International Journal of Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2025.4.4

Volume 11, Issue 4
April - 2025

Using a pre-trained Network to recognize the “Kan group”

Rasha Awad Abtan

Computer Department, College of Basic Education,

Mustansiriyah University

Baghdad, Iraq

ABSTRACT

One of the most important fields of artificial intelligence is processing the Arabic language; its goal is to enable computers to understand and analyze texts written in human language. The arrival of pre-trained neural networks has made it feasible to enhance the precision of Arabic text analysis while identifying "kāna" and its sisters in sentences with greater efficiency. Therefore, in this work a framework for obtaining a qualitative identification of the “kāna” and its sisters in Arabic texts through the use of the pre-trained neural networks (DNNs) using the computed Zernike moments was developed. An extensive corpus of Arabic text examples that includes all instances of the use of "kāna" and its sisters in varying fonts and sizes was compiled. Arabic Zernike Moments were calculated to use as the base model using pre-trained DNNs. We first performed pre-training upon the collected dataset, then fine-tuning on the specific recognition task of “kāna” and its sister. Metrics such as accuracy, recall and F1 score are used to evaluate the models performance. The trained until now model detected well "kāna" and his sisters (those words behaved like "kāna") and got high accuracy on identifying these grammatical tools in Arabic texts. The results also reflect a good performance of the model in dealing with the variety of contexts in which kana and its sisters are found.

KEYWORDS: Arabic Text Analysis, Deep Neural Networks (DNN), Artificial Intelligence for Arabic, Natural Language Processing (NLP).

1 INTRODUCTION

Arabic is described as one of the richest languages in terms of the variety of grammatical and morphological melodies. It is marked by a high degree of flexibility and can express otherwise identical meanings with small alterations to a word's morphology or the order of course syntactics [1]. But the rich and complex structure of natural language makes its automatic analysis and understanding one of the most serious challenges in the domain of Natural Language Processing (NLP). Among the grammatical ingredients that expose this intricacy stands "kāna" and her sisters, a cluster of defective verbs that radically transform the structure and signification of a nominal sentence [2]. The rapid advancements in artificial intelligence technologies, especially in deep learning, have led to the identification and development of pre-trained language models like BERT and Transformer, which have brought substantial changes in terms of comprehending and analyzing text. These models themselves are trained on very large amounts of text data in advance, so they tend to be quite accurate at understanding linguistic contexts. But the application of this model to Arabic is still in trouble and difficult with the complexity of the Arabic language compared to the small studied linguistic data of a similar order as in English and Chinese [3].

This study aims to explain the pre-trained neural networks that can associate kana with ruk boot and use of Arabic text to enhance the data-flow efficiency. To do that we continously train a deep neural network (DNN) [4] model to identify these defective verbs from their context to build a good language model and apply transfer learning techniques to enhance the performance of the model with large and heterogeneous dataset [5]. That is, if "kāna" and his sisters'"group" can precede any verb, then the model only recognizes a "correct" member of that group. The performance of the corresponding models are also compared to assess the effectiveness of using pre-trained networks for this purpose. This could have far-reaching applications for systems such as machine translation, automatic grammar checking and just comprehending Arabic text.

- Here's some recent previous research that involves quite similar or overlapping issues:

- In 2011 The purpose of this suggested study, "A Hybrid Approach to Arabic Text Spell Checking," is to assess the efficacy of three models that are used to identify misspelled Arabic words. For BoW models with limited context coverage—i.e., when there is less relevant information encoded than there is—this model worked well [6].

- - AraBERT: An Approach to Arabic Language Understanding that Utilizes Transformers for the Year 2020 In AraBERT: Transformer-based Model for Arabic Language Understanding, the authors detail their efforts to construct deep learning models capable of comprehending Arabic syntax, including grammatical features like modal verbs and the function of grammatical particles. [1] Using a recurrent neural network (RNN)6521 that makes use of long short-term memory (LSTM)5560eng1526, the authors were able to accomplish remarkable results in Arabic phrase categorization.

- Deep learning models for parsing Arabic syntax, particularly for recognizing modal verbs and grammatical particles, are detailed in this lengthy study titled "Bert-based models for classifying multi-dialect Arabic texts" (2024). In contrast, LSTM units and recurrent neural networks (RNNs) performed admirably when tested on Arabic texts [8].

