A Machine Learning Approach for Named Entity Recognition in Classical Arabic Natural Language Processing

Ramzi Salah[1], Muaadh Mukred[1,2,*], Lailatul Qadri binti Zakaria[1], and Fuad A. M. Al-Yarimi[3]
[1] Faculty of Information Science and Technology, University Kebangsaan Malaysia,
43600 Bangi, Selangor, Malaysia.
[2] Department of Business Analytics, Sunway Business School, Sunway University,
5, Jalan University, 47500, Petaling Jaya, Bandar Sunway, Selangor, Malaysia.
[3] Department of Computer Science, King Khalid University, Muhayel Aseer, Kingdom of Saudi Arabia.
[E-mail: muaadhm@sunway.edu.my]
*Corresponding author: Muaadh Mukred

Received March 27, 2024; revised August 27, 2024; accepted September 23, 2024; published October 31, 2024

Abstract

A key element of many Natural Language Processing (NLP) applications is Named Entity Recognition (NER). It involves categorizing and identifying text into separate categories, such as identifying a location or an individual's name. Arabic NER (ANER) is also utilized in numerous other Arabic NLP (ANLP) tasks, such as Machine Translation (MT), Question Answering (QA), and Information Extraction (IE). ANER systems can often be classified into three major groups: rule-based, Machine Learning (ML), and hybrid. This study focuses on examining ML-based ANER developments, particularly in the context of Classical Arabic, which presents unique challenges due to its complex morphological structure and limited linguistic resources. We propose a supervised approach that integrates word-level, morphological, and knowledge-based features to improve NER performance for Classical Arabic. Our method was evaluated on the CANERCorpus, a specialized dataset containing annotated texts from Classical Arabic literature. The Naive Bayes (NB) approach achieved an F-measure of 80%, with precision and recall levels at 86% and 75%, respectively. These results indicate a significant improvement over traditional methods, particularly in dealing with the intricate structure of Classical Arabic. The study highlights the potential of ML in overcoming the challenges of ANER and provides directions for further research in this domain.

Keywords: Arabic Named Entity Recognition (ANER); Classical Arabic; Modern standard Arabic; Machine Learning (ML); Natural Language Processing (NLP).

1. Introduction

In various applications with regard to Arabic Natural Language Processing (ANLP), the issue of Named Entity Recognition (NER) is crucial. Machine Translation (MT), Question Answering (QA), Information Retrieval (IR), as well as Information Extraction (IE) are only several of the critical activities that NER may be used for. Applications that use NER as an essential initial step can perform significantly overall [1, 2].

The first NER task presentation took place during the MUC-6 or the Sixth Message Understanding Conference. However, other names from certain domains, such as People, Location, Organisation, Sports, and many more, can be discovered in the text. Named Entities (NEs) are the terms used to describe these names. NER seeks to automatically identify and group these names into specified categories in text. Approaches based on Machine Learning (ML) are more beneficial since the system may be trained and quickly adjusted to other language domains [3, 4].

A specific identifier, or NE, is a word or phrase that clearly differentiates one entity from a collection of others that have similar traits. The term "named" with regard to the phrase NE limits the entities' range that has one or more definite designators that serve as referents. Proper names are typically included in fixed designators, although it depends on the area of interest. Other than that, NE can also be used as the reference term for objects in the domain [5].

In this context, NLP is essential to many applications, including information retrieval, question-answering, and MT [1, 2]. NER is typically referred to in NLP as a sequence labelling task where each word in a phrase is provided with a different label. It is normal practice to model and address NLP tasks using sequence labelling, where the input values are frequently words. Nevertheless, depending on the task, the input values might potentially be smaller units like individual characters [3, 4].

However, implementing NER in Arabic presents unique challenges due to the language's complex morphological structure, extensive use of inflexion, and the prevalence of dialectal variations. Arabic's rich morphology, where words can have multiple forms due to the addition of prefixes, suffixes, and infixes, makes extracting named entities more difficult than languages like English. Additionally, the scarcity of annotated corpora and linguistic resources for Classical Arabic further complicates the development of robust NER systems.

With almost 420 million speakers worldwide, Arabic is among the most widely spoken languages. According to [6], it is the official language of 24 nations, most in North Africa and the Middle East. The United Nations has six official languages, with Arabic as one of them [7]. Languages now play a crucial role in technology due to the growing usage of technology and tools for translation and information retrieval. Research in Arabic language processing is becoming crucial to keep current with new technologies as the use of Arabic language in social media as well as technology continues to expand. The study of NER in Arabic is still in its infancy, although there has been a significant amount of research on NER in English. The processing of Arabic has specific challenges in addition to a dearth of resources and annotated corpora. Arabic's morphological structure makes it exceptionally challenging to extract NE from the text [8, 9]. Extraction of NE in Arabic might be difficult due to its morphological structure. This is so because Arabic words can contain any combination of prefixes, stems, and suffixes. In Arabic, a letter's shape can also vary based on where it appears in a word. This makes it more challenging to recognize NE in Arabic text [10, 11].

This research suggests a NER approach for ML-based analysis of classical Arabic. The approach is based on word level, morphology, and knowledge-based aspects, which are produced from linguistic data on Arabic NE. The method uses several strategies to enhance

the performance of NLP tasks on traditional Arabic texts, such as text categorization and feature extraction. Experiments have revealed that this strategy is superior to both traditional methods and other cutting-edge strategies in terms of effectiveness. The report also identifies difficulties in dealing with classical Arabic and recommends possible directions for further investigation. Additionally, the study analyses previous research on Arabic NER (ANER) utilising ML systems, including the types of linguistic, domain, and entity, methods employed, and performance reported. The research also examines certain difficulties with ANER in text, and the models and characteristics utilized in ML methodologies. These contributions aim to advance the state-of-the-art in Arabic NER and provide a foundation for further research in this area.

The remainder of this study is structured as the following: An overview of relevant work on the topic is provided in Section II. The ML NER approach suggested in this study, and the many processes involved in its execution are described in Section III, along with the methodology utilized to detect NE in Arabic text. The effectiveness of the suggested strategy is assessed in Section IV. Section V brings the paper to a conclusion.

2. Related Work

The application of ML methods to recognize NE in classical Arabic, the dialect used in literature and official writings, has been the subject of several studies. Because ML methods have been shown to be successful in extracting characteristics, several latest research investigations have included them [12].

Benajiba and Rosso [13] performed a study that attempted to use the CRF approach rather than Maximum Entropy to improve the performance with regard to NER for classical Arabic. They employed elements like nationality, gazetteers, Base Phrase Chunks, as well as POS tags in the system they designed. Their study's findings demonstrated great accuracy, with recall, precision, as well as F-measure being, respectively, 72.77%, 86.90%, and 79.21%.

