# Evaluating Transformer-Based Pretrained Models for Classical Arabic Named Entity Recognition

Mariam Muhammed, Shahira Azab
Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

***Abstract*—This study presents a comprehensive comparative evaluation of transformer-based pretrained language models for Named Entity Recognition (NER) in Classical Arabic, an underexplored linguistic variety characterized by rich morphology, orthographic ambiguity, and the absence of diacritics. The main objective of this work is to identify the most effective transformer model for Classical Arabic NER and to analyze the linguistic factors influencing model performance. Using the CANERCorpus, which contains Hadith texts annotated with twenty fine-grained entity types, ten transformer-based models were fine-tuned and evaluated under consistent experimental settings. The study benchmarks models such as AraBERT, ArBERT, and multiple CAMeLBERT variants, comparing their precision, recall, and F1-scores. The results demonstrate that all models achieve strong performance (F1 > 96%), while CAMeL-CA-NER attains the highest score (F1 = 97.78%), confirming the advantage of domain-specific pretraining on Classical Arabic data. Error analysis further reveals that domain-adapted models better handle ambiguous entities and religious terminology. A comparative analysis with traditional and non-transformer approaches, including rule-based and BERT-CRF models from previous studies, shows that CAMeL-CA-NER surpasses earlier methods by more than 3% in F1-score, highlighting its superior capability in handling Classical Arabic text. However, this study is limited to the CANERCorpus, which primarily consists of Hadith texts; results may vary for other Classical Arabic genres or domains. These findings provide a valuable benchmark for future research and demonstrate the adaptability of modern NLP architectures to linguistically complex, low-resource domains.**

***Keywords*—Classical Arabic; Named Entity Recognition; transformer models; pretrained models; CANERCorpus**

## I. Introduction

Named Entity Recognition (NER) is an important task in natural language processing (NLP) that aims to identify and categorize entities such as persons, locations, and organizations. Since its introduction in the Message Understanding Conferences (MUC) in the 1990s [1], NER has become essential for various NLP applications, including information retrieval, machine translation, question answering, and text summarization [2], [3], [4].

Extensive research has been conducted in English and other widely spoken languages, including Spanish and Chinese. NER systems are generally built using three approaches: rule-based, machine learning, and hybrid methods. Rule-based systems rely on handcrafted linguistic rules and dictionaries [1], [5], [6], but they are limited by language-specific expertise and lack flexibility [7], [8]. Deep learning, a subset of machine learning, introduced neural architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which learn representations at both the character and word levels [9].

The rapid growth of Arabic content on the internet has created a strong need for accurate NLP tools for Arabic. Arabic serves as the official language in the Arab World, covering 22 countries [2]. It is a Semitic language with rich vocabulary, complex morphology, and challenging syntax. It exists in three forms, including Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA). CA represents the original Arabic language from the seventh to the early eleventh century CE [10]. MSA is used in formal contexts such as books, formal communications, and news articles, while DA is spoken in everyday life. Although CA is less used in modern contexts, it remains highly important as the language of the Quran (the holy book of Islam), the Hadith (the sayings of Prophet Muhammad), and the Islamic heritage.

Arabic NER faces unique challenges compared to other languages due to its morphological richness, allowing suffixes and prefixes to function as conjunctions, prepositions, and pronouns, as in the word "فأسقيناكموه" (faskenakomoh), meaning "So we gave you water to drink". Unlike languages that rely on capitalization to identify proper nouns, Arabic does not use capitalization. Moreover, Arabic utilizes short vowels (diacritics) to disambiguate word sense, but it is usually written without them, leading to ambiguity in word meaning. For example, the word "درس" can mean "دَرْس" (lesson) as a noun or "دَرَسَ" (has studied) as a verb [6]. Some examples of ambiguous words in Arabic text are listed in Appendix A.

Classical Arabic, in particular, poses additional linguistic challenges beyond those of Modern Standard Arabic. In addition to its rich morphology, absence of diacritics, and orthographic ambiguity, Classical Arabic exhibits archaic vocabulary, complex syntactic structures, and context-dependent expressions that differ significantly from Modern Standard Arabic usage. These characteristics increase word-level ambiguity and make NER in Classical Arabic notably more difficult than in contemporary Arabic varieties.

Furthermore, the scarcity of datasets and other reliable resources for Arabic increases the challenges of Arabic NER [2], [11].

Recent advances in deep learning and pretrained models have significantly improved Arabic NER. Contextualized embeddings such as ELMo, BERT, and GPT, along with sequence-to-sequence and convolutional neural network models, have enhanced system performance by capturing the

complex linguistic features of Arabic and learning effectively from annotated data. Transformer-based pretrained language models have transformed NLP and had a major impact on Arabic NER. Trained on massive text corpora, they can identify complex linguistic patterns. When fine-tuned on Arabic NER datasets, these models achieve state-of-the-art results with less data and lower computational cost.

The main contribution of this study is a comparative evaluation of transformer-based pretrained models for Classical Arabic NER using the CANERCorpus dataset. It investigates their predictive capabilities and performance on Hadith texts. This study provides a systematic benchmarking of transformer-based pretrained models for Classical Arabic NER. Unlike previous studies that mainly target Modern Standard Arabic, this work focuses on Classical Arabic, analyzing model behavior in a linguistically rich and underexplored domain. This benchmarking effort serves as a foundation for future research aiming to develop specialized models and resources for Classical Arabic.

