NGram Approach for Semantic Similarity on Arabic Short Text

Rana Husni Al-Mahmoud
Faculty of Information Technology
Applied Science Private University
Amman, Jordan

Ahmad Sharieh
Computer Science
Department King Abdullah II School of Information Technology
The University of Jordan
Amman, Jordan

***Abstract*—Measuring the semantic similarity between words requires a method that can simulate human thought. The use of computers to quantify and compare semantic similarities has become an important research area in various fields, including artificial intelligence, knowledge management, information retrieval, and natural language processing. Computational semantics require efficient measures for computing concept similarity, which still need to be developed. Several computational measures quantify semantic similarity based on knowledge resources such as the WordNet taxonomy. Several measures based on taxonomical parameters have been applied to optimize the expression for content semantics. This paper presents a new similarity measure for quantifying the semantic similarity between concepts, words, sentences, short text, and long text based on NGram features and Synonyms of NGram related to the same domain. The proposed algorithm was tested on 700 tweets, and the semantic similarity values were compared with cosine similarity on the same dataset. The results were analyzed manually by a domain expert who concluded that the values provided by the proposed algorithm were better than the cosine similarity values within the selected domain regarding the semantic similarity between the datasets' short texts.**

***Keywords*—Arabic text; Ngram; semantic sentences similarity; short text; ALMaany; natural language; semantic similarity of words; corpus-based measures**

I. Introduction

In this paper, semantic similarity is estimated by considering the similarity between bigrams synonyms related to the same domain. Paraphrase identification can detect different linguistic expressions with the same meaning or similar meanings [1]. Analyzing the similarity of meanings is part of the semantic text similarity task.

Recent advances have made social media a major source of news, with users flooded with news about similar events. Paraphrasing news articles and recognizing semantic similarities between them is a useful practice both in many general natural language processing applications and in new event detection (first story detection) on specific events [1].

For a long time, word semantic similarity has been essential to the processing of natural language and information retrieval (IR) [2]. For instance, in academic and industrial communities alike, semantic similarity has become a vital aspect of various applications in various fields. Word sense disambiguation, information retrieval, semantic searches, and explorations of biological macromolecules are prominent examples of semantic similarity applications [3]. Furthermore, it is possible to understand and categorize documents and obtain informative knowledge using semantic annotation [2].

Semantic text similarity measures the semantic similarity between two texts (documents, paragraphs, sentences, or a combination thereof). Most of the work to date on such measures has been done at the document level (that is, comparing two long texts, or one long text and one short text). Sentence-level analysis has received a lot of attention recently. As a result, training and test data were provided in multiple languages, and different approaches were developed for detecting sentence similarity. These approaches are generally classified into three types: vector space approaches, registration approaches, and other approaches such as: B. Use topic modeling for feature extraction [1].

Typically, the process of detecting two text segments' level of similarity involves, first, employing a straightforward lexical matching method and then detecting how many lexical units are contained in both input segments to calculate the similarity score [4]. This method can be improved by employing various techniques (e.g., stemming, part-of-speech tagging) or by considering different weighting and normalization factors. While these lexical similarity methods have been somewhat successful, they sometimes fail to adequately identify the semantic similarity between two texts. For instance, even though the phrases "I own a dog" and "I have an animal" are clearly similar, most contemporary similarity detection techniques do not recognize this. Often, knowledge-based or corpus-based approaches are used to detect semantic similarity at the word level [4]. These approaches have shown some success, particularly when applied to language processing tasks. Two of the most popular text-based semantic similarity approaches are to use approximations generated by query expansion and to employ the latent semantic analysis method. The former is often used for information retrieval tasks, while the latter is used to detect the similarity between texts by automatically acquiring second-order word relations from extensive collections of texts [4]. Other noteworthy methods for detecting semantic similarity are listed below [5]:

1) Structure-based measures, which use a function that computes the semantic similarity measures on ontology hierarchy structures.
2) Information content measures, which are based on the frequency of terms in a given document.
3) Feature-based measures, by which each term is described by a set of features and the similarity measure between two terms is defined as a function of their

properties.

4) Hybrid measures, which combine the structural characteristics of the previous methods to compute semantic similarity.

The following are basics and backgrounds knowledge:

WordNet created as part of a research project at Princeton University [6]. This is an extensive English vocabulary database. In WordNet, nouns, verbs, adverbs, and adjectives are organized by semantic relationships into synsets, each representing a concept.

Semantic similarity (or topological similarity) detects the similarity between terms, sentences, and documents. Similarity between sentences and documents is calculated by considering terms that describe internal concepts. Similarity at the sentence level is detected using syntactical and lexical measures [7].

The syntactic approach primarily uses syntactic dependencies to recognize semantic similarities and build a more comprehensive picture of the meaning of the compared texts. In this way, these approaches identify whether a noun is the subject or object of a verb. Lexical-based similarity approaches, on the other hand, measure similarity between texts based on character matching.

Three problems with the existing semantic measurements are the primary motivations for this work. The first issue relates to how text is represented and similarity calculations are performed. Text representation mainly concerns converting text to vectors by using lexical representation or word embedding representation. The drawback of the first one is that it depends on the occurrences of words in the text, either it occurred in the same order of occurrence or not. And it is a critical point in the semantic similarity measure and the limitation of word embedding. Most word embedding models are trained on corpora in different domains and the semantic similarity degree of the keywords depending on the domain concepts. The second issue relates to the external dictionary and ontologies. Most of them were static and not concerned about the same topic. For example, in Arabic, there is a need for more presence of these dictionaries. The third issue is the need to represent the two texts. In some applications, like in plagiarism, there is a need to find similarities between a fragment of sentences.