Using annotated text corpora, which marks the text with the entities it includes, is a frequent method for training NER models for classical Arabic. Using this method and training a model on a corpus of articles, Al-Twairesh, Al-Khalifa [14] were able to reach a high degree of accuracy, as indicated by the F1-score of 89.3%.

On the other hand, Abdul-Hamid and Darwish [15] established a CRF-based method for identifying three categories of NE in Arabic, including organizations, locations, and people. The ACE2005, as well as ANERcorp datasets, were utilized to evaluate their method, which exclusively employed surface features. The system's accuracy was 89%, the recall was 74%, and F-measure was 81%, according to the findings. These findings suggest that the system is more precise compared to the one described in Benajiba and Rosso [13] study.

Additionally, AbdelRahman, Elarnaoty [16] utilized a mixture of two ML methods, CRF and bootstrapping, to enhance ANER. The system made use of morphological features, BPC, gazetteers, and POS tags. It was able to correctly recognize a variety of NE, including people, locations, organizations, devices, dates and times. When evaluated on the ANERcorp dataset, the system outperformed the Ling Pipe NER tool, with F-measures for the various NE of 74.06%, 89.09%, 75.01%, 69.47%, 77.52%, 80.95%, 80.63%, 98.52%, 76.99%, as well as 96.05%, correspondingly.

Similarly, Bidhend, Minaei-Bidgoli [17] developed the Noor CRF-based NER system to extract people's names from religious scriptures. They established a corpus called NoorCorp that was concentrated on three Arabic-language texts on Islamic law and jurisprudence. Noor-Gazette, a gazetteer with regard to religious names, was also created by them. In terms of F-

measure, the system's overall performance was determined to be 99.93%, 93.86%, and 75.86%, respectively, when evaluated on the corpora of history, hadith, and law.

Morsi and Rafea [18] examined how various features affected the performance of a conditional random field-based ANER system for text written in Modern Standard Arabic. They created a baseline model for comparison and employed CRF-based models. The algorithm was able to extract four different categories of entities from the dataset utilized in the study, including people, locations, organizations, and other sorts. The system's F-measure performance was the highest at 68.05.

Research by Zirikly and Diab [19] stated that a dialectal ANER system was established utilising Egyptian colloquial Arabic. To identify people and locations, the ML method used the CRF method. NER characteristics comprised lexical and contextual features, distance from certain keywords, gazetteers, as well as Broun clustering. The authors additionally used web blogs from the Linguistic Data Consortium, concentrating on those published in the Egyptian dialect, to produce an annotated dataset for the Egyptian dialect. The findings revealed that the system had an F-measure of 49.18% for names of people and 91.429% for locations.

Recently, Alduailaj and Belghith [20] employed ML to examine Arabic-language cyberbullying. They used actual data from Twitter as well as YouTube to train and assess a Support Vector Machine (SVM) classification algorithm. With a detection rate of 95.742%, the findings presented that the SVM method outperformed the Naive Bayes (NB) algorithm. Due to this model's high precision, users will be better protected from online bullying.

Alsayadi and ElKorany [21] provide an integrated ML model with regard to ANER based on semantics. To infer the semantic connections between NE, the authors utilised a variety of linguistic features and made use of syntactic dependencies. Moreover, to extract people, organizations, and locations, they used a CRF classifier. The findings showed that this strategy had an overall F-measure of 87.86% for ANERCorp datasets and 84.72% for ALTEC datasets.

Dahan, Touir [22] developed an ANER system to handle inflection and ambiguity in the Arabic language, based on Hidden Markov Models (HMM) and employing a stemming mechanism. The technology is entirely automated and is capable of identifying Arabic names for people, organizations, and locations. The method was tested by the authors using a corpus created by the Al-Hayat, Assabah, and France Press Agency. The accuracy and recall performance indicators for the system were determined to be 73% and 77%, respectively, while the F-measure for people, organizations, as well as locations was 79%, 67%, and 78%.

Consequently, Al-Shoukry and Omar [23] utilised an ANER ML system having Decision Trees (DT) in the area of crime in Modern Standard Arabic (MSA) in research. Their method used a DT Classifier (DTC) with extraction features to extract NE of people, locations, crime types, times, and dates. The highest F-measure obtained was 81.35%. Note that the dataset was sourced from internet sources.

Aoumeur, Li [24] performed research utilising the CASAD dataset of Arabic literature for sentiment analysis, which was statistically evaluated using a respectable Cronbach's alpha. Additionally, they used six ML approaches to assess the word2vec and TF-IDF feature extraction methods of CASAD. The results revealed that when utilising LR with a word2vec feature extraction approach for binary classification (positive and negative), the accuracy of categorising Arabic text using well-known ML methods like SVM, LR, and NB was 71.42%.

In the research by Koulali and Meziane [25], the authors demonstrated an ANER that combined an SVM classifier and pattern extractor using patterns from POS-tagged text. Using CoNLL conference data, the system was able to recognize various types of NE and it used both dependent and independent features. Note that 90% of the ANERCorp data were utilised for training, while the remaining data were employed for testing. The maximum F-measure

obtained was 83.20% when the system was assessed as having different feature combinations.

Research by Mohammed and Omar [26] described a method using a neural network for extracting NE from Arabic. They divided identified entities into four categories—people, location, organization, and others—using ANERcorp and extra online resources. Additionally, the researchers used the same data to assess the performance of their neural network model with a decision tree model. According to the findings, the DT model had an accuracy of 87% compared to 92% for the neural network model.

On the other hand, Alanazi [27] extracted disease names, symptoms, treatments, and diagnostic methods from current Arabic medical texts using a hybrid methodology. The system's outcomes demonstrate the BBN's effectiveness, having an overall F-measure of 71.05%. When tested on the F-measure for the capacity to recognize disease names, a score of 98.10% was obtained, but when tested on the ability to recognize symptoms, only a score of 41.66% was attained.

Despite these advancements, a significant gap remains in the effective handling of NER for Classical Arabic [63-65]. Most existing research has concentrated on MSA or dialectal Arabic, leaving the specific challenges of Classical Arabic underexplored. This study addresses this gap by proposing a supervised machine learning approach that incorporates word-level, morphological, and knowledge-based features tailored for Classical Arabic. Our research provides a novel contribution by focusing on the underrepresented area of Classical Arabic NER, offering a method that improves upon the limitations identified in prior work. This research has demonstrated that high performance outcomes may be achieved using a variety of strategies, including feature engineering, preprocessing, and ensemble methods. Nevertheless, there is potential for development, especially when it comes to problems with data sparsity and annotation quality. Future research should concentrate on resolving these issues and investigating novel approaches to improve the accuracy and robustness of NER algorithms for classical Arabic.