The remainder of this study is organized as follows: Section II reviews related work on Arabic NER, covering rule-based, traditional machine learning, and deep learning approaches, and highlights gaps in previous studies. Section III presents the pretrained transformer models used in this study, including their architectures, training data, and fine-tuning procedures for NER. Section IV describes the dataset, the experimental setup, and evaluation measures. Section V reports the experimental results, providing quantitative performance comparisons across models. Section VI offers a discussion of the results, analyzing model behavior and domain-specific challenges. Section VII presents a qualitative error analysis, highlighting common mistakes and ambiguous cases. Section VIII provides a comparative evaluation with previous Arabic NER approaches, emphasizing the benefits of domain-specific pretraining. Section IX discusses the limitations of the study, including the dataset and methodological constraints. Finally, Section X concludes the study and outlines potential directions for future research. Supplementary material, such as lists of ambiguous words and extended experiment details, is provided in Appendix A.

## II. Related Work

Many approaches have been proposed to perform Arabic NER. Qu et al. (2024) classified them into four main paradigms: rule-based methods, machine learning, deep learning, and pretrained language models [10].

Before the rise of machine learning, the rule-based approach was the prime choice utilized in NER tasks. These systems depended on manually written linguistic rules and dictionaries. Examples of features commonly used include morphological analyzers [8], [12], [13], lexical triggers [14], regular expressions and gazetteers [15], as well as transliteration techniques [16]. Rule-based systems usually perform well in narrow domains but fail to generalize across domains. They also require heavy manual effort, which makes them less scalable [1].

With the advancement of statistical approaches, machine learning (ML) became widely adopted. Researchers explored several ML techniques, including conditional random fields (CRF) [17], [18], support vector machines (SVM) [19], [20], and meta-classifiers [21], [22], [23]. These approaches combined features such as lexical, contextual, morphological, gazetteer, and part-of-speech tags to improve performance. Although machine learning methods achieve satisfactory performance, they struggle to learn complex and high-level features from data when using linear models such as log-linear HMM or linear chain CRF. However, these approaches often underperform on Classical Arabic due to the language's rich morphology, absence of diacritics, and the scarcity of domain-specific datasets.

To address this limitation, deep learning methods have emerged as a solution for automatically discovering hidden features. The advancement of Arabic NER through deep-learning methods can be analyzed from three perspectives: i) input representations such as word-level embeddings [22], character-level embeddings [24], and additional features ; ii) context encoders, such as CNN [25], RNN [24], [26], or multi-attention networks [26] that capture dependencies between tokens; and iii) label decoders that assign the correct entity labels [25], [27]. While these methods reduce reliance on manual feature engineering, they still struggle to disambiguate entities in Classical Arabic texts, especially in religious and historical domains.

In recent years, pretrained language models (PLMs) have transformed Arabic NER. By learning contextual representations from massive corpora, models such as BERT [27], AraBERT [28], ARBERT and MARBERT [29], ArabicBERT [30], CAMelBERT [31], and Multilingual BERT [32] achieved state-of-the-art performance in many NLP tasks, including NER [33]. Despite their success, most studies focus on Modern Standard Arabic or social media text. Classical Arabic, particularly in Hadith texts, poses unique challenges not fully addressed by existing models.

Our study builds on this prior work by applying transformer-based pretrained models specifically to Classical Arabic NER using the CANERCorpus. This approach leverages domain-specific pretraining to improve recognition of ambiguous words, religious entities, and morphologically complex tokens, addressing gaps left by previous approaches.

## III. Pretrained Models

We conducted training on the dataset using multiple pretrained language models. Table I provides an overview of the used models and their training data.

- AraBERT: Developed by Antoun et al. [28], AraBERT is a BERT-based model specialized for Arabic. It has several versions (v0.1/v1 and v0.2/ v2). Versions v0.1 and v1 were trained on 23GB of text, while v0.2 and v2 used a larger 77GB corpus. In this study, we employed AraBERT (v2).
- ARBERT/MARBERT: Introduced by Abdul-Mageed et al. [29], these models are based on the BERT architecture. ARBERT was trained on 66GB of news text, while MARBERT was trained on a larger 128GB

dataset that included 50% tweets. MARBERT is particularly strong for social media Arabic.

- ARBERT/MARBERT: Introduced by Abdul-Mageed et al. [29], these models are based on the BERT architecture. ARBERT was trained on 66GB of news text, while MARBERT was trained on a larger 128GB dataset that included 50% tweets. MARBERT is particularly strong for social media Arabic.
- ArabicBERT: Developed by Safaya et al. [30], this model was trained on 95GB of Arabic text, mostly from the OSCAR corpus. It captures broad Arabic patterns and structures.
- GigaBERT: Developed by Lan et al. [34], GigaBERT is a bilingual model trained on English and Arabic. It was trained on ACE2005 data, including broadcast news and newsgroups.
- QARiB: Provided by Abdelali et al. [35], QARiB was trained on a huge dataset consisting of around 420 million tweets gathered from Twitter and 180 million text sentences gathered from Arabic GigaWord, Abulkhair, and OPUS corpora.
- CAMeLBERT Models: Developed by Inoue et al. [31], CAMeLBERT has three variants: CAMeLBERT-MSA-NER, CAMeLBERT-CA-NER, and CAMeLBERT-MIX-NER. They are fine-tuned for NER tasks on MSA, CA, and a mix of MSA/CA/DA texts.
- Multilingual BERT: Proposed by Devlin et al. [27], this model supports 104 languages, including Arabic. Although trained on Wikipedia across languages rather than Arabic-specific data, it still provides useful performance for Arabic NER.