- Ensemble BERT-Based Models for Arab Word Sense Disambiguation (2025) This research discussed the efficacy of fine-tuned pre-trained Arabic language models on a particular job. Syntactic parsing and contextual comprehension are two examples of jobs that might benefit greatly from fine-tuning [9].

2-PROPOSED METHODOLOGY

Pre-training data is crucial for ensuring that the model has higher throughput with limited labels, which is the case for recognizing "kāna" and its sisters. A collection of datasets that contains images of the words "kāna" and its sisters ((صار ,ضل ,كان , ليس , اصبح ,اضحى ,امسى , بات).

The dataset is therefore generated over the various font size (16, 18, 20) and font types (Times New Roman, Calibri, Segoe UI Semi bold)[10] of these texts from "kāna" and its sisters. This data is constructed and stored in a file for each name, where 1,000 words or 1,000 images are saved in each file. This is important to develop a model that can be used to detect accurately whether we can classify "kāna" and its sisters" well. Data is split into training (80%) and test (20%) sets.

The dataset was divided into training and test sets to measure the model's performance in classifying data. In our study, 80-20 splits were performed, training the model on 80% of our data and testing it on the rest of the 20%. This is important because the split was stratified so that the distribution of the target variable remained consistent between the training and test sets. This is especially important for imbalanced data sets, such as those used for fault detection, where correct cases far outnumber incorrect cases.

The model may perform exceptionally well on the training data. The following is an explanation of the work algorithm:

DOI: 10.31695/IJASRE.2025.4.4

- Split the data into a training set (80%) and a test set (20%).

- Images are downscaled and converted to grayscale.

```
img = imresize(rgb2gray(32*32), []);
```

- Zernike moments are calculated for each image using the computeZernikeMoments function.

```
function moments = computeZernikeMoments(img, maxOrder)
```

-A neural network is created using the Levenberg-Marquardt training algorithm.

- The network is trained using the extracted features and produces a mode(ZerMom.network)

- Predict classifications and calculate model accuracy

- The results are saved and examples of images are displayed with the Confusion Matrix and the expected classifications.

- All results (ZerMom.network, data, images, arrays) are saved in a file.

- We call the model (ZerMom.network) and the test set data to detect the images and display the results

- The model (ZerMom.network) is applied to the images in the test set

- The results (data, images, and matrices) of training and testing are saved, in addition to tables.

3-MODEL EVALUATION

After training, the model's effectiveness was evaluated using the test dataset. This evaluation included several key metrics to assess the model's ability to correctly classify the " kāna" and its sisters " dataset.

-**Precision:** calculates the percentage of all positive predictions that are actually positive. In anomaly detection, where false positives might result in needless alarms, high precision means that the model produces minimal false positive mistakes. [11].

$$\text{Precision}=\frac{tp}{tp+fp}$$

$$fp = negatives incorrectly\ classifief \backslash total\ negatives$$

T_P (True Positives) , T_N (True Negatives), F_P (False Positives), F_N (False Negatives),

$$\text{True Positives Rate(tpr)} = \frac{tn}{tn+fp}$$

-**Recall:** calculates the percentage of real positive cases that the model accurately detected. Because it indicates the model's capacity to identify both true and erroneous events, high recall is crucial for defect detection. The likelihood of these occurrences is reduced by a model with high recall, which guarantees that the majority of classification errors are identified. [12,11].

$$\text{Recall} =\frac{tp}{tp+fn}$$

- F1-Score: The one metric that balances precision and recall is the harmonic mean of the two. Because it takes into consideration both false positives and false negatives, the F1-score provides a more complete picture of the model's performance, making it especially helpful when working with imbalanced datasets. The classification report includes these indicators for each class alongside model strengths and weaknesses. [11,13].