3. Methodology

The development of ML methods has considerably aided the recent rapid growth of NLP. In this research, we investigate the use of ML in classical Arabic, a formal as well as literary variety of Arabic [12, 20].

Supervised learning is a method that has been frequently applied in NLP for classical Arabic. In this case, a model is trained on a labelled dataset, and predictions are made on new text using the patterns discovered from the data. Using a dataset of text with NE labelled, for instance, a NER model may be trained to identify NE in a new text [20].

Unsupervised learning, which includes training a model on a dataset devoid of labels, is another prominent method. Algorithms for unsupervised learning can be used to identify patterns concerning the data and combine related data points [28]. According to [29, 30], an unsupervised MT model may, for instance, train to translate text by observing patterns in the source and target language text and translating texts based on those patterns.

NLP for classical Arabic has also used semi-supervised learning, which entails training a model on a dataset with both labelled and unlabeled input [31]. This method may be helpful when there is a shortage of labelled data.

Another ML method that has been used for NLP for classical Arabic is active learning. The model is iteratively improved through active learning by choosing the most relevant data points to label [32]. When labelling data is time-consuming or expensive, this can

effectively boost a model's performance [62].

The success of ML depends on the annotations of NEs by language experts [33, 34]. It is essential to have access to a considerable number of domain-relevant texts that can be examined manually in order to develop effective rules [35]. The expertise and abilities of the knowledge engineer are crucial for creating an effective system.

Several iterative methods had to be used in order to produce an accurate system. Every method begins with the development of rules for a training corpus of texts. This evaluation's goal is to determine if the rules should be altered in light of these trials' results [36, 37]. This section explained the knowledge sources needed to recognize NEs in classic Arabic texts.

3.1 Dataset

A collection of documents written in the classical Arabic language and annotated with names of people, organisations, and locations is known as a classical Arabic NER corpus. For the purpose of NER in classical Arabic, this type of corpus is utilised for training and assessing models. To provide the model with a varied range of NE and situations to learn from, the corpus often contains a number of text types, including articles, historical records, and literary works. The corpus includes annotations for the specified entities in addition to the text, such as information about their kind (such as people or organisation) and the amount of text they take up in the text. The NER corpus is a crucial tool for NLP researchers to create and test NER models for classical Arabic, a language with limited resources [38].

The CANERCorpus [38], a classical Arabic NER corpus annotated by human subject matter experts, served as the study's dataset. Approximately 7,000 Hadiths (prophetic sayings) obtained from the Sahih Al-Bukhari book are included in the collection, which has been annotated with 21 distinct entity classifications. There are several categories accessible, including People, Location, Organisation, Unit, Money, Book, Date, Time, Family, Natural Item, Offence, Sect, Religion, Prophet, God, Number, Day, Heaven, Hell, Month, and Other. Approximately 258,264 words and 72,108 recognised entities make up the corpus. **Table 1** lists the NE in each tag.

In order to investigate the CANERCorpus, NE was divided into two primary categories, as seen in **Fig. 1**. In the first category, which is referred to as "general," there are classifications for people, locations, organizations, money, books, dates, times, crimes, days, and numbers. Numerous fields, including politics, economics, sports, and even crime, contain NE. The second category, referred regarded as "specific domain," is exclusive to the Islamic world. It has subcategories for "sect," "religion," "prophet," "Allah," "heaven," and "hell." On the other hand, the corpus's background is primarily concerned with the Islamic world. As a result, NE in the Islamic domain has names, meanings, and functions that are substantially distinct from those in other domains.

3.2 Features

Features are referred as words 'characteristics that the algorithm considers [39]. ML and rule-based methods both use similar implementation strategies. Both of these linked traits are used in NER in various ways. An example illustration of a feature is as follows:

1) Boolean features are either false or true.
2) A numeric feature is associated with how many characters are in a word.
3) A nominal feature resembles the word in lower case.

The supervised method, instead of relying on a collection of predetermined features,

employs a classification methodology to manage the NERs. Multiple categorization methods were used to identify which words were NEs because Arabic is a complicated language [40]. The ANER algorithm was required to decipher the initial settings of the NEs due to the complexity of Arabic. Several characteristics were used for this.

Using a variety of feature types is essential for supervised NER algorithms to function effectively. The word's own characteristics, such as its part of speech, shape, and capitalization, as well as morphological and knowledge-based qualities, might be considered among these features. Word-level features, such as proper nouns being capitalized or words in names or locations having a certain form, might help discover patterns in the text that are suggestive of NE [41, 42].

The existence of affixes or the word's root are examples of morphological elements that provide information about the structure of words. In languages with rich inflectional and derivational morphology, like Arabic, these elements can be helpful for distinguishing NE [43, 44]. For instance, Arabic NEs may frequently be recognized by their distinctive morphological patterns.

Knowledge-based features, often referred to as external features, are based on external knowledge sources like ontologies, gazetteers, and Wikipedia. The NER model's performance may be enhanced using these elements to include external information. A gazetteer of well-known NE, for instance, can be used to locate NE in the text [45, 46].

In conclusion, obtaining excellent performance in supervised NER methods requires the capacity to manage many feature types, including word-level, morphological, and knowledge-based features. Important to keep in mind is that the feature management process depends on the nature of the NER problem and the resources that are available.

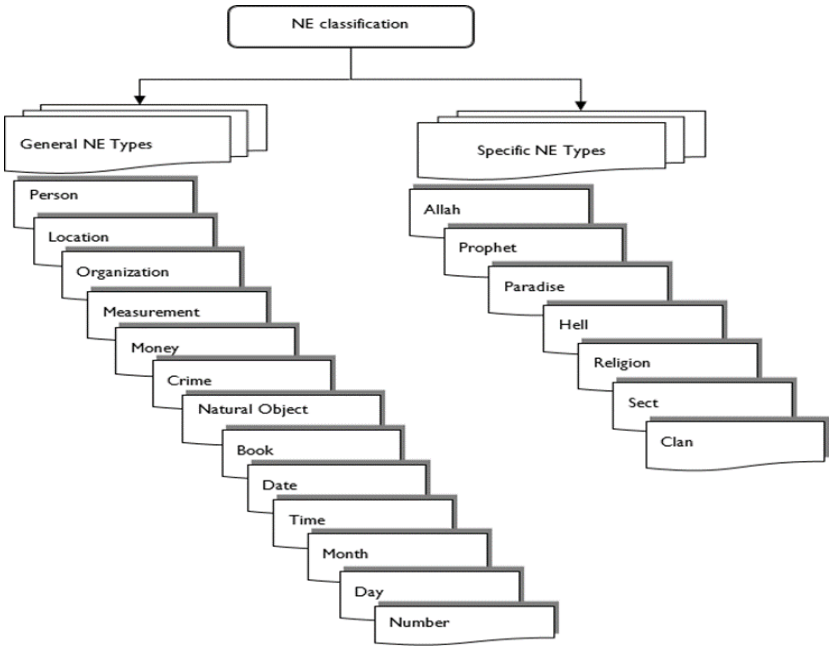

Fig. 1. CANERCorpus work, NE Classification

Table 1. Number with regard to Named Entities in every Tag.