TABLE I. OVERVIEW OF THE PRETRAINED MODELS

| Model Name | Model Information | | |
|---|---|---|---|
| | ***Authors*** | ***Training Data*** | ***Arabic Text Source*** |
| AraBERT (v2) | Antoun et al. [28] | 77 GB | OSCAR, Arabic Wikipedia, OSIAN, Arabic Corpus |
| ARBERT | Abdul-Mageed et al. [29] | 66 GB | News articles |
| MAR-BERT | Abdul-Mageed et al. [29] | 128 GB | Tweets, News |
| Arabic-BERT | Safaya et al.[30] | 95 GB | OSCAR corpus |
| Giga-Bert: | Lan et al. [34] | ACE-2005 | Broadcast news, Newsgroups |
| QARIB | Abdelali et al. [35] | 420M tweets + 180M sentences | Twitter API, GigaWord, Abulkhair, OPUS |
| CAMeLBERT-MSA | Inoue et al. [31] | 107 GB | Gigaword, Wikipedia, OSCAR, OSIAN |
| CAMeLBERT-CA | Inoue et al. [31] | 6 GB | OpenITI corpus (v1.2) |
| CAMeLBERT-MIX | Inoue et a. Mach | 167 GB | Various Sourses Combined (MSA, CA, DA) |
| Multi-lingual BERT | Devlin et al. [27] | ~16 GB total (all languages), Arabic ≈ 3 GB | Wikipedia (104 languages) |

## IV. EXPERIMENTS

In this section, we describe the dataset, experimental setup, and evaluation metrics used in this study.

### A. Dataset

Table II presents a list of available annotated datasets for Arabic NER. In this study, we employed the "CANERCorpus", which will be further detailed in this section.

TABLE II. LIST OF ANNOTATED DATASETS FOR ARABIC NER

| Corpus | Year | Text Source | Tags | words | Availability |
|---|---|---|---|---|---|
| ANERcorp[1] | 2007 | Website, news, magazines | 4 | 150 k | Free |
| ACE 2003[2] | 2004 | Broadcast News, newswire genres | 7 | 42 k | Required fees |
| ACE 2004[3] | 2004 | Broadcast News, newswire genres | 7 | 151 k | Required fees |
| ACE 2005[4] | 2005 | Transcripts, news | 7 | ~300 k | Required fees |
| ACE 2007[5] | 2007 | Transcripts, news | 7 | ~300 k | Required fees |
| REFLEX[6] | 2009 | Reuters news | 4 | 22.5 k | Required fees |
| AQMAR[7] | 2012 | Arabic Wikipedia | 7 | 1 M | Free |
| OntoNotes 5.0[8] | 2013 | News | 18 | 300 k | Required fees |
| WDC[9] | 2014 | Wikipedia | 4 | 6 M | Free |
| **CANER[10]** | **2018** | **Religion** | **20** | **258 k** | **Free** |
| Wojood[11] | 2022 | Wikipedia | 21 | 550 K | Under request |

We used the CANERCorpus [36], a Classical Arabic NER dataset annotated by experts. It contains 7,000 Hadiths from *Sahih Al-Bukhari*. The corpus includes 20 entity classes such as person, location, organization, date, book, and others. In total, the dataset has about 258K words, with 72K entities (≈23%). Fig. 1 and Table III show word count for each Named Entity.

[1] URL: https://camel.abudhabi.nyu.edu/anercorp/
[2] URL: https://catalog.ldc.upenn.edu/LDC2004T09
[3] URL: https://catalog.ldc.upenn.edu/LDC2005T09
[4] URL: https://catalog.ldc.upenn.edu/LDC2006T06
[5] URL: https://catalog.ldc.upenn.edu/LDC2014T18
[6] URL: https://catalog.ldc.upenn.edu/LDC2009T11
[7] URL: http://www.cs.cmu.edu/~ark/ArabicNER/
[8] URL: https://catalog.ldc.upenn.edu/LDC2013T19
[9] URL: https://github.com/Maha-J-Althobaiti/Arabic_NER_Wiki-Corpus
[10] URL: https://github.com/RamziSalah/Classical-Arabic-Named-Entity Recognition-Corpus
[11] URL: https://ontology.birzeit.edu/Wojood/

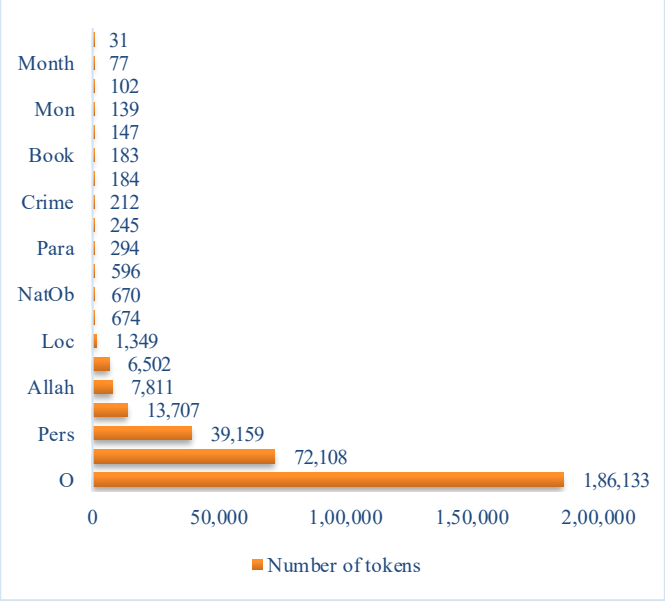


Fig. 1. Distribution of Named Entity tokens in CANERCorpus.