These problems motivate us to current work. This work depends on an updated dictionary ALMaany [8], and all needed synonyms extracted depending on the same domain. In addition, the proposed algorithm can be applied to various varieties of text length, and consider the order of keywords by taking NGram words from all texts. The present work describes a new method for measuring semantic similarity between words and concepts that uses Ngram synonyms connected to the same domain.

The first step of the proposed algorithm is crawling articles from sites to collect the most frequent words for the same domain. The extracted keywords used in the following steps:

1) Searching for tweets depending on extracted words.
2) Extracting synonyms from ALMaany [8] and concentrate on the selected domain or topic.

The proposed algorithm depends on the synonyms, and NGrams was evaluated based on 700 tweets and compared the proposed algorithm's output with the cosine similarity values. An expert assessed the results and determined that the proposed algorithm detected the semantic similarity between the dataset's short sentences more accurately than the cosine similarity values.

The remainder of the paper is organized as follows. Section II discusses previous works related to semantic similarity measures. The proposed method is described in Section III. The experimental results are then discussed in Section IV. Finally, the paper is concluded and directions for future studies are recommended in Section V.

II. Related Works

Semantic text similarity is a measure of the degree of semantic similarity between two texts, such as documents, paragraphs, sentences, words, or a combination of them [1]. Various semantic similarity approaches have been described and summarized in many surveys. For example, [9] presented the fundamental aspects of the theoretical and practical backgrounds of semantic similarity assessments of texts. They also discussed the general technology used for sophisticated text analyses (i.e., text mining), alongside discussions of relevant methodologies, architectures, and challenges. In other work, [10] explored the development of semantic similarity methods. They classified different methods as knowledge-based, corpus-based, deep neural network-based, and hybrid methods, according to their underlying principles. It starts with traditional NLP techniques (e.g., kernel-based methods) and progresses to the most recent research on transformer-based models while examining each approach's merits and disadvantages.

[11] reviewed existing approaches to measuring semantic similarity at either the document, sentence, or word level, focusing on Arabic text. The approach utilized by [1], [12] employs a set of extracted features based on lexical, syntactic, and semantic computations to detect the similarity between tweet pairs. One approach uses knowledge and corpora to express the meanings of terms to solve the issue of polysemy and includes a constituency parse tree to capture the syntactic structures of short texts. The approach also uses word alignment features to detect the similarity between tweet pairs.

Semantics is an essential aspect of studies on natural language processing. Previously, in [13], they surveyed various deep learning approaches that have been used to detect the semantics of words, sentences, and documents. However, most previous studies have considered the semantics only of documents (i.e., either two long texts or one long text and one short text have been compared). Recently, though, comparisons between individual sentences have become more common [1]. Previous researchers have also used measured semantic similarity to compare words or concepts. However, such measures are rarely used to compare multi-word phrases [4]. Three broad categories of semantic similarity detection methods are used to determine the level of similarity between words: Dictionary/ontology-based methods consider knowledge bases to gather the semantic information that is compared when determining semantic similarity [14]. Meanwhile, corpus-based methods primarily use word frequencies to determine semantic similarity. This is done based on statistics taken from extensive corpora. Finally, hybrid methods consider more than one information source to determine semantic similarity [14].

The statistical methods employed by corpus-based approaches have recently evolved. Thus, such approaches can follow one of two principal orientations [14]:

- The first is the unsupervised orientation, which involves the use of training sets and unannotated corpora. Approaches that follow this orientation can be further divided depending on the method of discrimination used (type-based or token-based). When type-based discrimination is employed, the similarity is measured by an algorithm after the contexts have been represented, which is done via high-dimensional spaces, which are defined by word co-occurrences. Meanwhile, when token-based discrimination is used, all contexts containing the target word are clustered together. Each resultant cluster comprises contexts that contain similar usages of the target word.
- The second orientation includes supervised and semi-supervised approaches, by which an annotated training corpus with the appropriate classification models is applied. Supervised methods include probabilistic methods, and they typically employ the naive Bayes algorithm and follow the maximum entropy approach [14]. Which methods are followed when using such an approach depends on how similar the evaluated examples are. These methods compare sets of learned vector prototypes using a similarity metric. This is done for each word sense. Meanwhile, other methods consider discriminating rules to make comparisons. Such methods rely on specific rules that apply to each word sense. In turn, methods based on these specific rules merge heterogeneous learning modules [14].

The Arabic language is an official language used by the United Nations, and more than 450 million people in the world speak Arabic as their first language [1]. The vocabulary of this language is rich, and its morphology is complex. It is also a synthetic language, meaning that a given morpheme can comprise a stem and affixes, which can indicate different aspects (e.g., tense, gender, and what word class a word belongs to). Moreover, different parts of speech can be affixed to each other. Arabic is a derivational, flexional, and agglutinative language. These characteristics make it difficult to conduct research on language processing and text mining, as special tools and resources are needed. An additional problem arises from the fact that the lexical and morphological features of Arabic have a profound effect on sentence analyses. If the research question addressed in this study is to be adequately answered, such challenges must be overcome. Much research has focused on various problems related to analyses of semantic similarity and developed methods for overcoming these problems. However, most of these methods are either domain-dependent or language-dependent. Moreover, little research on this matter has examined Arabic [1]. Another shortcoming of previous studies focusing on the Arabic language is that the semantic similarity analyses employed have not utilized enough resources (e.g., tools and benchmark data) due to a lack of availability. One such research work was conducted by [15], who determined semantic similarity at the sentence level using supervised learning. Specifically, their method analyzed semantic, lexical, and syntactic-semantic features, which were extracted using an Arabic dictionary, a lexical markup framework, and a learning corpus. After the method was used, its outcomes were assessed by Weka; the assessment showed that the proposed model produced highly accurate results [1]. However, the results were not as favorable when the method was used to detect semantic similarity between phrases and sentences. This is because, compared to word-level estimations, sentence-level estimations are substantially more challenging to perform since sentence-level semantics are noncompositional and involve many more possible interpretations.