“Type	**Count**	**Percentage**
Allah	7811	12.95
Prophet	6502	10.77
Pers	39159	64.87
Loc	1349	2.09
Org	9	0.01
Meas	147	0.24
Mon	139	0.23
Book	183	2.24
Date	596	0.95
Time	102	0.17
Rlig	184	0.31
Sect	17	0.03
Clan	674	1.11
NatOb	670	1.11
Crime	212	0.35
Para	294	0.49
Hell	245	0.41
Month	77	0.13
Day	31	0.05
Num	13707	1.51
Named Entity	72108	100.00”

3.3 Features

Features are referred as words 'characteristics that the algorithm considers [39]. ML and rule-based methods both use similar implementation strategies [62, 63]. Both of these linked traits are used in NER in various ways. An example illustration of a feature is as follows:

4) Boolean features are either false or true.
5) A numeric feature is associated with how many characters are in a word.
6) A nominal feature resembles the word in lowercase.

The supervised method, instead of relying on a collection of predetermined features, employs a classification methodology to manage the NERs. Multiple categorization methods were used to identify which words were NEs because Arabic is a complicated language [40]. Due to the complexity of Arabic, the ANER algorithm was required to decipher the initial settings of the NEs. Several characteristics were used for this.

Using a variety of feature types is essential for supervised NER algorithms to function effectively. The word's own characteristics, such as its part of speech, shape, and capitalization, as well as morphological and knowledge-based qualities, might be considered among these features. Word-level features, such as proper nouns being capitalized or words in names or locations having a certain form, might help discover patterns in the text that suggest NE [41, 42].

The existence of affixes or the word's root are examples of morphological elements that provide information about the structure of words. In languages with rich inflectional and derivational morphology, like Arabic, these elements can be helpful for distinguishing NE [43, 44]. For instance, Arabic NEs may frequently be recognized by their distinctive morphological patterns.

Knowledge-based features, often referred to as external features, are based on external knowledge sources like ontologies, gazetteers, and Wikipedia. The NER model's performance may be enhanced using these elements to include external information. A gazetteer of well-known NE, for instance, can be used to locate NE in the text [45, 46].

In conclusion, obtaining excellent performance in supervised NER methods requires the capacity to manage many feature types, including word-level, morphological, and knowledge-based features. Important to keep in mind is that the feature management process depends on the nature of the NER problem and the resources that are available. The features that are utilised to spot NEs in Arabic literature are described in the next section.

3.3.1 Word Level

During the learning phase, the word level feature explains the target word's characteristics [16, 21, 47]. In this research, the word level included the word length [48-50] n-word sliding wind and POS tagging, sentence word order, before and after stemming [16, 19, 21, 27, 47, 50-52], as well as the frequency or distribution of each NE type with respect to the training data set [47]. The word distribution in a data set offered useful information in examining an NE that often appears in a dataset. For instance, 'المنورة' (almunauarah) refers to a location NE. Compared to other NEs, it appears often. This indicates that the classifier would be enabled to identify this word as a Location NE by recognizing the distribution or frequency. The word level features are listed in **Table 2**.

Table 2. Word level features.

Feature	Description
prevword3	Third word prior to the certain NE.
prevwordAcual3	The NE type of the third word prior to the certain NE.
prevwordPOS3	The POS of the third word prior to the certain NE.
prevword2	Second word prior to the certain NE.
prevwordAcual2	The NE type of the second word prior to the certain NE.
prevwordPOS2	The POS of the second word prior to the certain NE.
prevword	First word prior to the certain NE.
prevwordAcual	The NE type of the first word prior to the certain NE.
prevwordPOS	The POS of the first word prior to the certain NE.
nextword	First word following the certain NE.
nextwordAcual	The NE type of the first word following the certain NE.
nextwordPOS	The POS of the first word following the certain NE.
nextword2	Second word following the certain NE.
nextwordAcual2	The NE type of the second word following the certain NE.

nextwordPOS2	The POS of the second word following the certain NE.
nextword3	Third word following the certain NE.
nextwordAcual3	The NE type of the third word following the certain NE.
nextwordPOS3	The POS of the third word following the certain NE.
lengthWord	This feature may be utilised to check if the length of a word is less than three or not because it has been discovered that very short words are not NE.
freqNE	Infrequent words are collected by calculating the word frequency in the utilised corpus during the training phase and then selecting the cut off frequency to build the binary feature.”

3.3.2 Morphological Features

Given Arabic's inflectional nature, its broad morphemes' diversity may be utilized as attributes by ANER to generate a vocabulary [40]. Morphological features are what constitute Arabic as a distinct language [3, 53]. Previous studies suggested that the usage of morphological feature sets (MADAMIRA) might enhance ANER's approach to morphological analysis as well as disambiguation for Arabic [21, 27, 47, 49, 53-55]. The morphological specifics, often referred to as features that MADAMIRA was able to extract, are discussed below:

1) The prefix and suffix feature may be used to determine if a word has any suffixes or prefixes. This is crucial since suffixes and prefixes are uncommon in Arabic NEs. For instance, “عبد الله”, “مكة/ Makah”, as well as “الحجاز/ The Hijaz are both Arabic NEs with no affix [16, 50].
2) In Arabic, there are three forms with respect to the verb that can be differentiated by the aspect feature: the perfect (ماضي), imperfect (مضارع), as well as imperative (أمر). Arabic nominative (مرفوع), accusative (منصوب), as well as genitive (مجرور) cases, are all supported by this case characteristic. Note that binary feminine and masculine values were recognized by the gender feature [47].
3) For every word, the number function may take on one of three possible values to present its single, dual, or plural forms.
4) The mood function determined the three idiomatic forms with regard to the imperfective verb in Arabic. They are the indicative (مرفوع), the subjunctive (منصوب), as well as the jussive (مجزوم).
5) The definite, indefinite, as well as construct states were all acknowledged by the state function.
6) The speech recognition system comprehends the binary labels for the two voice types, active as well as passive.
7) The Enclitics, as well as Proclitic, feature exact clitics attached to a stem.
8) The part-of-speech (POS) feature identifies a binary value that is present if the POS tag refers to a proper noun or noun.
9) The diptote (“الممنوع من الصرف”) feature recognizes dipotated words. Generally, diptoted Arabic words, for example, أحمد /Ahmed” and مكة/ Makah are NEs in Arabic.
10) The definite article (“ال ”/“the”) feature is crucial to determine NEs due to the names of many organizations begin with an article, for example, “ الأمم المتحدة ”/ “The United Nations”.
11) The interjection feature identifies interjections, for example, “يا”/“oh”. It is utilised to

recognize NEs related to a person because this type of interjection usually happens before the person's name.

