B. Marinho, C. E. C. Campelo, A. Veloso, and A. S. da Costa Melo, "Fake news classification based on subjective language," in *Proc. 21st Int. Conf. Inf. Integr. Web-Based Appl. Services*, Dec. 2019, pp. 15–24.

[34] F. J. F.-B. Penuela, "Deception detection in Arabic tweets and news," in *Proc. FIRE*, 2019, pp. 122–126.

[35] G. Weir, K. Owoeye, A. Oberacker, and H. Alshahrani, "Cloud-based textual analysis as a basis for document classification," in *Proc. Int. Conf. High Perform. Comput. Simul. (HPCS)*, Jul. 2018, pp. 672–676.

[36] B. Cartwright, G. R. Weir, and R. Frank, "Cyberterrorism in the cloud," in *Security, Privacy, and Digital Forensics in the Cloud*. Hoboken, NJ, USA: Wiley, 2019, p. 217.

[37] Z. Khanam, B. Alwasel, H. Sirafi, and M. Rashid, "Fake news detection using machine learning approaches," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1099, no. 1, 2021, Art. no. 012040.

[38] D. H. Abd, W. Khan, K. A. Thamer, and A. J. Hussain, "Arabic light stemmer based on ISRI stemmer," in *Proc. Int. Conf. Intell. Comput.* Cham, Switzerland: Springer, 2021, pp. 32–45.

[39] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Ann. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, Jun. 2014, pp. 55–60. [Online]. Available: https://aclanthology.org/P14-5010

[40] H. Rabiee, *Arabic Language Analysis Toolkit* (Final Year Project–University of Leeds School of Computing Studies, 2010/2011). Univ. of Leeds, School of Computing Studies, 2011. [Online]. Available: https://books.google.com.sa/books?id=lE2YMwEACAAJ

[41] M. Mustafa, A. S. Eldeen, S. Bani-Ahmad, and A. O. Elfaki, "A comparative survey on Arabic stemming: Approaches and challenges," *Intell. Inf. Manage.*, vol. 9, no. 2, pp. 39–67, 2017.

[42] I. Pak and P. L. Teh, "Text segmentation techniques: A critical review," in *Innovative Computing, Optimization and Its Applications: Modelling and Simulations*. Cham, Switzerland: Springer, 2018, pp. 167–181.

[43] S. G. K. Patro and K. K. Sahu, "Normalization: A preprocessing stage," 2015, *arXiv:1503.06462*.

[44] S. Naseer, M. M. Ghafoor, S. bin Khalid Alvi, A. Kiran, S. U. Rahmand, G. Murtazae, and G. Murtaza, "Named entity recognition (NER) in NLP techniques, tools accuracy and performance," *Pakistan J. Multidisciplinary Res.*, vol. 2, no. 2, pp. 293–308, 2021.

[45] D. Kumawat and V. Jain, "POS tagging approaches: A comparison," *Int. J. Comput. Appl.*, vol. 118, no. 6, pp. 32–38, May 2015.

[46] M. K. Saad and W. M. Ashour, "Arabic morphological tools for text mining," in *Proc. Corpora, 6th ArchEng Int. Symp., 6th Int. Symp. Elect. Electron. Eng. Comput. Sci. (EEECS)*, vol. 18, 2010, pp. 1–6.

[47] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, Emotions*. Cambridge, U.K.: Cambridge Univ. Press, Jun. 2015.

[48] T. I. Jain and D. Nemade, "Recognizing contextual polarity in phrase-level sentiment analysis," *Int. J. Comput. Appl.*, vol. 7, no. 5, pp. 12–21, Sep. 2010.

[49] J. Islam, L. Xiao, and R. E. Mercer, "A lexicon-based approach for detecting hedges in informal text," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 3109–3113.

[50] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic Inquiry and Word Count: LIWC*, vol. 71. Mahway, NJ, USA: Lawrence Erlbaum Associates, 2001.

[51] L. Van Wissen and P. Boot, "An electronic translation of the LIWC dictionary into Dutch," in *Proc. Electron. Lexicography 21st Century (eLex)*, 2017, pp. 703–715.

[52] P. Boot, "Machine-translated texts as an alternative to translated dictionaries for LIWC," *OSF Preprints*, Jan. 2021.

[53] S. Alharthi, R. Siddiq, and H. Alghamdi, "Detecting Arabic fake reviews in e-commerce platforms using machine and deep learning approaches," *J. King Abdulaziz Univ. Comput. Inf. Technol. Sci.*, vol. 11, no. 1, pp. 27–34, Sep. 2022.

[54] Google Cloud. (2022). *Translation AI*. [Online]. Available: https://cloud.google.com/translate

[55] M. Al-Ayyoub, Y. Jararweh, A. Rabab'ah, and M. Aldwairi, "Feature extraction and selection for Arabic tweets authorship authentication," *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 3, pp. 383–393, Jun. 2017.

[56] H. Cao, P. K. Sen, A. F. Peery, and E. S. Dellon, "Assessing agreement with multiple raters on correlated Kappa statistics," *Biometrical J.*, vol. 58, no. 4, pp. 935–943, Jul. 2016.

[57] M. Al-Sanabani and S. Al-Hagree, "Improved an algorithm for Arabic name matching," *Open Trans. Inf. Process.*, vol. 2015, pp. 2374–3778, 2015.

[58] Z. K. Igaab and I. A. Kareem, "Affixation in English and Arabic: A contrastive study," *English Lang. Literature Stud.*, vol. 8, no. 1, pp. 92–103, Feb. 2018.

[59] Q. Yaseen and I. Hmeidi, "Extracting the roots of Arabic words without removing affixes," *J. Inf. Sci.*, vol. 40, no. 3, pp. 376–385, Jun. 2014.

HANEN T. HIMDI received the Ph.D. degree in computer science from the University of Strathclyde, U.K. He is currently an Assistant Professor of computer science and artificial intelligence with the College of Computer Science and Engineering, University of Jeddah, Saudi Arabia. He is also enthusiastic about pioneering new ideas and developing cutting-edge technologies. Several of the academic papers are devoted to the topic of creating AI models that are useful for the Arabic language. His research interests include machine learning, natural language processing, textual analysis, deep learning, and the creation of AI models that make use of cutting-edge learning techniques.

FATMAH Y. ASSIRI received the Ph.D. degree in computer science from Colorado State University, Fort Collins, CO, USA. He was a Consultant with the Entrepreneurship and Innovation Center, University of Jeddah, Saudi Arabia. He is currently an Associate Professor of software engineering and a former Supervisor with the Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah. He translated and published a software testing book, and created the first English-Arabic glossary for software testing terminologies. His research interests include software testing and validation, automation, and data and machine learning to develop smart solutions. He was an invited speaker in many local events, and participated in other local and international competitions and hackathons as a mentor and judge; passionate about innovation and new technologies.