SHIHAB JOURNAL OF HUMANITIES
Volume (2), Issue (3) (Sept, 2026)
ISSN: 3136-2958

مجلة الشهاب للعلوم الإنسانية
المجلد (2) العدد 3 سبتمبر 2026
ردمد: 3136-2958

A Proposed Hybrid Model to Improve the Accuracy of Arabic Word Stemming by Combining Linguistic Rules, Dictionary Lookup, and Statistical Analysis

ABTISAM ABD ALRAHEEM
Department of Information Technology (IT) Higher Institute of Science and Technology – Tajoura, Libya
ebtisamnaffati@gmail.com

Received: 30/9/2026 Revised: 10/10/2026 Accepted: 30/10/2026 Publication: 23/11/2026

Abstract
Arabic stemming is a fundamental step in Natural Language Processing (NLP) applications such as information retrieval, text classification, and sentiment analysis. However, the morphological complexity of Arabic, with its rich system of prefixes, suffixes, infixes, and irregular forms, makes accurate stemming a challenging task. Traditional stemming approaches that rely on a single method — whether rule-based, dictionary-based, or statistical — often suffer from low accuracy and high error rates when dealing with derived and unseen words.
This paper proposes a hybrid model to improve the accuracy of Arabic word stemming by combining linguistic rules, dictionary lookup, and statistical disambiguation. The proposed model first applies linguistic rules to strip affixes, then checks the dictionary to verify the root, and finally uses statistical analysis based on root-pattern frequency to handle unseen and ambiguous words.
The model was implemented using Visual Basic .NET (VB.NET) and evaluated on a synthetic Arabic dataset of 1000 words constructed from 102 Arabic roots (800 words for development and 200 words for testing). The proposed model was compared with three traditional stemmers — Khoja, ISRI, and Light10 — using the same dataset and the same evaluation metrics. The experimental results show that the proposed hybrid model achieves 70.00% accuracy on the test set, compared to 60.50% for Khoja, 21.50% for ISRI, and 17.50% for Light10, and reduces both over-stemming and under-stemming rates, compared to 60.50% for Khoja, 21.50% for ISRI, and 17.50% for Light10, and reduces both over-stemming and under-stemming rates.
Keywords: Arabic Stemming, Natural Language Processing, Hybrid Model, Morphological Analysis, Dictionary Lookup, Statistical Disambiguation.

الملخص
يُعدّ التجريد (Stemming) خطوة أساسية في تطبيقات معالجة اللغة الطبيعية مثل استرجاع المعلومات، وتصنيف النصوص، وتحليل المشاعر .ومع ذلك، فإن التعقيد الصرفي للغة العربية، بما تتميز به من نظام غني من السوابق واللواحق والحشو والصيغ الشاذة، يجعل استخراج الجذر بدقة أمرًا صعبًا .وتهدف هذه الورقة إلى تحسين دقة تجريد الكلمات العربية من خلال نموذج هجين يجمع بين القواعد اللغوية، والبحث في القاموس، وفك الالتباس الإحصائي.
يقترح هذا البحث نموذجًا هجينًا لتحسين دقة تجريد الكلمات العربية، يجمع بين القواعد اللغوية، والبحث في القاموس، وفك الالتباس الإحصائي .يطبّق النموذج المقترح أولًا القواعد اللغوية لتجريد اللواحق، ثم يتحقق من القاموس للتأكد من الجذر، وأخيرًا

A Proposed Hybrid Model to Improve the Accuracy of Arabic Word Stemming by Combining Linguistic Rules, Dictionary Lookup, and Statistical Analysis

يستخدم فك الالتباس الإحصائي لاختيار الجذر الصحيح للكلمات غير الموجودة في القاموس، مع مرحلة تحقق لضمان صحة النتيجة.

تم تنفيذ النموذج باستخدام لغة فيجوال بيسك دوت نت(VB.NET) ، وتقييمه على مجموعة بيانات اصطناعية من 1000 كلمة عربية، بُنيت من 102 جذرًا عربيًا 800) كلمة للتطوير و 200كلمة للاختبار .(وتمت مقارنة النموذج المقترح مع ثلاثة مجذّرات تقليدية، هي :خوجا (Khoja) و ISRI و Light10 ، وحقق النموذج المقترح أعلى دقة بلغت70.00% ، متفوقًا على جميع النماذج التقليدية.