12) The relative pronoun feature identifies when a word is preceded by pronouns, for example, "الذي، التي، الذين" since most relative pronouns are followed by a NE.

Number	Feature	Feature value definition
1	Aspect	*Verb aspect*: Command, Imperfective, Perfective or Not applicable (NA)
2	Case	*Grammatical case*: Nominative, Accusative, Genitive, NA or Undefined
3	Gender	*Nominal gender*: Feminine, Masculine or NA
4	Mood	*Grammatical Mood*: Indicative, Jussive, Subjunctive, NA or Undefined
5	Number	*Grammatical number*: Singular, Plural, Dual, NA or Undefined
6	Person	*Person information*: 1st, 2nd, 3rd or NA
7	State	*Grammatical state*: Indefinite, Definite, Construct/Poss/Idafa, NA or Undefined
8	Voice	*Verb voice*: Active, Passive, NA or Undefined
9	Proclitic 3	*Question proclitic*: No proclitic (NP), NA or Interrogative Particle > a
10	Proclitic 2	*Conjunction proclitic*: NP, NA, Conjunction fa, Response conditional fa, Subordinating conjunction fa, Conjunction wa, Particle wa or Subordinating conjunction wa
11	Proclitic 1	*Preposition proclitic*: NP, NA, Particle bi, Preposition bi, Preposition ka, Emphatic Particle la, Preposition la, Response conditional la, Jussive li, Preposition li, Future marker sa, Preposition ta, Particle wa, Preposition wa, Preposition fy, Negative particle lA, Negative particle mA, Vocative yA, Vocative wA or Vocative hA
12	Proclitic 0	*Article proclitic*: NP, NA, Determiner, Negative particle lA, Negative particle mA, Relative pronoun mA or Particle mA
13	Enclitics	*Pronominal*: No enclitic, NA, 1st person (plural|singular), 2nd person (dual|(feminine (plural|singular))|(masculine (plural|singular))), 3rd person (dual|(feminine (plural|singular))|(masculine (plural|singular))), Vocative particle, Negative particle lA, Interrogative pronoun (ma|mA|man), Relative pronoun (ma|mA|man) or Subordinating conjunction (ma|mA)
	POS	*POS definition*: Nouns, Number Words, Proper Nouns, Adjectives, Adverbs, Pronouns, Verbs, Particles, Prepositions, Abbreviations, Punctuation, Conjunctions, Interjections, Digital Numbers or Foreign/Latin

Fig. 2. MADA and MADAMIRA morphological features

3.3.3 Knowledge-based Features

In NER, the lists are the main tools utilised with regard to the rule-based as well as ML methods [56]. The term "gazetteers" is expressed as lists of knowledge sources that may comprise different NEs types [1]. There are several Arabic gazetteers that may be employed in defining NEs.

As demonstrated in **Table 3**, knowledge-based features rely on a number of knowledge sources with various features. If a word may be discovered in a gazetteer, a list of keywords or a list of trigger words depends on knowledge-based features. A word is "found" in a gazetteer if at least one unit in the gazetteer exactly matches it. Given greater flexibility, this highly sensitive matching criterion performs better. There are three methods for creating binary values when using gazetteers.

Table 3. Knowledge-based features.

feature	Description
Stop Word	Stop words are words that regularly appear yet are not allowed in NE. This feature is employed to assess if a word appears in the stop words list.
Gazetteers	Lists of particular data, such as names of people, organisations, locations, and days of the week, constitute a gazetteer. A targeted word's likelihood of appearing in a gazetteer is determined by this feature.
Key Word Prev	A NE is recognised using the Key Word Prev and Key Word Next features. These features might include verb or noun lists, among other things. The indication features

Key Word Next	are another name for them. These features established whether a word is on the list of lexical triggers. In texts written in natural language, NEs are identified using a variety of Arabic terminology.
Freq NE	When building the binary feature, the cut off frequency is chosen after determining how frequently a word appears in the corpus during the training phase.
Black list	The Blacklist feature is conducted utilising Blacklist dictionaries that include words that should be rejected as NE.
Pattern	If no pattern is identified, a collection of patterns drawn from the corpus research is either zero or the pattern index in which the word is present.

3.4 Supervised method

The supervised method utilizes the tagged corpus to examine each NE type in the text [57]. Several research have established the effect with regard to different classifiers as well as features on ANER [13, 15, 19, 22, 23, 58-61]. The methods utilised in these investigations typically included phases for evaluation, learning, feature extraction, as well as preprocessing. In this investigation, the pre-processing, as well as feature extraction processes were carried out using MADA [54] as well as MADAMIRA [53]. To represent the corpus, these technologies generated a table with a variety of attributes. In order to categorize the NEs using Weka, the data were normalized. A baseline for the supervised method was established using these tools. According to **Fig. 3**, the methodology suggested in this study employs a supervised methodology.

Pre-processing is the first step in **Fig. 3**. This step is useful for navigating the nuances and complexity of Arabic. The steps included in MADA, in addition to the preprocessing step, are as follows:

1) A lemmatization process to recognize the lemma.
2) POS tagging to analyze each part of speech.
3) A thorough morphological disambiguation procedure that identifies complete or absent morphological features using a ranked list of the most probable feature values of a word in a text.
4) A procedure known as stemming reduced each word to its morphological stem.
5) A tokenization procedure that separates clitics and makes appropriate spelling corrections. The tokenization procedure specifies the output format and the tokenization separation rules.

These preprocessing steps are not just preliminary tasks but are crucial in ensuring the robustness and accuracy of the feature extraction process [62, 63]. By carefully preparing the data through these methods, the model is better equipped to handle the complexities of Classical Arabic, ultimately leading to more reliable and precise NER results [63]. This foundational work is vital for the success of the subsequent classification and evaluation phases, ensuring that the NER system can effectively recognize and categorize entities in a linguistically challenging context [64, 65].