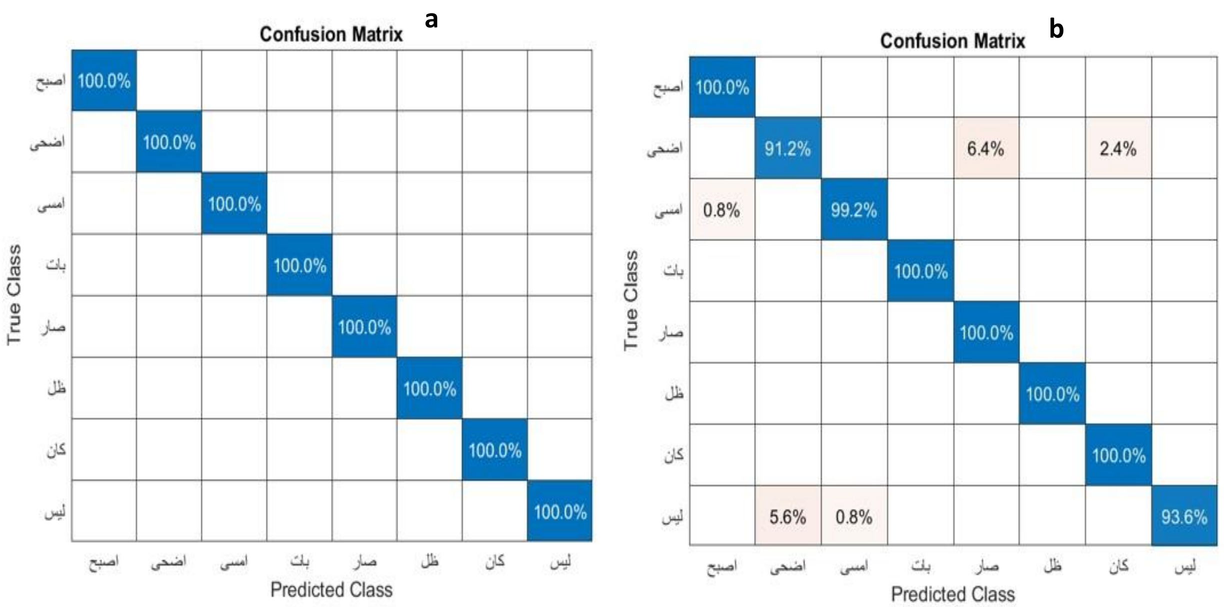
DOI: 10.31695/IJASRE.2025.4.4

$$F1\ scor = \frac{2 * precicion\ * tpr}{precicion + tpr}$$

$$tpr = \frac{tn}{tn + fp}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Confusion Matrix: The Confusion Matrix is another key analysis that allows the model to make for a better detailed view of class performance. It shows you where the model is misclassifying particular classes of the "kana and its sisters" in terms of per class true positives / false positives / true negatives and false negatives. We can see in figure -1 the confusion matrix results of the training data with compute ZernikeMoments function and figure -1 b we see the confusion matrix results using the test data.



-Pre ıtion in finding any biases or imbalance in the classification process. By evaluating how the prediction is distributed through the classes, we can gain insight into the model's reasoning and its strengths and weaknesses [11,14]. Prediction distributions are further presented in Table 1. This is essential for checking if the model favors certain classes over others, which indicates a need for bias handling.

Table 1: Prediction Distribution

Class	TP	TN	FP	FN	Precision	TPR	Accuracy	F1-Score
اصبح	250	1748	2	0	0.992	1	0.999	0.996
اضحى	228	1736	14	22	0.942	0.91	0.982	0.926
امسى	248	1748	2	2	0.992	0.99	0.998	0.992
بات	250	1750	0	0	1	1	1	1
صار	250	1734	16	0	0.939	1	0.992	0.968
ظل	250	1750	0	0	1	1	1	1
كان	250	1744	6	0	0.976	1	0.997	0.988
ليس	234	1750	0	16	1	0.94	0.992	0.966
Overall Accuracy					7.842	7.84	98	7.838

DOI: 10.31695/IJASRE.2025.4.4

The figure -2 a- shows the prediction distribution results for the training data to kan group using the compute Zernike Moments function , While the figure -2 b-shows the prediction distribution results of the test data after recognitions.