When considering the Arabic language, similarity approaches face several significant challenges [16]:

- Arabic is a complicated (and often ambiguous) language.
- Arabic WordNet is a multilingual concept dictionary that maps Arabic word senses with their equivalents in English WordNet [17]. However, the Arabic database was built manually and does not contain sufficient essential information. It also contains many fewer concepts than English WordNet, and it is lacking several important semantic relations between synsets.
- Few Arabic corpora consider all possible domains and words. This is because each Arabic corpus focuses on only one domain; thus, these corpora do not contain all essential information.

Based on the above points, the cosine similarity measurement has been employed in many Arabic systems. Results show that this measure outperforms other lexical measurements.

Lexical similarity methods are unreliable when assessing the Arabic language because of the language's unique features, such as its morphology. Furthermore, the semantic similarity approach is undesirable when considering Arabic because of the aforementioned shortcomings of Arabic WordNet and Arabic corpora. Recently, the hybrid similarity approach has been considered potentially useful for examining semantic similarity in Arabic since it utilizes multiple measurement methods, thus providing more robust analyses than other techniques [16].

Twitter is a fast-growing social media tool with which people can connect and share microblog posts called tweets [18]. This tool also produces vast amounts of information. We have considered tweets in our research because tweets are limited to 280 characters. Thus, compared to the text posted on other social media platforms like Facebook (which has no post length limitations), tweets are brief yet tell complete stories that can be compared relatively easily.

Different methods for semantic similarity approaches have been recently proposed based on the aforementioned algorithms. Most of recent works based on word embedding techniques. Authors in [19] applied Word2Vec model on an English corpus to represent words in vector form. Then a Cosine Similarity method was used to calculate the similarity value. Authors in [20] presented an approach that combining LDA topic model approach with BERT word embedding for pairwise semantic similarity detection. A hybrid approach based on Word Embedding and External Knowledge Sources was used to find the semantic similarity value between two

short text. Another hybrid approach based on a WordNet proposed in [21] to measure concept semantic similarity.

This is clearly evident from previous works that most of suggested approached applied on English datasets. Therefore, more research effort is needed for computing semantic similarity for Arabic Language. In summary, a good amount of work has been invested to calculate semantic similarity either depending on external static external knowledge or by using word-embedding models that created on large corpus that is not related with the tested datasets. The reduction in problem dimension at the expense of the real values' interpretability is one of the key drawbacks of word embedding that form the vector representations [22]. And, due to Arabic WordNet limitation in keyword synonyms [23], [24], we chose ALMaany [8] to extract synonyms related to Arabic keywords within specific domain. We depended on ALMaany because it is one of the most recent dictionary and continuously updated. In addition, ALMaany is fast, free, electronic and easy to use [25], [26].

Most previous related works have considered a single corpus or a dataset when detecting the semantic similarity between sentences or documents. Differently, this work proposes a new method that considers n-gram synonyms within a single domain to detect the semantic similarity between concepts and words. Furthermore, the contributions of this work are relevant to any sentence, paragraph, or document.

III. Proposed Work

Due to its nature, the Arabic language may allow more than one meaning (and sometimes opposite meanings) to be assigned to the same word. Therefore, semantic similarity detection methods should find similarities between words related to the same domain. Because Arabic WordNet is limited to extract synonyms for Arabic terms within a particular domain [16], we choose ALMaany's [8]. ALMaany is essential to us because it is one of the most modern dictionaries and is regularly updated it is quick to access, free, computerized, and simple to use.

In this work, we considered a common news topic, namely the current relationship between Qatar and the UAE during the last quarter of 2017.

The following steps were conducted to find synonyms of the most frequently used keywords. First, sources were searched for relevant articles through Google. This was done because Twitter has a short-text format, and this step ensured that we would consider all keywords that could be found in tweets mentioning the news topic of interest. The data sources considered in this work were the websites of news agencies such as Reuters, news channels such as Aljazeera, and online versions of printed newspapers such as the Middle East. The sources used to obtain articles are presented in Table II. We searched for the main keywords, such as those found in Table I. Initially, we found almost 10,000 articles from online sources. After the main keywords were used to filter the articles, around 3000 remained. Table III shows a sample of the search results.

Second, we found the keywords most frequently used in the articles after removing stop words. Table V shows some of these words.

Third, we manually extracted synonyms related to the topic under investigation from ALMaany [8]. Table VI presents some of these synonyms.

The dataset of articles was used to find the most frequent words that needed to be used to extract tweets from Twitter. In an initial step, we need to prove our algorithm on short Arabic text; and then, in future works, it will be applied in paragraphs, and after that, on long articles

REST APIs and Streaming APIs make up most of the Twitter APIs[1]. Use the RESTful state transfer (REST) search API to search tweets from Twitter's search index. The REST API offers historical results going back as long as the search index allows (usually last seven days). The streaming API, however, returns information from the query's starting point. Real-time monitoring of a particular query is possible using streaming API. According to their website, Twitter's search API contains several restrictions.[2]. We developed a number of searches that include all combinations of the most frequent keywords extracted from websites' articles. Table IV shows a sample of the tweets. The results were filtered, and only tweets containing "**إمارات**" and "**قطر**" or "**امارات**" and "**يمن**" were kept.