TABLE III. WORD COUNT FOR EACH NAMED ENTITY CLASS

| Named Entity tag | Number of tokens |
|---|---|
| Allah | 7,811 |
| Prophet (Pro.) | 6,502 |
| Pers | 39,159 |
| Num | 13,707 |
| Loc | 1,349 |
| Clan | 674 |
| NatOb | 670 |
| Crime | 212 |
| Date | 596 |
| Para | 294 |
| Hell | 245 |
| Rlig | 184 |
| Book | 183 |
| Means | 147 |
| Mon | 139 |
| Time | 102 |
| Month | 77 |
| Day | 31 |
| Sect | 17 |
| Org | 9 |
| Other Words (O) | 186,133 |
| Named Entity | 72,108 |
| Total | 258,241 |

For our experiments, we focused on person, God, prophet, location, clan, date, natural object, and other entities. The "number" entities were excluded since most of them were page numbers and did not add value to the task.

The CANERCorpus dataset did not provide clear sentence boundaries. We estimated the average sentence length in Sahih Al-Bukhari (38 words) and used this to split the text into sentences as in [36]. To avoid splitting related entities, the last token of each segment was labeled as "other" (O). We also removed punctuation, special characters, and diacritics. The dataset was then divided into training, validation, and test subsets. We randomly split the corpus into 80% training, 10% validation, and 10% testing. Table IV presents the distribution of named entities across each subset.

TABLE IV. THE DISTRIBUTION OF NAMED ENTITIES ACROSS THE TRAINING, TESTING, AND VALIDATION DATASETS

| Named Entity | Training | Testing | Validation | Total |
|---|---|---|---|---|
| Pers | 31,362 | 3,754 | 4,043 | 39,159 |
| Allah | 6,269 | 765 | 777 | 7,811 |
| Prophet | 5,162 | 669 | 671 | 6,502 |
| Loc | 1,087 | 127 | 135 | 1,349 |
| Clan | 601 | 103 | 63 | 767 |
| NatOb | 550 | 61 | 59 | 670 |
| Date | 491 | 56 | 49 | 596 |
| Other | 158,119 | 20,288 | 19,296 | 197,703 |
| Total | 203,641 | 25,823 | 25,093 | 254,557 |

## B. Experimental Setup

The experiments were conducted on the Google Colab platform12 using a Tesla T4 GPU. Models were trained on the training dataset, with hyperparameters tuned on the validation dataset. Table V summarizes the parameters.

TABLE V. HYPERPARAMETER VALUES FOR THE MODELS

| Parameter | Value |
|---|---|
| Learning Rate (LR) | 1e-5 |
| Maximum Length Size | 256 |
| Batch size | 16 |
| Optimizer | Adam |
| No. of epochs | Varied by model |

## C. Evaluation Measures

To assess the effectiveness of the pretrained models, we employed various metrics, including precision, recall, and F-measure. Eq. (1) outlines precision (P), which measures the correct identification of named entities by the model among all identified entities. Eq. (2) illustrates recall (R), indicating the correct identification of named entities by the model among all entities present in the corpus. Eq. (3) describes the F-measure (F), employed to harmonize the conflicting relationship between precision and recall [9], [37].

$$precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1-Score = \frac{2\ (Precicion*Recall)}{(precsion+Recall)} \quad (3)$$

where, TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

[12] https://colab.research.google.com/

## V. RESULTS

In this section, we present the results of training ten pretrained models on the CANERCorpus dataset. Table VI and Fig. 2 summarize the models' performance in terms of Precision, Recall, and F1-score.

TABLE VI. RESULTS ACHIEVED BY PRETRAINED LANGUAGE MODELS

| Model Name | P | R | F | No. of Epochs |
|---|---|---|---|---|
| AraBERT | 96.11 | 96.94 | 96.53 | 13 |
| ArBERT | 96.89 | **98.09** | 97.47 | 7 |
| MARBERT | 96.51 | 97.44 | 96.97 | 7 |
| GigaBERT | 96.93 | 97.50 | 97.21 | 11 |
| Arabic-BERT | 96.63 | 97.50 | 97.06 | 8 |
| QARIB | 96.31 | 96.70 | 96.50 | 7 |
| Camel-msa-ner | 96.85 | 97.75 | 97.30 | 8 |
| Camel-ca-ner | **97.51** | 98.05 | **97.78** | 9 |
| Camel-mix-ner | 96.94 | 97.84 | 97.39 | 8 |
| Multilingual | 96.11 | 96.94 | 96.53 | 11 |

The bold numbers mark the best results.

All models achieved strong results, with F1 scores ranging from 96.53% to 97.78%. The best performance was obtained by CAMeLBERT-CA-NER with an F1 score of 97.78%. QARiB also performed strongly, and models such as AraBERT, ArabicBERT, ARBERT, and MARBERT achieved high results with F1 scores above 96.5%. In contrast, Multilingual BERT obtained the lowest F1 score (96.53%).

These results demonstrate the robustness of transformer-based models in handling the linguistic challenges of Classical Arabic.

## VI. DISCUSSION

This section provides an analysis and interpretation of the results presented in Section V.