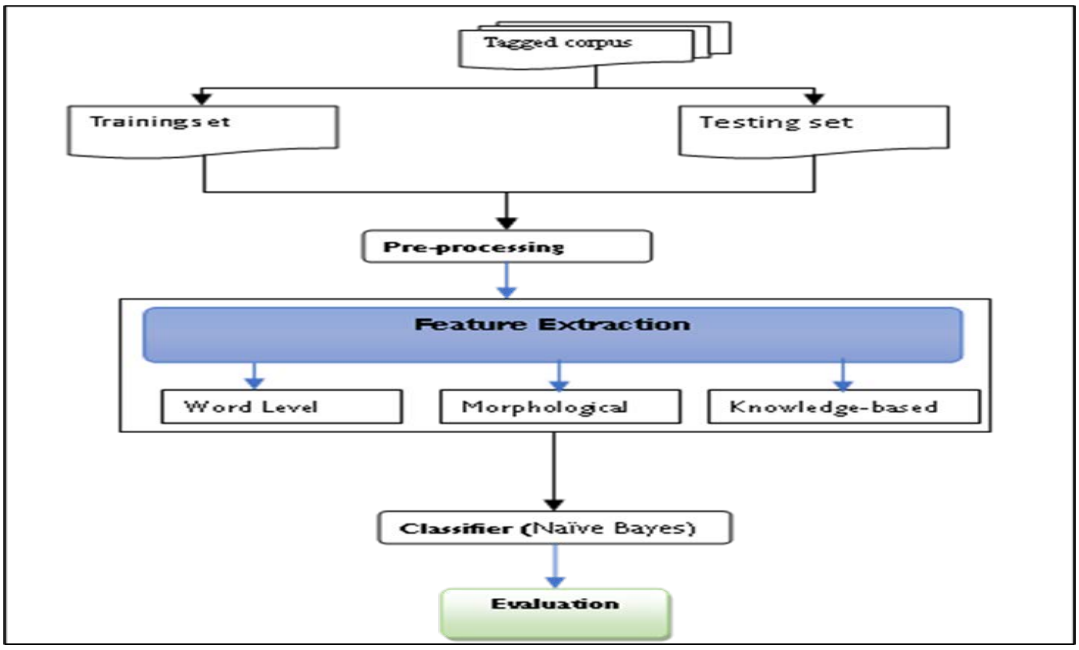

Fig. 3. Supervised method with regard to ANER.

3.5 Multinomial naïve Bayes (NB) Classifier

The Multinomial Naïve Bayes (NB) classifier was chosen for this study due to its simplicity and effectiveness in handling text classification tasks, particularly when dealing with word frequency as a feature. This model assumes that the presence of a particular feature in a class is independent of the presence of other features, which simplifies the computation and allows for efficient training even with large datasets. The Multinomial Naïve Bayes classifier was implemented by first converting the extracted features into numerical vectors, where each feature corresponds to a specific dimension in the vector space. These feature vectors were then used to train the classifier on the annotated CANERCorpus, with the training data comprising the feature vectors and their associated entity labels. During the prediction phase, the classifier computed the posterior probability for each class, such as Person or Location, for each word in the test data and assigned the class with the highest probability to the word, thereby identifying the named entities within the text.

Strong independent assumptions are referred to be "Naive Bayes" in models. Each feature is presumed to be conditionally independent of every other feature in an NB model. The following phrase is used to calculate the probability of observing features *f1* through *fn* in an NB model, given class *c*:

$$p(f1,\ldots,fn|c) = \prod i = 1np(fi|c)p(f1,\ldots,fn|c) = \prod i = 1np(fi|c)$$

To utilise an NB model to categorize a new example, the posterior probability is frequently employed due to its simplicity:

$$p(c|f1,\ldots,fn) \propto p(c)p(f1|c)\ldots p(fn|c)p(c|f1,\ldots,fn) \propto p(c)p(f1|c)\ldots p(fn|c)$$

Because the independence assumptions given by NB models are frequently incorrect, they have earned the title of the "Idiot Bayes" model. NB models still continue to function successfully even when employed to finish demanding tasks when strong independence assumptions are false.

To this point, the distribution with regard to each feature was not discussed. Simply put, $p(fi|c)p(fi|c)$ were not defined. The term "Multinomial Naive Bayes" is utilised when $p(fi|c)p(fi|c)$ resembles a multinomial distribution. This configuration is used for countable data, such as document word counts.

In summary, the conditional independence of each feature in a model is referred to as an NB classifier. A Multinomial NB classifier, in contrast, is a particular type of NB classifier that makes use of a multinomial distribution for each feature.

4. Results and Discussion

The outcomes of the supervised method for ANER are covered in this section. This method draws features from the tagged corpus. Consequently, word level, morphological, and Lock-up list features were used to determine the outcomes. The outcomes of the supervised method with the word level feature of #NC, #NE, and #CN are displayed in **Table 4**.

Table 4. Results of Word level features.

Name	#NC	#NE	#CN	Precision	Recall	F-measure
Allah	4968	7811	6210	0.636	0.800	0.709
Prophet	3785	6502	5201	0.582	0.728	0.647
Pers	32012	39159	38310	0.817	0.836	0.826
Loc	862	1349	920	0.639	0.937	0.760
Meas	35	147	49	0.238	0.714	0.357
Money	12	139	28	0.086	0.429	0.144
Book	59	183	66	0.322	0.894	0.474
Date	169	596	210	0.284	0.805	0.419
Time	88	102	99	0.863	0.889	0.876
Rlig	120	184	168	0.652	0.714	0.682
Sect	6	17	9	0.353	0.667	0.462
Clan	412	674	521	0.611	0.791	0.690
NatOb	464	670	598	0.693	0.776	0.732
Crime	151	212	179	0.712	0.844	0.772
Para	217	294	298	0.738	0.728	0.733
Hell	186	245	203	0.759	0.916	0.830
Month	55	77	81	0.714	0.679	0.696
Day	19	31	33	0.613	0.576	0.594
Num	10754	13707	11281	0.785	0.953	0.861
Total				0.584	0.772	0.665

Table 4 presents that the Sect entity had the fewest f-measures and that the Time and NE had the greatest. These results are of different patterns, depicted in **Fig. 4**.