TABLE I. List of Main Keywords that are used in Searching for Articles

امارات	الامارات	إمارات	الإمارات	قطر
السعودية	الرياض	دبي	أبوظبي	الدوحة

TABLE II. List of sites that are used in searching for articles

Site Url
www.alarabiya.net
www.skynewsarabia.com
www.dw.com/ar
www.bbc.com/arabic
www.france24.com/ar
www.alhurra.com
ara.reuters.com/
www.trt.net.tr/arabic
www.anb-tv.net/Arabian
www.arab48.com
www.arabi21.com
www.thenewkhalij.net
www.alhayat.com
www.alkhaleej.ae
www.aawsat.com
www.alarab.qa
arabic.rt.com
www.afp.com/ar
alkhaleejonline.net
www.cnbcarabia.com
www.middle-east-online.com
www.moheet.com
www.anntv.tv
www.huffpostarabi.com
arabic.cnn.com
www.aljazeera.net

In general, the proposed algorithm was employed to estimate the semantic similarity value of two short texts via the following process:

1) Take bigrams and trigrams.

[1]https://developer.twitter.com/en/docs/basics/getting-started
[2]https://developer.twitter.com/en/docs/basics/rate-limiting

TABLE III. SAMPLE OF ARTICLES

Article URL	Article Title	Article Content
http://www.huffpostarabi.com/mohammed-jamea/story_b_10885308.html	جولة نتنياهو الإفريقية والغياب العربي	اكتست الجولة التي قام بها رئيس الوزراء الإسرائيلي بنيامين نتنياهو خلال اليومين الماضيين إلى أربع دول إفريقية، أهمية بالغة لعدة أسباب، حيث إنها الزيارة الأولى لأرفع مسؤول إسرائيلي للمنطقة منذ فترة طويلة، إضافة إلى أن وفد نتنياهو يضم 80 رجل أعمال يمثلون 50 شركة إسرائيلية، وهذا مما يدلل على سعيه لتعزيز التبادل التجاري والمزيد من التغلغل في القارة السمراء.
http://www.huffpostarabi.com/2017/01/19/story_n_14265640.html	اكتشاف خلل تقني في نسخة iOS 10 يسبب انهيار هواتف آيفون	اكتشف أحد مستخدمي موقع يوتيوب وجودَ خلل غريب في النسخة العاشرة من نظام تشغيل الأجهزة المحمولة iOS 10، حيث يسمح ذلك الخلل للمخربين بتعطيل أي هاتف آيفون أو حاسب آيباد، عبر إرسال رسالة نصية تحتوي على الرموز التعبيرية العلم وقوس القزح.
http://www.huffpostarabi.com/2017/04/04/story_n_15808882.html	ليس العداء للإرهاب فقط سرّ الكيمياء الشخصية بين ترامب والسيسي.. إليك نقاط التشابه بينهما	الاستقبال الحميم الذي تلقاه الرئيس المصري عبد الفتاح السيسي في البيت الأبيض من قبل نظيره الأميركي دونالد ترامب جذب انتباه وسائل الإعلام الغربية والتي لفتت إلى أنه جاء بعد أسبوعين فقط مما بدا أنه رفض من قبل ترامب لمصافحة المستشارة الألمانية المرموقة أنغيلا ميركل.
http://www.middle-east-online.com/?id=256805	السيسي يوسع جهوده الدبلوماسية لإحياء السلام	الامم المتحدة (الولايات المتحدة) - حض الرئيس المصري عبدالفتاح السيسي الفلسطينيين في خطاب له امام الجمعية العامة للأمم المتحدة على "الاتحاد"، وان يكونوا مستعدين "لقبول التعايش" بسلام مع الإسرائيليين".
http://www.middle-east-online.com/?id=256121	إما حرب اقتصادية عالمية على بيونغيانغ أو يتفرق مجلس الأمن	الامم المتحدة (الولايات المتحدة) - دعت واشنطن مجلس الأمن الدولي إلى البت الاثنين بشأن عقوبات جديدة مشددة ضد كوريا الشمالية المتهمة بتهديد السلام من خلال برامجها للأسلحة النووية والتقليدية.
http://www.middle-east-online.com/?id=256064	واشنطن تطلب أقسى العقوبات على بيونغيانغ	الامم المتحدة (الولايات المتحدة) - طلبت واشنطن رسميا التصويت الاثنين في مجلس الامن على مشروع قرار بفرض عقوبات جديدة ومشددة ضد كوريا الشمالية على الرغم من معارضة الصين وروسيا، وسط دعوات الاعلام الرسمي الكوري الشمالي لتطوير قدرات البلاد النووية.
http://www.huffpostarabi.com/2017/08/05/story_n_17685688.html	واشنطن بوست: لهذه الأسباب ترى الولايات المتحدة الإمارت حليفاً مزعجاً	الإمارات حليف مهم للولايات المتحدة، لكنه في نفس الوقت سبب لها صداعاً وإزعاجاً، بسبب عدد من الممارسات الضارة لمصالح واشنطن خاصة فيما يتعلق بالأوضاع في اليمن هكذا ترى صحيفة "واشنطن بوست" الأميركية في تقرير لها الخميس 3 أغسطس/آب 2017.
https://arabic.cnn.com/health/2017/05/05/ime-050517-eman-abdel-atti	أسمن" امرأة في العالم تصل إلى أبوظبي" لاستكمال علاجها	الإمارات العربية المتحدة السمنة أمراض أمراض وأدوية صحة وحياة قد يعجبك أيضا عصائر "مضغوطة" يصل سعرها إلى 10 دولارات.
http://www.huffpostarabi.com/gamal-nassar/post_15480_b_17567306.html	مآلات الأزمة الخليجية على المنطقة العربية	الأزمة الخليجية أتت بظلالها على اضطراب الأوضاع والاستقرار في المنطقة، فالبرغم من المساعي الإقليمية المتمثلة في الكويت وتركيا، والجهود الدولية المتمثلة في أميركا وبريطانيا وألمانيا وفرنسا، وغيرها من الدول، فإن الأزمة لا تزال تراوح مكانها، ومرشحة للاستمرار لفترة طويلة.
http://www.huffpostarabi.com/2015/11/15/story_n_8558504.html	علامة تدل على أنك الطفل الأصغر في 16 عائلتك	الأشقاء الأصغر سناً زمرةٌ خاصة جداً، فالطفل الأصغر لم يعرف عالماً كان فيه الطفل الوحيد.