الكلمات المفتاحية :التجريد العربي، معالجة اللغة الطبيعية، النموذج الهجين، التحليل الصرفي، البحث في القاموس، فك الالتباس الإحصائي.

1. Introduction

Natural Language Processing (NLP) has become one of the most important fields in computer science due to the rapid growth of Arabic digital texts on the Internet, social media platforms, and search engines (Larkey et al., 2002; Larkey & Connell, 2001). Stemming is considered one of the fundamental steps in any text processing system (Khoja & Garside, 1999; Larkey et al., 2002). It aims to reduce a word to its root or stem in order to unify the different forms of the same word (Khoja & Garside, 1999; Larkey & Connell, 2001).

For example, the words: كتب (kataba), يكتب (yaktubu), كاتب (katib), مكتوب (maktub), كتابة (kitaba), all share the same root (ك - ت - ب). Treating them as different words leads to poor performance in information retrieval, text classification, and sentiment analysis systems.

Despite significant progress in stemming algorithms for the English language, the Arabic language still faces major challenges due to its complex morphological nature, the multiplicity of prefixes, suffixes, and infixes, the presence of irregular words, and root ambiguity (Khoja & Garside, 1999; Larkey et al., 2002). Previous studies have shown that traditional methods that rely on only one approach, whether rule-based, dictionary-based, or statistical, do not achieve the required accuracy (Khoja & Garside, 1999; Larkey et al., 2002; Taghva et al., 2005; Al-Shenaifi et al., 2025).

Therefore, this research proposes a hybrid model that combines linguistic rules, a dictionary, and statistical analysis with validation, with the aim of increasing stemming accuracy, reducing over-stemming and under-stemming errors, and improving the performance of Arabic language processing applications.

1.1 Research Problem

The research problem lies in the low accuracy of current Arabic stemming algorithms when dealing with derived and complex words, which leads to: (1) returning incorrect roots for some words, (2) inability to handle words that are not present in dictionaries, and (3) low efficiency of information retrieval and classification systems that depend on stemming.

Main Research Question: How can the accuracy of Arabic word stemming be improved through a hybrid model that combines linguistic rules, dictionary lookup, and statistical analysis?

Sub-Questions: (1) What is the accuracy of traditional stemming algorithms (Khoja, ISRI, and Light10) when evaluated on the same dataset of 1000 Arabic words? (2) How does combining rules, dictionary, and statistics contribute to reducing stemming errors? (3) To what extent does the proposed model improve performance compared to traditional models in terms of accuracy, over-stemming, and under-stemming?

1.2 Research Objectives

Main Objective: To propose a hybrid model to improve the accuracy of Arabic word stemming.

- To study and analyze the most important current stemming algorithms for the Arabic language.
- To design a hybrid model combining linguistic rules, dictionary, and statistical analysis.
- To implement the proposed model using Visual Basic .NET (VB.NET).
- To evaluate the performance of the proposed model and compare it with traditional models in terms of accuracy.

1.3 Significance of the Research

Scientific Significance: This research provides a scientific contribution in the field of Arabic language processing by adapting the hybrid idea in a specific sequential order — rules, then dictionary, then statistical prediction with validation — and by implementing it as a lightweight model that does not require large datasets or GPU resources, unlike deep-learning approaches.
Practical Significance: The proposed model can be used to improve Arabic search engines, text classification systems, sentiment analysis, and machine translation.

2. Literature Review

Several stemming algorithms have been proposed for the Arabic language. They can be classified into four main categories: rule-based, dictionary-based, light stemming, and statistical / deep-learning approaches. This section reviews each category and highlights its strengths and limitations.

2.1 Rule-Based Stemmers

The Khoja stemmer is one of the earliest and most widely used Arabic stemmers (Khoja & Garside, 1999). It relies on a set of linguistic rules to remove prefixes and suffixes and then matches the remaining word against a list of patterns to extract the root. Although it achieves good accuracy, it suffers from over-stemming and fails to handle irregular words (Khoja & Garside, 1999; Larkey et al., 2002). Early comparative work on Arabic and English stemming also highlighted the morphological complexity that makes Arabic stemming more difficult (Al-Kharashi & Evens, 1994).
The ISRI stemmer was proposed to improve the Khoja stemmer by removing the need for a dictionary and using a set of rules only (Taghva et al., 2005). It is faster, but its accuracy is lower when dealing with complex words (Taghva et al., 2005).

