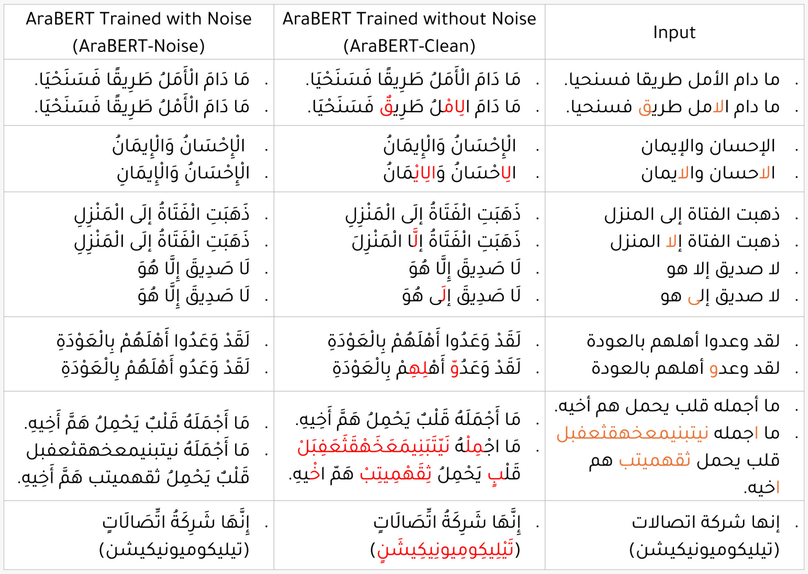
EINPRESSWIRE

New AI model enables native speakers and foreign learners to read undiacritized Arabic texts with greater fluency

Scientists report that they have developed a new machine-learning system designed to overcome challenges encountered in the diacritization of Arabic texts.

SHARJAH, EMIRATE OF SHARJAH, UNITED ARAB EMIRATES, February 4, 2026 /EINPresswire.com/ -- By Ifath Arwah, University of Sharjah

Reading an Arabic newspaper, a book, or academic prose fluently, whether digital or in print, remains challenging for many native speakers, let alone learners of Arabic as a foreign language.



AraBERT Trained with Noise (AraBERT-Noise)	AraBERT Trained without Noise (AraBERT-Clean)	Input
مَا دَامَ الْأَمَلُ طَرِيقًا فَسَنَحْيَا. . مَا دَامَ الْأَمَلُ طَرِيقًا فَسَنَحْيَا. .	مَا دَامَ الْأَمَلُ طَرِيقًا فَسَنَحْيَا. . مَا دَامَ الِامْلُ طَرِيقٌ فَسَنَحْيَا. .	ما دام الأمل طريقا فسنحيا. . ما دام الامل طريق فسنحيا. .
الْإِحْسَانُ وَالْإِيمَانُ . الْإِحْسَانُ وَالْإِيمَانِ .	الْإِحْسَانُ وَالْإِيمَانُ . الِاحْسَانُ وَالِايْمَانُ .	الإحسان والإيمان . الاحسان والايمان .
ذَهَبَتِ الْفَتَاةُ إِلَى الْمَنْزِلِ . ذَهَبَتِ الْفَتَاةُ إِلَى الْمَنْزِلِ . لَا صَدِيقَ إِلَّا هُوَ . لَا صَدِيقَ إِلَّا هُوَ .	ذَهَبَتِ الْفَتَاةُ إِلَى الْمَنْزِلِ . ذَهَبَتِ الْفَتَاةُ إِلَّا الْمَنْزِلَ . لَا صَدِيقَ إِلَّا هُوَ . لَا صَدِيقَ إِلَى هُوَ .	ذهبت الفتاة إلى المنزل . ذهبت الفتاة إلا المنزل . لا صديق إلا هو . لا صديق إلى هو .
لَقَدْ وَعَدُوا أَهْلَهُمْ بِالْعَوْدَةِ . لَقَدْ وَعَدُو أَهْلَهُمْ بِالْعَوْدَةِ .	لَقَدْ وَعَدُوا أَهْلَهُمْ بِالْعَوْدَةِ . لَقَدْ وَعَدُوّ أَهْلِهِمْ بِالْعَوْدَةِ .	لقد وعدوا أهلهم بالعودة . لقد وعدو أهلهم بالعودة .
مَا أَجْمَلَهُ قَلْبٌ يَحْمِلُ هَمَّ أَخِيهِ. . مَا أَجْمَلَهُ نيتبنيمعخهقثعفبل قَلْبٌ يَحْمِلُ ثقهميتب هَمَّ أَخِيهِ. .	مَا أَجْمَلَهُ قَلْبٌ يَحْمِلُ هَمَّ أَخِيهِ. . مَا اجْمِلْهُ نَيّتَبَنِيمَعَخَهْقَثَعَفِبَلْ قَلْبٍ يَحْمِلُ ثِقَهْمِيتِبْ هَمّ اخْيه. .	ما أجمله قلب يحمل هم أخيه. . ما اجمله نيتبنيمعخهقثعفبل قلب يحمل ثقهميتب هم اخيه. .
إِنَّهَا شَرِكَةُ اتِّصَالَاتٍ (تيليكوميونيكيشن) .	إِنَّهَا شَرِكَةُ اتِّصَالَاتٍ (تَيْلِيكِومِيونِيكِيشَنٍ) .	إنها شركة اتصالات (تيليكوميونيكيشن) .