TABLE IV. SAMPLE OF TWEETS

Tweet Id	Tweet Text	Search Keywords
94155135639282892	دبلوماسي أميركي: السعودية والإمارات ارتكبتا خطأ بافتعال الأزمة الحالية مع #قطر https://t.co/WQPr6mxRoP	السعودية
94155135639282892	دبلوماسي أميركي: السعودية والإمارات ارتكبتا خطأ بافتعال الأزمة الحالية مع #قطر https://t.co/WQPr6mxRoP	قطر
94064821343822233	#الإمارات تُقيل مسؤولاً رياضياً بارزاً لمصافحته قطرياً! https://t.co/SpwrXbnlZ1	الامارات
94124963979320115	القطرية_للتأمين" تدير ظهرها لـ #دبي وتعلن الخروج من #الإمارات#" https://t.co/nvE5FQHfq4	الامارات
94213725979252327	وقائع رياضية تكشف صداقة #الإمارات لـ"إسرائيل" والعداء لـ #قطر https://t.co/PTDpMrCrE2	الامارات
94235116669167206	وقائع رياضية تكشف صداقة #الإمارات لـ"إسرائيل" والعداء لقطر https://t.co/PTDpMrU2vA	الامارات
94223791969935360	وقائع رياضية تكشف صداقة #الإمارات لـ"#إسرائيل" والعداء لقطر https://t.co/PTDpMrU2vA	الامارات
94197843956063027	وقائع رياضية تكشف صداقة #الإمارات لـ"إسرائيل" والعداء لـ #قطر https://t.co/PTDpMrCrE2	الامارات
94205373357121945	إقالة السركال كشفت "المستور" | #يوسف_السركال #الإمارات #حصار_قطر	الامارات
94076320381390028	#الإمارات تُقيل مسؤولاً رياضياً بارزاً لمصافحته قطرياً! https://t.co/SpwrXbnlZ1	الامارات

TABLE V. SAMPLE OF MOST FREQUENTLY KEYWORDS FOUND IN THE ARTICLES

رئيس	خلال	سعوديه	عام	دولة	دول
اول	والإمارات	يوم	منطقة	قطر	يوم
حيث	قبل	مجلس	دبي	ذلك	اخرى
وقت	اكثر	والحرين	خليج	ماضي	انها
امارات	وقال	محمد	عربية	متحده	عالم

TABLE VI. SAMPLE OF SYNONYMS FOR MOST FREQUENTLY WORDS

Keyword	Synonyms			
اخر	عرقل	مختلف	ابطأ	نهاية
مثل	شبيه	حكمة	عذب	مساو
عده	ادوات	متعدد	عدد	فترة من الزمن
امام	قبل	زعيم	رئيس	عالم
حول	سنة	عكس	ابدل	احاط
امن	سلام	اقر	هدأ	
ماضي	قديم	آنف	هالك	
عبد	رقيق	اصلح	مهد	
عمل	فعل	صنع	ممارسة	
تعاون	تضامن	تآزر	مشاركة	
مركز	رتبة	وسط	مسكن	
غير	آخر	سوى	مختلف	
ازمه	شدة	ضيق	محنة	
شهر	فضح	نشر	مجموع الايام	
كبير	ضخم	عظيم	كهل	
جميع	عامة	سواء	كل	
دون	اقل	رذيل	كتب	
اعلى	ارفع	اكرم	قمة	
بعض	قليل	جزء	قسم	
قرار	ادنى	حكم	قاع	
اولى	اجدر	اناب	فوض	
بيان	تصريح	منشور	فصيح	
جديده	حديثة	طازج	عصرية	
مجال	نطاق	موضع	شأن	
شيخ	استاذ	كهل	زعيم	
سابق	بارز	متقدم	زاحم	
قال	تحدث	تكلم	روى	
قبل	أمام	اقتنع	رضي	
امر	فرض	طلب	رآس	
وأضاف	اتبع	الف	دمج	
اضافه	ايواء	اكمال	جمع	
يمكن	ربما	لعل	توقع	
وكاله	انابة	تفويض	تخويل	
اكثر	معظم	أسرف	اوضح	
دعم	ساند	عون	اغاث	
اكبر	اضخم	اعظم	اعرض	
وقالت	تحدثت	روت	اخبرت	

2) Extract all synonyms of the most frequent unigrams.
3) Estimate the number of similarities between the bigram and trigram of two texts.

In this work, only unigrams and bigrams are used. The n-gram comparison can be increased if the degree of similarity between two sentences in terms of their meaning and structure is known. Algorithm 1 describes the steps taken to estimate semantic similarity.

TABLE VII. ABBREVIATIONS

Value	Abbreviation
Semantic Similarity for S1 UniGram	SS_S1U
Semantic Similarity for S2 UniGram	SS_S2U
Number if similar unigrams between S1 and S1	UniS1S2
Number of words in S1	S1L
Number of words in S2	S2L
Semantic Similarity for S1 BiGram	SS_S1B
Semantic Similarity for S2 BiGram	SS_S2B
Number if similar bigrams between S1 and S1	BiS1S2

$$SSV = (0.75 * B) + (0.25 * A) \quad (1)$$

UniSemilarity (A)

$$SS\_S1U = UniS1S2/S1L \quad (2)$$

$$SS\_S2U = UniS1S2/S2L \quad (3)$$

Algorithm 1 Semantic Related Words Extraction

Require: S1, S2 two Arabic short complete text (as Tweets), with lengths n and m, respectively
Ensure: Semantic similarity value (SSV)
1: Take only nouns and verbs as features from S1, S2
2: Apply the following preprocessing on S1 and S2
3: Remove non-Arabic characters
4: Remove stop words
5: Remove low-frequency tokens
6: Determine the stem of the remaining text
7: Take bigrams and trigrams of the two texts S1 and S2
8: Estimate the number of similarities between the bigrams and trigrams of the two texts (if the bigrams have the same token, give a higher value than if the two words are synonyms)
9: Estimate the SSV using Equations (1)–(5)(Depend on Table VII.