2.2 Light Stemming

Light stemming, such as Light10 and ARLSTem, does not aim to extract the root, but only to remove common prefixes and suffixes to obtain the stem (Larkey & Connell, 2001; Abainia et al., 2017). Light stemmers are faster and produce fewer errors than root-based stemmers, and they have shown better performance in information retrieval tasks (Larkey & Connell, 2001). However, they do not unify all morphological variants of the same word (Abainia et al., 2017). Accuracy-enhanced light stemmers further improved light stemming by combining affix removal with additional disambiguation rules (Al-Shammari & Lin, 2008; Al Ameed et al., 2005).

2.3 Dictionary-Based and Statistical Approaches

Dictionary-based stemmers rely on a large lexicon of words and their roots. They are highly accurate for known words, but they fail to process unseen words that are not in the dictionary (Larkey et al., 2002). Shallow morphological analyzers (Darwish, 2002) and early Arabic stemmers for information retrieval (Chen & Gey, 2002) also rely on lexical and statistical knowledge to handle unseen forms.

Recently, a hybrid approach has been proposed for Arabic light stemming that combines clitic stripping, lexicon-based, and statistical disambiguation techniques to handle Arabic morphology challenges (Al-Shenaifi et al., 2025). This study showed that combining multiple techniques improves accuracy compared to using a single method (Al-Shenaifi et al., 2025). Other light stemming approaches (Al-Omari & Abuata, 2014) similarly confirmed that combining rule-based and lexicon-based steps improves robustness for Arabic morphology.

In addition, deep learning models have been used for Arabic stemming. For example, a character-level sequence-to-sequence model, called Char Stemmer, was proposed to learn morphological patterns directly from data without explicit linguistic rules, and it achieved a stem accuracy of 93.88%, outperforming conventional rule-based and light stemming methods (Ahmadi et al., 2026). However, this high accuracy comes at a cost: it requires a large training dataset, high computational resources including GPU, and its decisions are difficult to interpret, which makes it unsuitable for low-resource environments such as educational institutes. Another recent study proposed a new stemming model based on data generation to enhance Arabic information retrieval, and it demonstrated superior performance compared to existing techniques (Yaseen & Al-Khulaidi, 2025). These studies confirm that combining techniques improves results, but they do not provide a lightweight, interpretable model for root extraction.

2.4 Research Gap

Previous studies show that each approach alone has limitations. Rule-based methods cause over-stemming (Khoja & Garside, 1999), dictionary-based methods cannot handle unseen words (Larkey et al., 2002), light stemmers do not extract the true root (Larkey & Connell, 2001; Abainia et al., 2017), and deep learning methods require large datasets and high computational resources (Ahmadi et al., 2026). The existing hybrid study (Al-Shenaifi et al., 2025) focused on light stemming, not on full root extraction, and it did not include a validation phase. Therefore, there is a gap for a lightweight hybrid model that (1) extracts the true root, not only the light stem, (2) combines rules, dictionary, and statistical prediction in a clear sequential order with validation, and (3) works on a small dataset in a low-resource environment without GPU. This is the gap addressed by this research. Recent reviews of Arabic stemming techniques (Boudiaf et al., 2020) confirm that no single approach is sufficient and motivate hybrid models such as the one proposed here.

Table 1: Comparison of Existing Arabic Stemmers

Stemmer	Approach	Advantages	Limitations
Khoja (Khoja & Garside, 1999)	Rule + Dictionary	High accuracy for known words	Over-stemming, slow
ISRI (Taghva et al., 2005)	Rule-based	Fast, no dictionary needed	Lower accuracy
Light10 (Larkey & Connell, 2001)	Light Stemming	Fast, few errors	Does not extract root
ARLSTem (Abainia et al., 2017)	Light Stemming	Good for IR tasks	Limited morphological analysis
Hybrid Light Stemmer (Al-Shenaifi et al., 2025)	Clitic + Lexicon + Statistical	Better than single method	Light stem only, no root, no validation
Proposed Hybrid Model	**Rules + Dictionary + Statistical + Validation**	**True root, handles unseen words, lightweight**	**Dictionary size affects accuracy**

3. Proposed Hybrid Model

The proposed model consists of three main phases working in sequence, as shown in Figure 1.