he effectiveness of noise incorporation by comparing the performance of AraBERT-Enhanced-Noisy with AraBERT-Enhanced on various examples. The noisy model correctly diacritized sentences with common spelling errors, distinguished between valid and nonsense

The difficulty largely stems from the nature of Arabic writing, which relies heavily on consonants. Without diacritics, which mark short vowels, it becomes extremely hard to achieve accurate pronunciation, proper contextual understanding, and clear meaning.

Now, scientists at the University of Sharjah report that they have developed a new machine-learning system designed to overcome these challenges.
The system mainly targets problems that existing programs face when encountering undiacritized Arabic script, writing that lacks the vowel marks necessary to pronounce words correctly, a process linguists refer to as diacritization.

The presence of diacritics in Arabic is vital not only for how a word is pronounced but also for semantics. A single word can have multiple, entirely different meanings, depending on how it is articulated.

"Diacritization in Arabic is crucial for correct pronunciation, for differentiating words, and for improving text readability. Diacritics, which represent short vowels, are placed above or below letters. Without them, Arabic becomes challenging for non-native speakers, language learners, and even many native speakers," the researchers explain in their study published in the journal Information Processing and Management. (https://doi.org/10.1016/j.ipm.2025.104345)

The study proposes "a framework for developing robust, context-aware Arabic diacritization models. The methodology included dataset enhancement, noise injection, context-aware training, and the development of SukounBERT.v2 using a diverse corpus," they note.

New leap in Arabic diacritization research

Linguists employ eight diacritics in Arabic orthography to produce distinct vocalizations of the same word to clarify its meaning and context. Classical Arabic texts typically go without diacritical marks, and the same is true for most standard Arabic materials as well as scripts representing the language's diverse dialects.

While recent years have seen considerable advances in Arabic diacritization research, "existing models struggle to generalize across the diverse forms of Arabic and perform poorly in noisy, error-prone environments," the authors note. Their work aims to remove current impediments by allowing existing AI models to furnish accurate vowel marks that support fluent, unambiguous reading.

According to the researchers, "These limitations may be tied to problems in training data and, more critically, to insufficient contextual understanding. To address these gaps, we present SukounBERT.v2, a BERT-based Arabic diacritization system that is built using a multi-phase approach."

SukounBERT is an AI-driven model designed to restore diacritics to Arabic writing. The authors' newly introduced SukounBERT.v2 builds on earlier models. It is specifically constructed to address earlier versions' shortcomings, such as poor generalization across different Arabic varieties and reduced performance in noisy or error-prone environments.