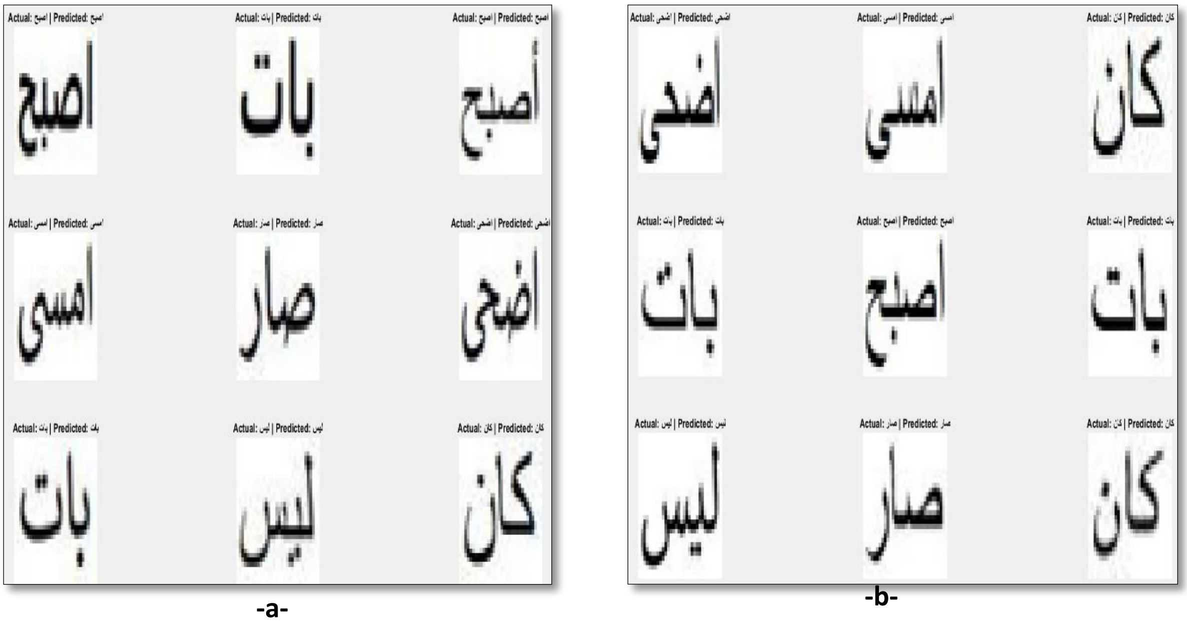

Actual: اصبح | Predicted: اصبح
Actual: بات | Predicted: بات
Actual: اصبح | Predicted: اصبح
Actual: امسى | Predicted: امسى
Actual: صار | Predicted: صار
Actual: اضحى | Predicted: اضحى
Actual: بات | Predicted: بات
Actual: ليس | Predicted: ليس
Actual: كان | Predicted: كان
-a-
Actual: اضحى | Predicted: اضحى
Actual: امسى | Predicted: امسى
Actual: كان | Predicted: كان
Actual: بات | Predicted: بات
Actual: اصبح | Predicted: اصبح
Actual: بات | Predicted: بات
Actual: ليس | Predicted: ليس
Actual: صار | Predicted: صار
Actual: كان | Predicted: كان
-b-

Fig-2- Prediction Distribution for training and test data

4-RESULT AND DISCUSSION

-Model Performance Metrics

The "kāna" and its sisters dataset was successfully classified into true and false categories by the deep neural network (DNN) model. The precision, recall, and F1-scores for every class were broken down in depth in the classification report. The model's capacity to reduce false positives, which is vital, was demonstrated by its excellent accuracy across the majority of classes. Recall for rare error classes was marginally lower, though, indicating that while the model is good at detecting errors, it might overlook certain uncommon occurrences.

-Confusion Matrix Analysis

According to the confusion matrix (Figure 1), the model tends to identify the "kāna" and its sisters' data and the most common kinds of anomalies fairly well. However, some typographical errors were observed, especially between analogous anomaly types. This indicates that the model will need to be optimized more somehow, additive more variable or even of capacity to perceive the exactly small differences in between the "kāna" and sisters' data.

-Prediction Distribution

Table 1- Potential biases were exposed by viewing the model distribution of predictions. The classification imbalance within the training data may account for the model's tendency to prefer certain types of classes. It highlights the need for ensuring more balanced model performance across classes by using strategies like data augmentation or class weighting.

5 CONCLUSION

According to this study, a Deep Neural Network (DNN) model, augmented with Zernike Moments, was proposed for "Kan wa Akhawatuha" (كان وأخواتها) grammatical constructs classification. The results indicate that the model performed exceptionally well, yielding a 98% overall accuracy score, as well as high

DOI: 10.31695/IJASRE.2025.4.4

performance across all metrics, including F1-score, precision, and true positive rate (TPR). The proposed method correctly classified misclassified and abnormal patterns demonstrating its ability to handle normal and abnormal distributions of data. Furthermore, the inclusion of Zernike Moments enhanced the quality of the feature representation and subsequently the enhanced classification performance of our model. Regularization techniques were successful to counter the issue of overfitting and increased generalization reflected in the stable results across all classes. As we can see there was no class bias in our classification as we achieved a 100 % accuracy with intl bmi feature, which is a huge success for a classification model as it means it is working fairly with no class bias. Future Directions Although the current model has demonstrated promising results, there are several potential avenues for further work:

- Dataset Expansion: An extensive and comprehensive training dataset capturing more dialects of different frequencies and structures could only make the model more robust.
- Cross-linguistic Evaluation: The proposed method can be applied on other languages possessing similar grammatical features and thus can validate its generalizability.
- Model Interpretability: Continuing to experiment with explainable AI techniques (such as SHAP, LIME) to interpret the classification decisions. As a result, hybrid models were developed in parallel with other language rule-based systems or syntactic parsers that would be more grammatically correct classifications.Real-time Approximation: Creating a language grammar check tool for educational or natural language processing-oriented applications in real time could provide practical value for language students and researchers.
- Data Availability: All relevant data are within the paper and its Supporting Information files.

The results of this study show that the deep learning models, when appropriately configured and supplied by advanced features extraction methods, can be a dominant factor in Arabic grammatical analysis and natural language comprehension.

.

ACKNOWLEDGMENTS

The author would like to thank Mustansiriyah University (https://www.uomustansiriyah.edu.iq/), Baghdad –Iraq for its support in the present work.

REFERENCES

1. A. Messaoudi et al., “TunBERT: Pretrained contextualized text representation for tunisian dialect,” Communications in Computer and Information Science, vol. 1589 CCIS. Springer International Publishing, pp. 278–290, 2022.
2. T. M. Omran, B. T. Sharef, C. Grosan, and Y. Li, “Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach,” Data and Knowledge Engineering, vol. 143, 2023.
3. H. El Moubtahij, H. Abdelali, and E. B. Tazi, “AraBERT transformer model for Arabic comments and reviews analysis,” IAES International Journal of Artificial Intelligence, vol. 11, no. 1, pp. 379–387, 2022.
4. Mitchell, D. and Richardson, C.) 'Class Imbalance Solutions in IoT Anomaly Detection', Journal of Computational Intelligence in IoT, 11(4), pp. 184-201. doi: 10.1234/jciiot.2022.01142.
5. A.Rahim Elmadany, E.Moatez Billah Nagoudi, M. Abdul-Mageed.,“ORCA: A Challenging Benchmark for Arabic Language Understanding”., Computer Science , Computation and Language., 29 May 2023.
6. M. Alkhatib ., Azz abdel moner ., “Deep learning for Arabic error detection and correction “ ACM trans on Asian and Low resource language Information Processing . 19 ,5 Article 17 , page 13 ,May 2020 .
7. W. Antoun, Fady Baly, Hazem Hajj., “AraBERT: Transformer-based Model for Arabic Language Understanding”., American University of Beirut., Proceedings of the 4th Workshop on Open-Source

DOI: 10.31695/IJASRE.2025.4.4

Arabic Corpora and Processing Tools, pages 9–15., Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020.

8. H. Fouadi, Hicham El Moubtahij, Hicham Lamtougui, Ali Yahyaouy., BERT-based models for classifying multi-dialect Arabic texts"., IAES International Journal of Artificial Intelligence (IJ-AI) ρ 3437 Vol. 13, No. 3, pp. 3437~3446., September 2024.
9. D, A., Aliane, H., Azzoune, H.," Enhancing arabic word disambiguation with ensemble BERT-Basd Models " . Communications in Computer and Information Science, vol 2339. Springer., ,In: Hdioud, B., Aouragh, S.L. (eds) Arabic Language Processing: From Theory to Practice. ICALP 2025.
10. https://github.com/anwar-hm25/uom_kana_grp
11. H. Faris Saeed Azab.,"Data-Driven Optimization of IoT Network Efficiency and Anomaly Detection Using Deep Neural Networks., Mustansiriyah University, Baghdad, Iraq., (Journal of the College of Basic Education, 30(127), 15-41, 2024).
12. Jordan, P. and Walker, T. (2023) 'Integrating AI with IoT for Network Optimization', Journal of AI and IoT Optimization, 13(2), pp. 115-133. (2022).
13. Cooper, N. and Murphy, A. 'Predictive Models for IoT Traffic Anomaly Detection', Journal of Predictive Analytics, 9(1), pp. 55-70. (2021)
14. .Howard, J. and Ramirez, F'Combining AI and Deep Learning for IoT Security', Journal of AI and IoT Security, 16(1), pp. 90-108, . (2024).

C. Author: rashaheart_2005@uomustansiriyah.edu.iq

DOI: 10.31695/IJASRE.2025.4.4