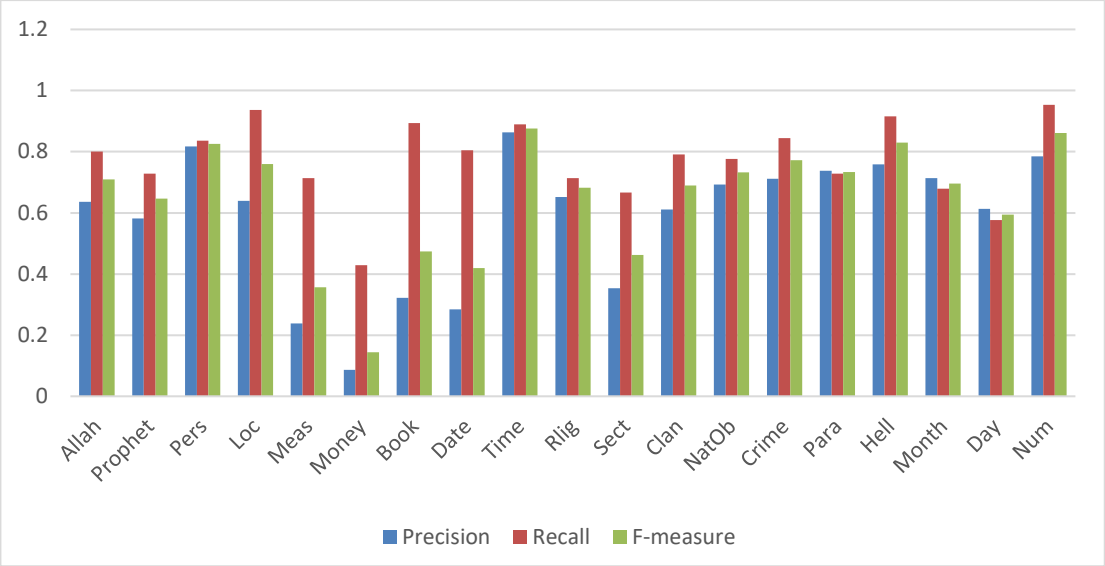

Fig. 4. Evaluation of word level features.

Table 5. Results for Morphological features.

Name	#NC	#NE	#CN	Precision	Recall	F-measure
Allah	3819	7811	4012	0.489	0.952	0.646
Prophet	3652	6502	3981	0.562	0.917	0.697
Pers	30101	39159	39851	0.769	0.755	0.762
Loc	711	1349	930	0.527	0.765	0.624
Meas	45	147	51	0.306	0.882	0.455
Money	29	139	31	0.209	0.935	0.341
Book	61	183	82	0.333	0.744	0.460
Date	155	596	231	0.260	0.671	0.375
Time	82	102	94	0.804	0.872	0.837
Rlig	115	184	171	0.625	0.673	0.648
Sect	6	17	8	0.353	0.750	0.480
Clan	399	674	467	0.592	0.854	0.699
NatOb	416	670	567	0.621	0.734	0.673
Crime	131	212	188	0.618	0.697	0.655
Para	188	294	312	0.639	0.603	0.620
Hell	160	245	225	0.653	0.711	0.681
Month	29	77	93	0.377	0.312	0.341
Day	16	31	33	0.516	0.485	0.500
Num	10321	13707	12365	0.753	0.835	0.792
Total				0.527	0.745	0.617

Table 5 provides the results with regard to the supervised method utilising the morphological features.

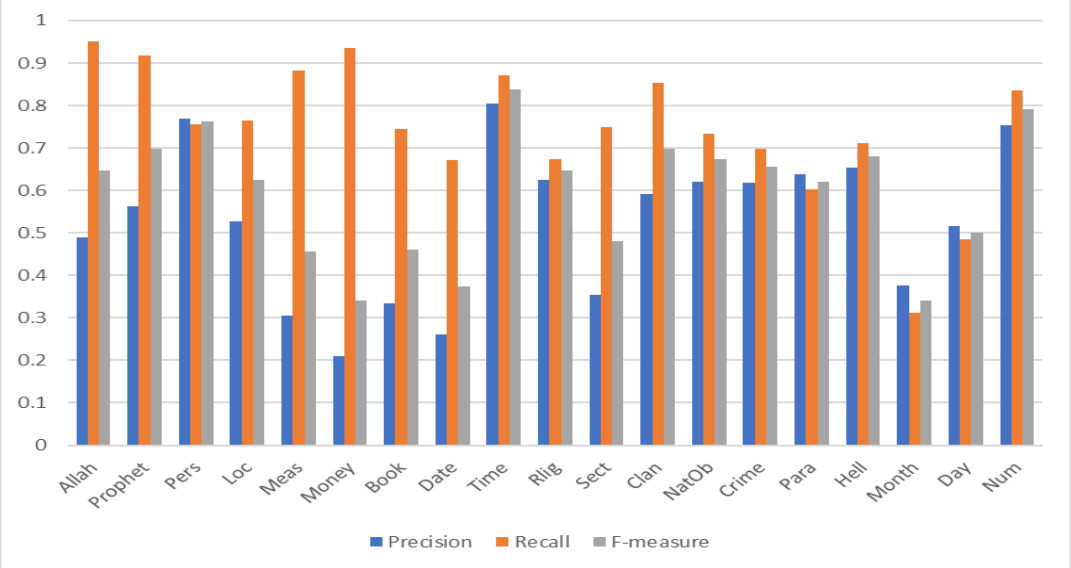

Fig. 5. Results for Morphological features.

The results with regard to the supervised method utilising knowledge-based features may be observed in **Table 6**, and these results are depicted in **Fig. 6**.

Table 6. Results for knowledge-based features.

Name	#NC	#NE	#CN	Precision	Recall	F-measure
Allah	5010	7811	5813	0.641	0.862	0.735
Prophet	3891	6502	5121	0.598	0.760	0.670
Pers	31251	39159	36102	0.798	0.866	0.830
Loc	895	1349	931	0.663	0.961	0.785
Meas	20	147	31	0.136	0.645	0.225
Money	10	139	27	0.072	0.370	0.120
Book	55	183	77	0.301	0.714	0.423
Date	161	596	201	0.270	0.801	0.404
Time	80	102	110	0.784	0.727	0.755
Rlig	139	184	166	0.755	0.837	0.794
Sect	7	17	13	0.412	0.538	0.467
Clan	381	674	532	0.565	0.716	0.632
NatOb	425	670	512	0.634	0.830	0.719
Crime	139	212	160	0.656	0.869	0.747
Para	216	294	301	0.735	0.718	0.726
Hell	191	245	261	0.780	0.732	0.755