The best performance of CAMeLBERT-CA-NER confirms the advantage of domain-specific pretraining on Classical Arabic texts, since the model was exposed to linguistic and stylistic patterns directly relevant to the Hadith corpus. This pretraining allowed CAMeLBERT-CA-NER to better recognize religious and historical entities (e.g., Prophet names, clans, and locations) and to handle morphological variations typical of Classical Arabic. In contrast, models trained primarily on Modern Standard Arabic or social media data showed slightly higher misclassification rates for such entities due to domain mismatch. CAMeLBERT-MIX-NER also achieved competitive performance, suggesting that multiple Arabic varieties (MSA, CA, and DA) can further improve generalization across heterogeneous texts.

QARiB performed strongly as well, benefiting from its large-scale training on diverse Twitter and formal Arabic corpora. This highlights that large amounts of heterogeneous data, even if not exclusively Classical Arabic, can still capture rich contextual representations transferable to CA NER tasks.

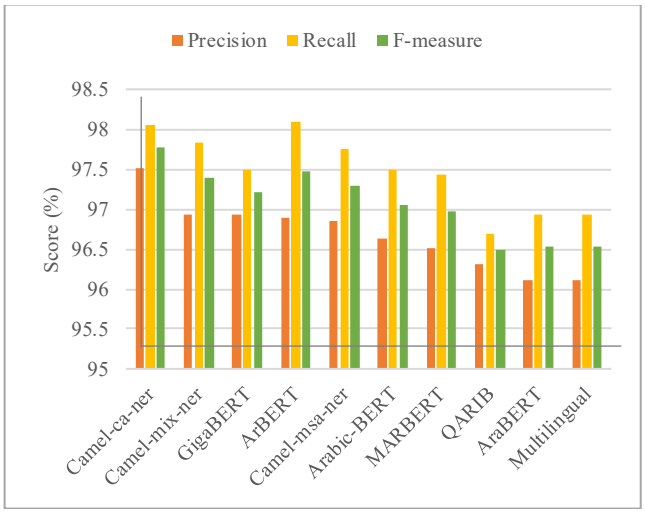


Fig. 2. Performance of pretrained models on CANERCorpus.

Models such as AraBERT, ArabicBERT, ARBERT, and MARBERT also achieved high results. Although primarily trained on MSA and social media text, they were able to generalize effectively to CA. This suggests that while surface differences exist between MSA and CA, pretrained models can leverage their shared morphological and syntactic features.

The weaker performance of Multilingual BERT can be attributed to the relatively small portion of Arabic in its training data (~3 GB from Wikipedia), compared to the much larger corpora used in Arabic-specific models. This result underlines the importance of language-specific and domain-focused training when dealing with morphologically rich languages like Arabic.

Beyond overall scores, a closer inspection of entity-level performance revealed interesting trends. Models trained on Classical Arabic data (e.g., CAMeLBERT-CA-NER) showed superior accuracy in recognizing religious and historical entities such as Prophet names, clans, and locations mentioned in Hadiths. On the other hand, models trained mainly on MSA and tweets occasionally misclassified such entities, reflecting a domain mismatch. However, these models performed comparably well on generic entities like dates, times, and numerical expressions, where the differences between CA and MSA are minimal.

Overall, the results highlight three important insights:

- Domain-specific pretraining matters. CAMeLBERT-CA-NER demonstrates clear advantages when the training domain matches the target text.
- Cross-variety transfer is feasible for MSA and DA-based models still generalize well to CA, especially for common entity types.
- Data size and diversity play a crucial role; larger and more varied corpora (e.g., QARiB, MARBERT) improve robustness, even if they are not domain-specific.

These findings emphasize that while transformer models are inherently powerful, their effectiveness for Classical Arabic NER is maximized when they are trained on domain-relevant corpora. At the same time, the strong performance of general-purpose models suggests promising opportunities for cross-domain transfer and low-resource adaptation in future research.

## VII. Error Analysis

To better understand the behavior of the pretrained models on Classical Arabic NER, we conducted an analysis of detected ambiguities in the test set. We selected the four highest-performing models based on F1-score: ArBERT, CAMeL-CA-NER, CAMeL-MSA-NER, and CAMeL-MIX-NER. The test set contained 24,966 words. Table VII summarizes the number of correctly and incorrectly tagged words for these models.

TABLE VII. Correct and Incorrect Tag Counts for Top-Performing Models

| Model | Correct Tags | Incorrect Tags |
|---|---|---|
| ArBERT | 24,813 | 153 |
| CAMeL-ca-NER | 24,829 | 137 |
| CAMeL-mas-NER | 24,805 | 161 |
| CAMeL-mix-NER | 24,813 | 153 |

Some examples of ambiguous words and their corresponding model predictions are shown in Table VIII.

- Words following "ك" (e.g., كرسول) are sometimes misclassified by ArBERT, which labels them as "O" instead of the correct tag "Prophet".
- Words preceded by "ب", such as (*بنخلة*), may be incorrectly tagged as "LOC" by ArBERT and, in some cases, by CAMeL-CA-NER and CAMeL-MIX-NER, though the correct tag is "O".
- Words preceded by "ف", such as (*فلصفوان*) in the phrase (*وإن لم يرض عمر فلصفوان*), are occasionally misclassified by ArBERT as "O" instead of "Pers".
- Words following (*نو*) in CAMeL-CA-NER are sometimes wrongly labeled as "Pers", for example ( *نو اليدين*).
- Some prophets' names, like (*ابراهيم*) and (*يوسف*), are occasionally tagged as "Pers" instead of "Prophet", and vice versa. This also occurs in ArBERT.
- Certain locations, for instance (*مزدلفة*) and (*جمرة العقبة*), are misclassified by CAMeL-CA-NER.
- Person names following words like (*امرأة*) or (*ابن*) may be incorrectly labeled as "Pers" even when they refer to prophets, as in (*ابن مريم*). This occurs in ArBERT, CAMeL-CA-NER, and CAMeL-MIX-NER.