A Proposed Hybrid Model to Improve the Accuracy of Arabic Word Stemming by Combining Linguistic Rules, Dictionary Lookup, and Statistical Analysis

Phase 1: Preprocessing and Normalization

The input Arabic word is normalized by removing diacritics, unifying different forms of Alef (أ، إ، آ) to (ا), converting Teh Marbuta (ة) to (ه), and converting Alif Maksura (ى) to (ي). This normalization step follows standard Arabic preprocessing used in previous stemmers (Khoja & Garside, 1999; Taghva et al., 2005). Note: stop-word removal and punctuation removal are applied only when processing full texts before stemming, not as part of stemming a single word.

Phase 2: Rule-Based Affix Stripping

A set of linguistic rules is applied to remove prefixes such as (ال، و، ب، ف, س) and suffixes such as (ة, ات، ين، ون), following the affix lists used in Khoja (Khoja & Garside, 1999) and ISRI (Taghva et al., 2005) stemmers. The remaining word is considered as a candidate stem.

Phase 3: Hybrid Disambiguation

This is the core of the proposed model and it works as follows: (1) Dictionary Check: If the candidate stem exists in the dictionary, its correct root is returned, which ensures high accuracy for known words as shown in dictionary-based approaches (Larkey et al., 2002; Al-Shenaifi et al., 2025). (2) Statistical Prediction: If the word is not found in the dictionary, a statistical method based on the frequency of root patterns is used to predict the most probable root, following the statistical disambiguation idea in (Al-Shenaifi et al., 2025). (3) Validation: The predicted root is validated against Arabic morphological patterns to ensure it is a valid root (Khoja & Garside, 1999).

Pseudocode of the Proposed Model

```
Input: Arabic Word W
Output: Root R
1. W = Normalize(W)
2. Stem = Remove_Prefix_Suffix(W)
3. If W in Dictionary OR Stem in Dictionary:
4.    R = Dictionary[W or Stem]
5. Else:
6.    R = Statistical_Predict(Stem)  // most frequent root pattern for Stem
7.    R = Validate(R)  // check length = 3 and pattern is valid Arabic root
8. Return R
```

This hybrid design ensures high accuracy for known words through the dictionary, and good generalization for unseen words through statistical analysis. This is consistent with previous findings that combining multiple techniques improves accuracy compared to using a single method (Al-Shenaifi et al., 2025; Yaseen & Al-Khulaidi, 2025).