"We refine the Arabic Diacritization (AD) dataset by correcting spelling mistakes, introducing a line-splitting mechanism, and by injecting various forms of noise into the dataset, such as spelling errors, transliterated non-Arabic words, and nonsense tokens," the authors note. They add, "Furthermore, we develop a context-aware training dataset that incorporates explicit diacritic markings and the diacritic naming of classical grammar treatises."

The Sukoun Corpus and diacritization research

The authors' method draws on the Sukoun Corpus, a large-scale, diverse dataset comprising over 5.2 million lines and 71 million tokens from a variety of Arabic written sources, including dictionaries, poetry, and purpose-crafted contextual sentences.

They further augment their corpus with a token-level mapping dictionary that enables minimal or micro-diacritization without sacrificing accuracy. “This is a previously unreported feature in Arabic diacritization research. Trained on this enriched dataset, SukounBERT.v2 delivers state-of-the-art performance with over 55% relative reduction in Diacritic Error Rate (DER) and Word Error Rate (WER) compared to leading models.”

According to the authors, their approach benefits both native speakers and learners of Arabic as a foreign language by reducing perceptual noise and avoiding “garden path” effects, a cognitive process that results in misleading linguistic cues that can momentarily lead readers to a false interpretation.

The approach does not recommend restoring excessive diacritics, as nearly every letter of the Arabic alphabet already carries a diacritic. Instead, it adopts the strategy of “minimal” rather than “full” diacritization, offering native speakers and learners of Arabic “essential phonetic cues that enhance word recognition and comprehension, bridging the gap between structured textbook language and authentic, largely unvowelized texts found in newspapers, literature, and everyday media.”

By striking a balance between semantic precision and cognitive efficiency, “minimal diacritization aligns with modern publishing practices and accommodates diverse reader profiles. As the authors emphasize, the approach makes it “an optimal strategy for enhancing real-world reading performance across proficiency levels.”

Revolutionizing modern Arabic diacritization

Research on automating Arabic diacritization has gained momentum as the number of the language’s more than 400 million native speakers and over 100 million people worldwide learning or using it as a second or foreign language increases. Moreover, manual diacritization remains both complex and time-consuming, and although linguists have historically depended on limited but useful rule-based systems to navigate Arabic language intricacies, the method is no longer practical for the massive proliferation of digital texts.

The authors point out that SukounBERT.v2 relies heavily on contextual clues to resolve ambiguities in meaning and pronunciation. A plethora of research shows that the presence of diacritics greatly enhances reading and comprehension skills, enabling readers to access a precise semantic representation of words that are otherwise difficult to infer from undiacritized script.

Describing SukounBERT.v2 as a “state-of-the-art” model, the authors report that it outperforms existing open-source models by a substantial margin. They note that “the implementation of minimal diacritization using a token-level mapping dictionary enhanced the system's practicality by providing accurate yet readable output with only essential diacritics."

Unlike earlier AI-driven models that primarily emphasize accuracy, SukounBERT.v2 "introduces a more comprehensive strategy that enhances robustness, context awareness, and adaptability."

One of the model's most notable innovations is its minimal diacritization approach, "which optimally balances readability and phonetic accuracy, ensuring that only essential diacritics are retained without compromising meaning. Moreover, the inclusion of context-aware training data allows the model to infer grammatical roles more effectively, resolving structural ambiguities in Arabic text."

Despite these advancements, the authors acknowledge limitations, notably the scarcity of diacritized modern standard Arabic datasets, which continues to impede the progress of research in the field.

They conclude that addressing this gap will require "the development of large-scale, open-source MSA datasets to enhance model performance across different Arabic varieties. Furthermore, while SukounBERT.v2 achieves high accuracy, its lack of interpretability remains a challenge, limiting transparency in decision-making."

LEON BARKHO
University Of Sharjah
+971 50 165 4376
email us here