Additional ambiguities observed specifically in ArBERT and CAMeL-MSA-NER include:

- Words like (*قوله*) followed by (*تعالى*) are misrecognized as "Allah" instead of "O".
- Nouns such as (*أسماء*) are sometimes tagged as "Pers" when they are not proper names, e.g., in ( *أحب أسماء على رضي الله عنه إليه*).
- Words like (*على*) are occasionally misclassified as "Pers", e.g., (*وكانت على بردة*).
- Certain verbs followed by person names can be misclassified as "Pers", for example ( *فنسي عوف ثم عمر بن الخطاب*).
- The word (*الله*) is sometimes labeled as "Allah" when it should be "*لنبي الله*" or "*لرسول الله*".

These errors mainly arise from differences between the models' training data and the Classical Arabic text in CANERCorpus. Models fine-tuned on Modern Standard Arabic (e.g., ArBERT, CAMeL-MSA-NER) show more ambiguities, while models trained on Classical Arabic (CAMeL-CA-NER) or a mixture of varieties (CAMeL-MIX-NER) exhibit fewer errors. The qualitative analysis in Table IX shows that CAMeL-CA-NER effectively handles Classical Arabic morphological patterns and religious expressions, such as prefixes (و, ب, ف) and embedded prophet names, that often confuse models trained solely on Modern Standard Arabic. Its superior handling of ambiguous cases like كرسول reflects its experience to Classical Arabic syntax and vocabulary during pretraining. In contrast, ArBERT and CAMeL-MSA-NER frequently misclassify tokens due to domain mismatch and limited exposure to archaic forms. Occasional over-generalization errors in CAMeL-CA-NER, such as mislabeling اليدين as Person, indicate sensitivity to morphological cues that resemble name patterns. These insights confirm that domain-relevant pretraining enhances contextual disambiguation and improves the recognition of specialized entity types in Classical Arabic texts.

TABLE VIII. Model Predictions for Ambiguous Words

| The ambiguous word | Correct tag | AraBERT | CAMeL-ca-NER | CAMeL-msa-NER | CAMeL-mix-NER | Weighted- ET-ANER |
|---|---|---|---|---|---|---|
| وبمكة | Loc | O<br>X | Loc<br>√ | Loc<br>√ | Loc<br>√ | Loc<br>√ |
| كرسول | Pro. | O<br>X | Pro.<br>√ | Pro.<br>√ | Pro.<br>√ | Pro.<br>√ |
| اليدين | O | O<br>√ | Pers<br>X | O<br>√ | O<br>√ | O<br>√ |
| قال فيوسف نبي الله | Pro. | O<br>X | Pers<br>X | Pro.<br>√ | Pro.<br>√ | Pro.<br>√ |

## VIII. Comparative Evaluation with Previous Arabic NER Approaches

Table IX and Fig. 3 compare the performance of the best model, CAMeL-CA-NER, with previously reported results on the CANERCorpus dataset. Early rule-based approaches achieved modest performance (F1 = 89.5%), reflecting the limitations of handcrafted rules for Classical Arabic. Later BERT-based models reached around 94 to 95% F1, demonstrating the benefits of contextualized embeddings. However, our CAMeL-CA-NER model achieved the highest score (F1 = 97.78%), outperforming earlier systems by more than 3%. This improvement can be attributed to its domain-specific pretraining on Classical Arabic texts (OpenITI corpus[13]), which provides a closer linguistic match to Hadith data. These findings confirm that transformer architectures with domain-relevant pretraining substantially advance the state of Classical Arabic NER.

TABLE IX. Comparison with Previous Works that Used CANERCorpus in Arabic NER Process

| Previous Work | Approaches | Precision | Recall | F1-score |
|---|---|---|---|---|
| [38] | Rule-based approach | 90.2 | 89.3 | 89.5 |
| [39] | BERT | 94.1 | 94.9 | 94.5 |
| | BERT-CRF | 94.0 | 95.4 | 94.7 |
| | BERT-BLSTM-CRF | 93.8 | 95.2 | 94.4 |
| | BERT-GRU-CRF | 94.1 | 95.5 | 94.8 |
| [31] | CAMel-CA-NER | **97.51** | **98.05** | **97.78** |

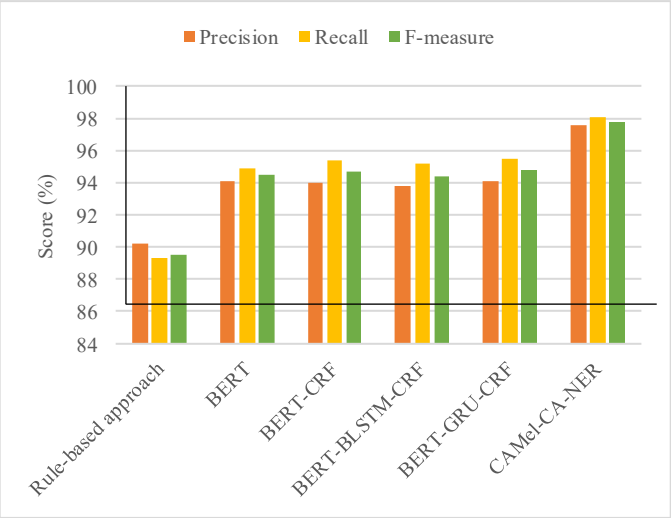


Fig. 3. Comparison of CAMel-CA-NER performance with previous Arabic NER approaches on the CANERCorpus dataset.