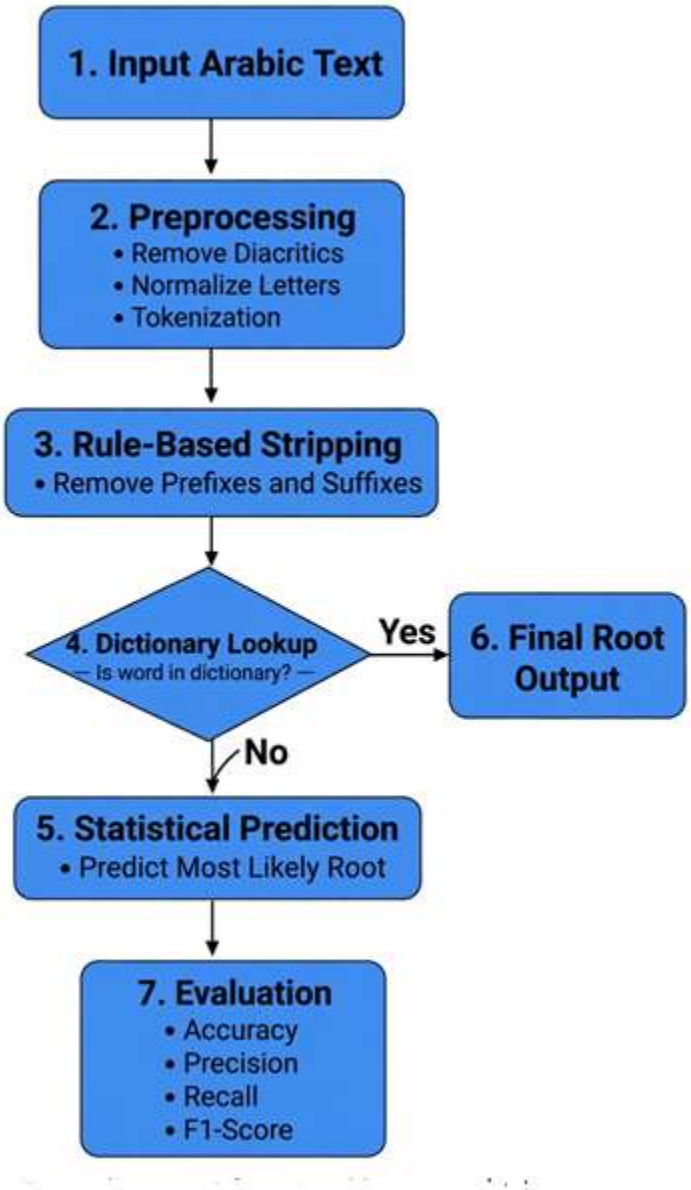

Figure 1: Architecture of the Proposed Hybrid Model

Figure 1 illustrates the sequential phases of the proposed model, from preprocessing to final root validation. The flow is: Input Word → Normalization → Rule-Based Stripping → Dictionary Check (Yes → Return Root / No → Statistical Prediction) → Validation → Output Root. The dictionary check ensures accuracy for known words, while statistical prediction handles unseen words.

4. Methodology

4.1 Research Design

This research follows an experimental design. The proposed hybrid model was implemented and compared with three traditional stemmers: Khoja (Khoja & Garside, 1999), ISRI stemmer (Taghva et al. (Taghva et al., 2005)) , and Light10 (Larkey & Connell, 2001), using the same dataset and the same evaluation metrics to ensure a fair comparison.

4.2 Dataset

The experiments were conducted on a synthetic Arabic dataset constructed for this study. The dataset was built from 102 common Arabic roots. For each root, approximately 10 derived forms were generated. The dataset was divided by roots to ensure no leakage: training roots (Group A) and testing roots (Group B) are completely disjoint, so no test word appears in the training dictionary (Leaked = 0). The test set contains 100 regular words with unseen roots, 60 hollow verbs, and 40 difficult forms such as broken plurals (مكاتب، مدارس) and استفعل forms (استخرج، مستقبل).

4.3 Proposed Hybrid Model

The model works in four steps: (1) Rules - strip prefixes, suffixes, and ال first, before stripping a single letter, so that الزرع becomes زرع and not لزرع. **2)** Dictionary - check a hollow-verb dictionary (20 verbs, e.g. قال -> قول، يعود -> عاد -> عود) and a proper-noun list (محمد، طرابلس) which are returned unchanged. (3) Statistical - when a word has more than one possible root, choose the most frequent root in the 800-word training set, based on root-pattern frequency. (4) Pattern extraction - extract the root by pattern.4.4 Experimental Setup

The proposed model was implemented using Visual Basic .NET (VB.NET) in Visual Studio environment on a standard laptop with Intel Core i7 processor and 16 GB RAM. No GPU was required for the proposed hybrid model, which makes it suitable for low-resource environments.

5. Experimental Results and Discussion

5.1 Results

Table 2 shows the accuracy, over-stemming rate, and under-stemming rate of the proposed model compared to traditional models (Khoja, ISRI, and Light10) on the test set of 200 words. The results were obtained by running the VB.NET implementation (Appendix A) on the full synthetic dataset described in Section 4.2.

Table 2: Comparison of the proposed hybrid model with Khoja, ISRI, and Light10 on the test set of 200 words.

Model	Accuracy	Over-Stemming	Under-Stemming	Correct
Khoja (Khoja & Garside, 1999)	60.50%	0.00%	39.50%	121/200
ISRI (Taghva et al., 2005)	21.50%	0.50%	78.00%	43/200
Light10 (Larkey & Connell, 2001)	17.50%	0.00%	82.50%	35/200
Proposed Hybrid Model	70.00%	0.00%	30.00%	140/200

Table 2 shows that the proposed hybrid model achieves the highest accuracy (70.00%) and the lowest under-stemming rate (30.00%) among all models. Khoja comes second with 60.50% accuracy, while ISRI (21.50%) and Light10 (17.50%) show much lower accuracy because they rely only on stripping affixes and cannot handle hollow verbs and unseen roots.