BiSimilarity (B)

$$SS\_S1B = BiS1S2/(S1L/2) \quad (4)$$

$$SS\_S2B = BiS1S2/(S2L/2) \quad (5)$$

Comparing two texts using bigrams provides an improved indication of their similarity because considering the meanings of two words gives a more accurate similarity value than considering the meaning of a single word.

Example

Assume the two texts given below are text1 and text2:

text1= وزير الخارجية البحريني: إقامة علاقات طيبة مع إيران مرهون بعدم تدخلها في الشؤون الداخلية للدول ووقف دعمها للإرهاب

text2= وزير خارجية البحرين: النظام الإيراني يدعم عددا من التنظيمات الإرهابية منها حزب الله اللبناني والميليشيات الانقلابية في اليمن

Starting from step 1 in Algorithm 1. Extracting only nouns and verbs causes text1 to become nVtext1 and text2 to become nVtext2, as follows:

nVtext1= وزير الخارجية إقامة علاقات طيبة مع إيران مرهون بعدم تدخلها في الشؤون الداخلية للدول ووقف دعمها للإرهاب

nVtext2 = وزير خارجية البحرين النظام يدعم عددا من التنظيمات منها حزب الله في اليمن

After completing all pre-processing steps(2-4) in Algorithm 1, nVtext1 becomes Text1Processed, and nVtext2 becomes Text2Processed, as follows: .

Text1Processed= وزر خرج قوم علق طوب رين رهن عدم دخل شون دخل دول وقف دعم رهب

Text2Processed= وزر خرج بحر نظم دعم عدد نظم منها حزب له يمن

Table VIII presents the results obtained from extracting unigrams, bigrams, and trigrams from Text1Processed and Text2Processed. By referring to ALMaany synonyms. The result will be as the following:

1) Number Of similarities using stemmed_Unigrams by computing how many similar words between the two

lists UniGramListText1Stemmed and UniGramListText1Stemmed and their synonyms equals 6.

2) Number Of similarities using stemmed_Bigrams by computing how many similar phrases between the two lists biGramListText1Stemmed and biGramListText2Stemmed and their synonyms equals 3.

Depending on Equation 1, SSV for text1 and text2 was 0.42 SSV=0.75 * (3/(14/2)) + 0.25*(6/14)= 0.42

Applying the cosine similarity measure to the same texts (text1 and text2) and considering term frequency as the features of words generated a similarity value between the two texts of 0.3.

In other research, they used Arabic WordNet to find the extent to which two concepts are related [3]. However, the similarity values they calculated did not depend on a specific domain, so the obtained values may have been substantially different from the actual values. For example, when extracting synonyms of an Arabic word (**قوات**), the related words and synonyms from Arabic WordNet, such as **قوات المارينز الأمريكية**, **أسطول** , and **وحدة** (see Table IX) were presented. Such comparisons between ambiguous words yield misleading values.

TABLE VIII. NGRAM OF TEXT1PROCESSED AND TEXT2PROCESSED

Text1Processed UniGram	Text2Processed UniGram	Text1Processed BiGram	Text2Processed BiGram	Text1Processed TriGram	Text2Processed TriGram
قوم	منها	وزر خرج	وزر خرج	علق طوب رين	وزر خرج بحر
رهن	عدد	رين رهن	نظم دعم	رين رهن عدم	بحر نظم دعم
شون	له	رهن عدم	دعم عدد	وزر خرج قوم	نظم دعم عدد
رهب	يمن	دعم رهب	عدد نظم	دخل شون دخل	عدد نظم منها
دعم	حزب	شون دخل	خرج بحر	دول وقف دعم	نظم منها حزب
علق	بحر	طوب رين	له يمن	دخل دول وقف	خرج بحر نظم
طوب	نظم	علق طوب	نظم منها	وقف دعم رهب	دعم عدد نظم
دخل	دعم	قوم علق	حزب له	قوم علق طوب	منها حزب له
رين	خرج	دول وقف	منها حزب	خرج قوم علق	حزب له يمن
عدم	وزر	عدم دخل	بحر نظم	شون دخل دول	
دول		خرج قوم		طوب رين رهن	
خرج		دخل شون		رهن عدم دخل	
وزر		دخل دول		عدم دخل شون	
وقف		وقف دعم			

TABLE IX. SYNONYMS OF ARABIC WORD FROM ARABIC WORDNET

Keyword	Synonyms
حَشْد	حِشْد
	حَشْدُ القُوَات
	تَجمع
فَرْد	حَافِظ للسِلام
	فَرْدٌ من قُوَّات حفظ السِلام
	جَيش
	جُنْدِي
نَشَرَ	نِشْر
	نَشَرَ القُوَات
	لَعِب
	وَضَع
شُرْطَة	شرطة
	الأمن
	بوليس
	قُوَّات الشرطة
	رجال الشرطة
	شرطي
	مجموع
قُوَّات	قُوات المارينز الأمريكية
	أسطول
	وحدة
جَيْش	ضَابط جَيش
	جَيش نِظَامِي
	قوات مسلحة
	عقيد
	لواء
	مشير
	مُلازم
	نقيب
	ضَابط
	جَيش

IV. EXPERIMENTAL RESULT

Java was utilized in implementing the proposed algorithm, and the implementation was run to collect full articles from websites (see Table II) using Google Search API [27]. We used the same keywords in Table I to search Twitter accounts. We utilized Twitter4J, an unauthorized Java tool for the Twitter API, to extract tweets. The proposed algorithm's ability to detect the semantic similarity between 700 tweets was tested.