## IX. Limitations of the Study

Despite the strong performance of transformer-based models for Classical Arabic NER demonstrated in this study, there are several limitations. First, the dataset used (CANERCorpus) is relatively outdated (2018), which may limit the applicability of our findings to more recent or varied Classical Arabic texts. Second, the corpus consists exclusively of Hadith texts, which means the models are primarily exposed to religious and historical language; this may affect their generalization to other Classical Arabic domains such as literature or formal documents. Third, while our models achieved high overall F1 scores, some entity types remain challenging, particularly ambiguous words or rare entities. Finally, the evaluation was limited to standard metrics (precision, recall, F1-score), and further work is needed to assess models' robustness in real-world NLP applications or cross-domain scenarios. Acknowledging these limitations provides transparency and helps readers interpret the scope of our findings accurately.

## X. Conclusion and Future Work

This study evaluated the performance of transformer-based pretrained language models for Classical Arabic NER using the CANERCorpus of Hadith texts. The results showed that all models achieved strong performance, with F1 scores above 96%, confirming the effectiveness of transformer-based approaches for morphologically rich and linguistically complex languages. Among the models, CAMeLBERT-CA-NER achieved the highest F1 score (97.78%), demonstrating the advantages of domain-specific pretraining on Classical Arabic texts. While models trained on MSA and social media text generalized reasonably well to CA, the error analysis revealed that they were more prone to misclassifying religious and historical entities, emphasizing the importance of domain-relevant data for accurate NER. The novelty of this study lies in its systematic benchmarking and detailed linguistic analysis rather than proposing a new architecture. By systematically comparing multiple transformer-based models on Classical Arabic data, this work establishes a valuable reference point for future advancements in Arabic NER and provides insights for developing domain-specific pretrained models.

Future research can build on these findings in several directions. First, expanding the dataset with additional Classical Arabic sources, including underrepresented entity types, would enhance model coverage and reduce ambiguity-related errors. In particular, while the CANERCorpus (2018) remains a valuable resource, incorporating more recent datasets would improve the relevance and applicability of the findings. Second, investigating cross-domain transfer and domain adaptation between Classical Arabic, Modern Standard Arabic, and dialectal varieties could further improve adaptability and robustness across diverse text genres. Third, incorporating strategies to handle ambiguous words and diacritic variations may enhance model accuracy. Fourth, exploring model interpretability techniques to understand how transformer models capture linguistic cues and contextual dependencies would provide valuable insights into their decision-making process. Finally, integrating these pretrained models into downstream applications such as semantic search, question answering, and historical text analysis would demonstrate their practical utility and support broader NLP research in Classical Arabic.

[13] https://github.com/OpenITI/RELEASE/tree/v2019.1.1

## REFERENCES

[1] A. Sharma, S. Chakraborty, and S. Kumar, "Named entity recognition in natural language processing: A systematic review," in Proceedings of Second Doctoral Symposium on Computational Intelligence, Springer, 2022, pp. 817–828.

[2] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, S. R. El-Beltagy, and W. El-Hajj, "A panoramic survey of natural language processing in the Arab world," Commun ACM, vol. 64, no. 4, pp. 72–81, 2021, doi: 10.1145/3447735.

[3] T. El Moussaoui and C. Loqman, "Advancements in Arabic Named Entity Recognition: A Comprehensive Review," IEEE Access, 2024.

[4] S. Albahli, "An Advanced Natural Language Processing Framework for Arabic Named Entity Recognition: A Novel Approach to Handling Morphological Richness and Nested Entities," Applied Sciences, vol. 15, no. 6, p. 3073, 2025.

[5] B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on named entity recognition—datasets, tools, and methodologies," Natural Language Processing Journal, vol. 3, p. 100017, 2023.

[6] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named entity recognition and relation extraction: State-of-the-art," ACM Computing Surveys (CSUR), vol. 54, no. 1, pp. 1–39, 2021.

[7] D. N. Shah and H. B. Bhadka, "Named entity recognition from Gujarati text using rule-based approach," in Intelligent Systems Design and Applications: 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017) held in Delhi, India, December 14-16, 2017, Springer, 2018, pp. 797–805.

[8] A. Anandika, S. Chakravarty, and B. K. Paikaray, "Named entity recognition in Odia language: a rule-based approach," International journal of reasoning-based intelligent systems, vol. 15, no. 1, pp. 15–21, 2023.

[9] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," IEEE Trans Knowl Data Eng, vol. 34, no. 1, pp. 50–70, 2020.

[10] X. Qu, Y. Gu, Q. Xia, Z. Li, Z. Wang, and B. Huai, "A Survey on Arabic Named Entity Recognition: Past, Recent Advances, and Future Trends," IEEE Trans Knowl Data Eng, vol. 36(03), pp. 943-959., 2024.

[11] M. Jarrar, M. Khalilia, and S. Ghanem, "Wojood: Nested arabic named entity corpus and recognition using bert," in Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3626–3636.

[12] K. Shaalan, "Rule-based approach in Arabic natural language processing," The International Journal on Information and Communication Technologies (IJICT), vol. 3, no. 3, pp. 11–19, 2010.

[13] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay named entity recognition based on rule-based approach," 2014.

[14] M. Hudhud, H. Abdelhaq, F. Mohsen, A. Rocha, L. Steels, and J. van den Herik, "ArabiaNer: A System to Extract Named Entities from Arabic Content.," in ICAART (1), 2021, pp. 489–497.