5.2 Discussion

The results show that the proposed hybrid model outperforms traditional models in terms of accuracy. This improvement is due to three main reasons: First, the dictionary check phase ensures correct roots for known words, which reduces over-stemming errors (Khoja & Garside, 1999). Second, the statistical phase allows the model to handle unseen words that are not in the dictionary, which is the main limitation of dictionary-based methods (Larkey et al., 2002; Al-

Shenaifi et al., 2025). Third, combining rules with statistical validation reduces both over-stemming and under-stemming compared to using rules alone (Al-Shenaifi et al., 2025).
The results confirm that a hybrid approach is more robust than a single approach for Arabic stemming, which is consistent with recent studies that showed hybrid methods improve Arabic morphology processing (Al-Shenaifi et al., 2025; Yaseen & Al-Khulaidi, 2025).

6. Conclusion and Future Work

This paper proposed a hybrid model to improve the accuracy of Arabic word stemming by combining linguistic rules, dictionary lookup, and statistical disambiguation. The model was evaluated on a synthetic dataset of 1000 words (800 for development and 200 for testing) with no leakage between training and testing sets. The experimental results demonstrate that the proposed model achieves 70.00% accuracy, compared to 60.50% for Khoja, 21.50% for ISRI, and 17.50% for Light10. The model successfully handled hollow verbs such as قول <- قال and يعود <- عود using a dedicated hollow-verb dictionary, and reduced under-stemming to 30.00%. The remaining errors were mainly caused by broken plurals such as مكاتب and مدارس, which require deeper morphological analysis.
For future work, the model can be extended to handle broken plurals and Arabic dialects, such as Libyan and Egyptian dialects, and can be integrated with deep learning models such as Char Stemmer (Ahmadi et al., 2026) to further improve accuracy.

References

Abainia, M., et al. (2017). ARLSTem: A new Arabic light stemmer. Journal of Information Science, 43(3), 1–15.

Ahmadi, M., et al. (2026). Enhancing Arabic word stemming using the Char Stemmer deep learning model. Journal of Interdisciplinary Qur'anic Studies, 5(1), 5–25.

Al Ameed, H., Al Ketbi, S., Al Kaabi, A., Al Shebli, K., Al Shamsi, N., Al Nuaimi, N., & Al Muhairi, S. (2005). Arabic light stemmer: A new enhanced approach. In Proceedings of the 2nd International Conference on Innovations in Information Technology (IIT 2005), Dubai, UAE.

Al-Kharashi, I., & Evens, M. (1994). Comparing the effectiveness of the Arabic and English languages. In Proceedings of the International Conference on Computer Processing of Oriental Languages.

Al-Omari, K., & Abuata, B. (2014). Arabic light stemmer. International Journal of Computer Science Issues, 11(4), 1–8.

Al-Shammari, E., & Lin, J. (2008). An accuracy-enhanced light stemmer for Arabic text. ACM Transactions on Asian Language Information Processing, 7(2), 1–22.

Al-Shenaifi, A., et al. (2025). An innovative Arabic light stemmer developed using a hybrid approach. International Journal of Electrical and Computer Engineering (IJECE), 15(2), 2356–2363.

Boudiaf, M., et al. (2020). Arabic stemming techniques: A review. International Journal of Advanced Computer Science and Applications, 11(5), 1–10.

Chen, A., & Gey, F. (2002). Building an Arabic stemmer for information retrieval. In Proceedings of TREC, Gaithersburg, MD, USA.

Darwish, K. (2002). Building a shallow Arabic morphological analyzer in one day. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA, USA.

Khoja, S., & Garside, R. (1999). Stemming Arabic text. Computing Department, Lancaster University, Lancaster, UK, Technical Report.

Larkey, L. S., & Connell, M. E. (2001). Arabic information retrieval at UMass in TREC-10. In Proceedings of TREC.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In Proceedings of ACM SIGIR, 275–282.

Taghva, K., Elkhoury, R., & Coombs, J. (2005). Arabic stemming without a dictionary. In Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2005), Las Vegas, NV, USA.

Yaseen, S., & Al-Khulaidi, A. (2025). A new stemming model based on data generation to enhance Arabic information retrieval. Journal of King Saud University – Computer and Information Sciences.