The semantic similarity measure used most often in previous work is cosine similarity. We applied cosine similarity to the same data set. Table X illustrates sample of the comparison between the SSV values and cosine similarity values. The results were analyzed manually by a domain expert who concluded that the values provided by the proposed algorithm were better than the cosine similarity values within the selected domain regarding the semantic similarity between the datasets' short texts. Trigrams and more n-grams can be considered to search for more equivalent documents. NGram can be increases as long as the length of the text increases. For examples For very short text, unigrams can be used. And for short text, bigrams can be used. So, as long as text length increases, n can be increased fot example triGram and FourGram.

Now, let us take a closer look at the values of the comparison results from examples in Table X. The semantic similarity values were enhanced based on the following:

1) We took into consideration synonyms of NGram words from the updated dictionary and concentrated on the synonyms from the same domain.
2) BiGrams similarity value increases the indication of similarity between the sentences. In our approach,BiGrams similarity value was given more weight over UniGram similarity.

As an initial step, the proposed algorithm was tested on short text. This algorithm can be applied to long text like documents and articles. Also, it can be applied to paragraphs, sentences,...etc. This algorithm has some limitations. One of these limitations it is based on an external dictionary. This problem can be solved by automatically extracting semantically related words depending on the same corpus.

V. CONCLUSION AND FUTURE WORK

Freely available semantic similarity measurements are essential for advancing many NLP research areas, especially for under-resourced languages such as Arabic. The lack of a commonly used, trustworthy, comprehensive dictionary and ontology of semantic similar words and phrases is recognized as one of the most challenging and exciting problems facing Arabic NLP applications. However, manually computing the similarity degree of two texts is costly and nearly impossible. In this study, we have sought to tackle the phenomenon of computing the degree of semantic similarity of Arabic short texts.

This work introduced a novel similarity measure based on n-gram synonyms connected to the same domain designed to quantify the semantic similarity between concepts and words. The proposed algorithm was evaluated on a dataset of 700 tweets, and the semantic similarity values and cosine

TABLE X. COMPARISON RESULTS OF PROPOSED APPROACH AND COSINE SIMILARITY

Text1	Text2	SSV	Time (MS)	Cosine	Time (MS)
بدء التصويت في حوار #rtarabic انفتاح الإمارات على حزب الإصلاح اليمني، هل سيغير موازين القوى في الميدان؟	السعودية والإمارات تسعيان للتحالف مع حزب الإصلاح وهما المتهمتان بإسقاط #اليمن بيد# https://t.co/6lzxxZjotJ …الحوثيين للتخلص منه عام 2014.. ف	1	14	0.2	31
الإصلاح و #الإمارات.. تقارب يرسم خارطة التحالفات الجديدة في #اليمن https://t.co/AWKp9Zmyic	بن سلمان وبن زايد يلتقيان رئيس حزب الإصلاح اليمني https://t.co/9p7LwqFGKA https://t.co/gvXcUUe1ZU السعودية #الإمارات #اليمن |	1	16	0.3	39
بدء التصويت في حوار #rtarabic انفتاح الإمارات على حزب الإصلاح اليمني، هل سيغير موازين القوى في الميدان؟	بن سلمان وبن زايد يلتقيان رئيس حزب الإصلاح اليمني https://t.co/9p7LwqFGKA https://t.co/gvXcUUe1ZU السعودية #الإمارات #اليمن |	1	12	0.3	33
بن سلمان وبن زايد يلتقيان رئيس حزب الإصلاح اليمني https://t.co/9p7LwqFGKA https://t.co/gvXcUUe1ZU السعودية #الإمارات #اليمن |	بدء التصويت في حوار #rtarabic انفتاح الإمارات على حزب الإصلاح اليمني، هل سيغير موازين القوى في الميدان؟	1	13	0.3	43
بن سلمان وبن زايد يلتقيان رئيس حزب الإصلاح اليمني https://t.co/9p7LwqFGKA https://t.co/gvXcUUe1ZU السعودية #الإمارات #اليمن |	السعودية والإمارات تسعيان للتحالف مع حزب الإصلاح وهما المتهمتان بإسقاط #اليمن بيد# https://t.co/6lzxxZjotJ …الحوثيين للتخلص منه عام 2014.. ف	1	19	0.4	52
السعودية والإمارات تسعيان للتحالف مع حزب الإصلاح وهما المتهمتان بإسقاط# https://t.co/6lzxxZjotJ …اليمن بيد الحوثيين للتخلص منه عام 2014.. ف#	بن سلمان وبن زايد يلتقيان رئيس حزب الإصلاح اليمني https://t.co/9p7LwqFGKA https://t.co/gvXcUUe1ZU السعودية #الإمارات #اليمن |	1	20	0.4	49
!الإمارات تُقبل مسؤولاً رياضياً بارزاً لمصافحته قطرياً# https://t.co/SpwrXbnlZ1 https://t.co/kU0i6zDii7 قطر #رياضة# |	!الإمارات تُقبل مسؤولاً رياضياً بارزاً لمصافحته قطرياً# https://t.co/SpwrXbnlZ1 https://t.co/SJ2iC2fyMv نيض الخليج العربي #قطر #رياضة# |	1	11	0.8	34
السعودية والإمارات تسعيان للتحالف مع حزب الإصلاح وهما المتهمتان# …بإسقاط #اليمن بيد الحوثيين للتخلص منه عام 2014.. ف https://t.co/6lzxxZjotJ	بدء التصويت في حوار #rtarabic انفتاح الإمارات على حزب الإصلاح اليمني، هل سيغير موازين القوى في الميدان؟	0.88	19	0.2	32

similarity were compared. A domain expert carefully examined the results and concluded that, considering the semantic similarity between the short sentences of the datasets, the values produced by the proposed algorithm were more accurate than the cosine similarity values within the chosen domain. This work can be a basis for other works investigating the same semantic similarity problems. Semantic similarity between texts written in the Arabic language can help determine, for example, who originally published a piece of news, who rephrased a previously published news article and claimed to be the original source, and how to extend it to solve problems related to plagiarism. Further research is a domain ontology that includes all relations between words in the domain. This domain will help determine how some words are related and how they are different. This knowledge can then be used for other purposes, such as to perform sentiment analyses of the Arabic language in this domain.