[15] K. Shaalan and H. Raza, "Person name entity recognition for Arabic," in Proceedings of the 2007 workshop on computational approaches to semitic languages: common issues and resources, 2007, pp. 17–24.

[16] D. Samy, A. Moreno, and J. M. Guirao, "A proposal for an Arabic named entity tagger leveraging a parallel corpus," in International Conference RANLP, Borovets, Bulgaria, 2005, pp. 459–465.

[17] M. A. Ali, A. B. A. Alwahhab, and Y. Farjami, "An Integrated Deep Learning Framework Combining LSTMCRF, GRU-CRF, and CNN-CRF with Word Embedding Techniques for Arabic Named Entity Recognition.," International Journal of Robotics & Control Systems, vol. 5, no. 2, 2025.

[18] M. S. Al-Qurishi and R. Souissi, "Arabic named entity recognition using transformer-based-crf model," in Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021), 2021, pp. 262–271.

[19] B. Ait Benali, S. Mihi, I. El Bazi, and N. Laachfoubi, "New approach for Arabic named entity recognition on social media based on feature selection using genetic algorithm," International Journal of Electrical and Computer Engineering, vol. 11, no. 2, p. 1485, 2021.

[20] R. M. Hamad and A. M. Abushaala, "Medical named entity recognition in arabic text using SVM," in 2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA), IEEE, 2023, pp. 200–205.

[21] Y. Benajiba, I. Zitouni, M. Diab, and P. Rosso, "Arabic named entity recognition: using features extracted from noisy data," in Proceedings of the ACL 2010 conference short papers, 2010, pp. 281–285.

[22] M. Gridach, "Character-aware neural networks for arabic named entity recognition for social media," in Proceedings of the 6th workshop on South and Southeast Asian natural language processing (WSSANLP2016), 2016, pp. 23–32.

[23] I. Keraghel, S. Morbieu, and M. Nadif, "A survey on recent advances in named entity recognition," arXiv preprint arXiv:2401.10825, 2024.

[24] M. N. A. Ali, G. Tan, and A. Hussain, "Bidirectional recurrent neural network approach for Arabic named entity recognition," Future Internet, vol. 10, no. 12, p. 123, 2018.

[25] D. Awad, C. Sabty, M. Elmahdy, and S. Abdennadher, "Arabic name entity recognition using deep learning," in Statistical Language and Speech Processing: 6th International Conference, SLSP 2018, Mons, Belgium, October 15–16, 2018, Proceedings 6, Springer, 2018, pp. 105–116.

[26] M. N. A. Ali, G. Tan, and A. Hussain, "Boosting Arabic named-entity recognition with multi-attention layer," IEEE Access, vol. 7, pp. 46575–46582, 2019.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, 2018, pp. 4171–4186.

[28] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 9–15.

[29] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: deep bidirectional transformers for Arabic," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 7088–7105.

[30] A. Safaya, M. Abdullatif, and D. Yuret, "Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media," in Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 2054–2059.

[31] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," in Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021, pp. 92–104.

[32] S. Bilal, "A Linguistic System for Predicting Sentiment in Arabic Tweets," in 2021 3rd International Conference on Natural Language Processing (ICNLP), IEEE, 2021, pp. 134–138.

[33] B. Kessentini, T. Alimi, and R. Boujelben, "Assessing BERT Models for Arabic Named Entity Recognition," in Advancements in Machine Learning and Natural Language Processing: Innovations and Applications: 3rd International Conference on Language Processing and Knowledge Management (LPKM'2024), Springer Nature, 2025, p. 194.

[34] W. Lan, Y. Chen, W. Xu, and A. Ritter, "Gigabert: Zero-shot transfer learning from english to arabic," in Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP), 2020.

[35] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-training bert on arabic tweets: Practical considerations," arXiv preprint arXiv:2102.10684, 2021.

[36] R. E. Salah and L. Q. B. Zakaria, "Building the classical Arabic named entity recognition corpus (CANERCorpus)," in 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), IEEE, 2018, pp. 1–8.

[37] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," Multimed Tools Appl, vol. 82, no. 3, pp. 3713–3744, 2023.

[38] R. Salah, M. Mukred, L. Qadri binti Zakaria, R. Ahmed, and H. Sari, "A new rule-based approach for classical arabic in natural language processing," Journal of Mathematics, vol. 2022, pp. 1–20, 2022.

[39] N. Alsaaran and M. Alrabiah, "Classical Arabic named entity recognition using variant deep neural network architectures and BERT," IEEE Access, vol. 9, pp. 91537–91547, 2021.

## APPENDIX A: AMBIGUOUS ARABIC WORDS AND THE IMPACT OF DIACRITICS

| Arabic form | English Meaning | Transliterations | Part of Speech (POS) |
|---|---|---|---|
| حُر | Freedom | Ḥurr | Noun |
| حَر | Hot | Ḥarr | Adjective |
| حَلْم | Patience | Ḥilm | Noun |
| حِلْم | Forgivingness | Ḥilm | Noun |
| حُلْم | Dream | Ḥulm | Noun |
| عِلْم | Knowledge | Elm | Noun |
| عَلَم | Flag | Alam | Noun |
| عَلِمَ | Knew | Alima | Verb |
| عُلِمَ | Is known | Ulima | Verb |
| عَلَّمَ | Taught | Allama | Verb |
| عُلِّمَ | Is taught | Ullima | Verb |