REFERENCES

[1] A.-S. Mohammad, Z. Jaradat, A.-A. Mahmoud, and Y. Jararweh, "Paraphrase identification and semantic text similarity analysis in arabic news tweets using lexical, syntactic, and semantic features," *Information Processing & Management*, vol. 53, no. 3, pp. 640–652, 2017.

[2] A. Faaza, D. James, A. Zuhair, A. Keeley *et al.*, "Arabic word semantic similarity," *International Journal of Cognitive and Language Sciences*, vol. 6, no. 10, pp. 2497–2505, 2012.

[3] F. A. Almarsoomi, J. D. OShea, Z. Bandar, and K. Crockett, "Awss: An algorithm for measuring arabic word semantic similarity," in *2013 IEEE international conference on systems, man, and cybernetics*. IEEE, 2013, pp. 504–509.

[4] R. Mihalcea, C. Corley, C. Strapparava *et al.*, "Corpus-based and knowledge-based measures of text semantic similarity," in *Aaai*, vol. 6, no. 2006, 2006, pp. 775–780.

[5] T. Slimani, "Description and evaluation of semantic similarity measures approaches," *arXiv preprint arXiv:1310.8059*, 2013.

[6] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[7] Y. Li, D. McLean, Z. A. Bandar, J. D. O'shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 8, pp. 1138–1150, 2006.

[8] almaany, "Translation and meaning in almaany," https://www.almaany.com/en/dict/ar-en/, July 2022.

[9] A. Rozeva and S. Zerkova, "Assessing semantic similarity of texts–methods and algorithms," in *AIP Conference Proceedings*, vol. 1910, no. 1. AIP Publishing LLC, 2017, p. 060012.

[10] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—a survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–37, 2021.

[11] M. Alian and A. Awajan, "Arabic semantic similarity approaches-review," in *2018 International Arab Conference on Information Technology (ACIT)*. IEEE, 2018, pp. 1–6.

[12] J. Yang, Y. Li, C. Gao, and Y. Zhang, "Measuring the short text similarity based on semantic and syntactic information," *Future Generation Computer Systems*, vol. 114, pp. 169–180, 2021.

[13] S. Zad, M. Heidari, P. Hajibabaee, and M. Malekzadeh, "A survey of deep learning methods on semantic similarity and sentence modeling," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021, pp. 0466–0472.

[14] A. Zouaghi, M. Zrigui, G. Antoniadis, and L. Merhbene, "Contribution to semantic analysis of arabic language," *Advances in Artificial Intelligence*, vol. 2012, 2012.

[15] W. Wali, B. Gargouri *et al.*, "Supervised learning to measure the semantic similarity between arabic sentences," in *Computational collective intelligence*. Springer, 2015, pp. 158–167.

[16] S. S. Aljameel, J. D. O'Shea, K. A. Crockett, and A. Latham, "Survey of string similarity approaches and the challenging faced by the arabic language," in *2016 11th International Conference on Computer Engineering & Systems (ICCES)*. IEEE, 2016, pp. 241–247.

[17] H. M. Alghamdi, A. Selamat, and N. S. A. Karim, "Arabic web pages clustering and annotation using semantic class features," *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4, pp. 388–397, 2014.

[18] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the American society for information science and technology*, vol. 60, no. 11, pp. 2169–2188, 2009.

[19] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, "Word2vec model analysis for semantic similarities in english words," *Procedia Computer Science*, vol. 157, pp. 160–167, 2019.

[20] N. Peinelt, D. Nguyen, and M. Liakata, "tbert: Topic models and bert joining forces for semantic similarity detection," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 7047–7055.

[21] Y. Cai, Q. Zhang, W. Lu, and X. Che, "A hybrid approach for measuring semantic similarity based on ic-weighted path distance in wordnet," *Journal of intelligent information systems*, vol. 51, no. 1, pp. 23–47, 2018.

[22] L. Gutiérrez and B. Keith, "A systematic literature review on word embeddings," in *International Conference on Software Process Improvement*. Springer, 2018, pp. 132–141.

[23] A. Saif, N. Omar, U. Z. Zainodin, and M. J. Ab Aziz, "Building sense tagged corpus using wikipedia for supervised word sense disambiguation," *Procedia Computer Science*, vol. 123, pp. 403–412, 2018.

[24] H. A. Abdeljaber, "Automatic arabic short answers scoring using longest common subsequence and arabic wordnet," *IEEE Access*, vol. 9, pp. 76 433–76 445, 2021.

[25] N. Altuwairesh, "Successful translation students' use of dictionaries," *International Journal of English Linguistics*, vol. 12, no. 2, 2022.

[26] N. Sabbah and R. Alsalem, "Female translation students' knowledge and use of online dictionaries and terminology data banks: A case study," *AWEJ for Translation & Literary Studies*, vol. 2, no. 2, 2018.

[27] Google, "Google," https://developers.google.com/custom-search/, July 2022.