Name	#NC	#NE	#CN	Precision	Recall	F-measure
Month	66	77	85	0.857	0.776	0.815
Day	23	31	30	0.742	0.767	0.754
Num	10120	13707	12100	0.738	0.836	0.784
Total				0.586	0.754	0.660

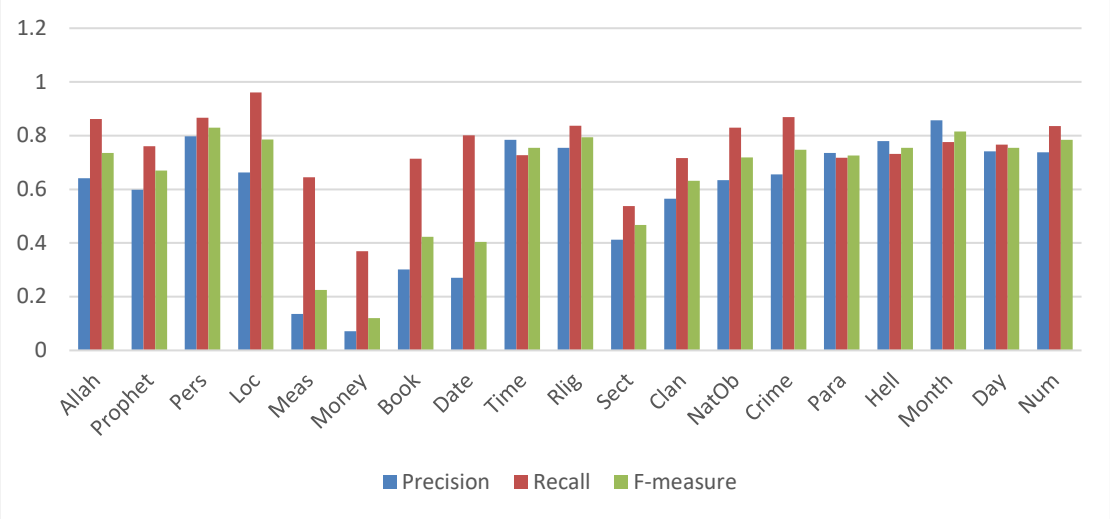

Fig. 6. Results of knowledge-based features.

Combined All Features

The supervised method relies on features. This section made a comparison of the results with regard to each feature. **Fig. 7**, as well as **Table 7**, portray the results from comparing the features.

Table 7. Overall results.

Name	#NC	#NE	#CN	Precision	Recall	F-measure
Allah	6104	7811	6410	0.9523	0.7815	0.858
Prophet	4512	6502	4760	0.9479	0.6939	0.801
Pers	31351	39159	36012	0.8706	0.8006	0.834
Loc	1130	1349	1304	0.8666	0.8377	0.852
Meas	97	147	113	0.8584	0.6599	0.746
Money	94	139	114	0.8246	0.6763	0.743
Book	124	183	141	0.8794	0.6776	0.765
Date	340	596	371	0.9164	0.5705	0.703
Time	92	102	94	0.9787	0.9020	0.939
Rlig	140	184	171	0.8187	0.7609	0.789
Sect	14	17	19	0.7368	0.8235	0.778
Clan	468	674	497	0.9416	0.6944	0.799
NatOb	469	670	597	0.7856	0.7000	0.740

Name	#NC	#NE	#CN	Precision	Recall	F-measure
Crime	151	212	179	0.8436	0.7123	0.772
Para	236	294	281	0.8399	0.8027	0.821
Hell	191	245	200	0.9550	0.7796	0.858
Month	66	77	80	0.8250	0.8571	0.841
Day	23	31	28	0.8214	0.7419	0.780
Num	10854	13707	11272	0.9629	0.7919	0.869
Total				0.8750	0.7507	0.805

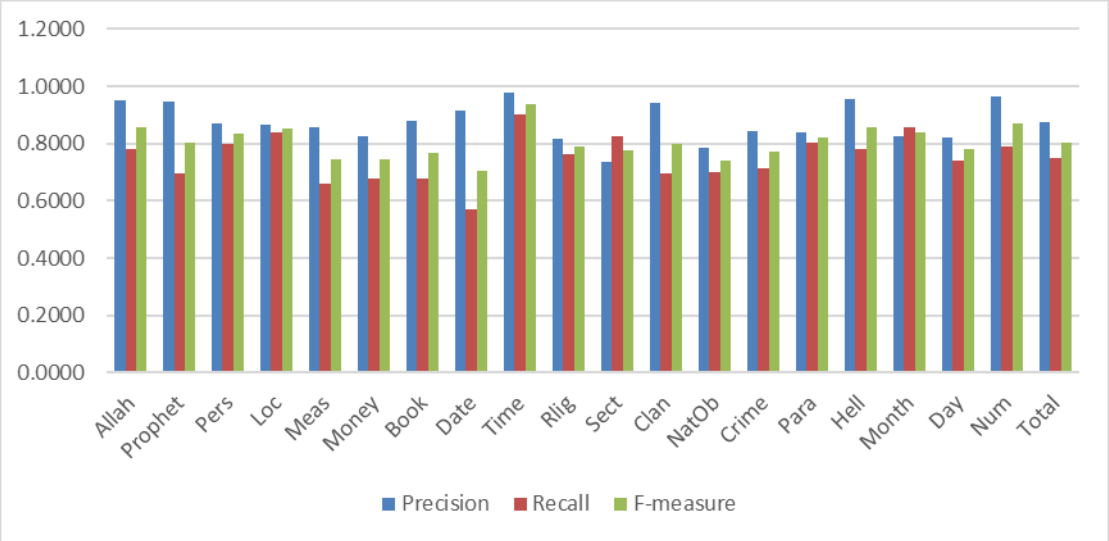

Fig. 7. Overall results.

These results highlight the strengths of the model in handling specific named entities while also pointing out areas where further refinement is needed. Future work could focus on addressing the identified weaknesses by enriching the dataset for the underperforming categories, fine-tuning the feature extraction process, or exploring alternative machine learning algorithms that might offer better performance for these more challenging categories. **Fig. 8** provides a comparison of different feature extraction approaches—Word level, Morphological, Knowledge-based, and a combined Overall method—across the same set of entity categories. This figure underscores the importance of an integrated approach, as the Overall method consistently outperforms individual feature types across most categories. The comparison illustrates that while individual features contribute valuable information, their combination yields the most effective results in terms of Precision, Recall, and F-measure.

Together, these figures and tables offer a comprehensive overview of the model's performance, revealing both its strengths and areas for improvement. Future work could focus on enhancing the model's capabilities in underperforming categories by refining the feature extraction process or incorporating more sophisticated machine learning techniques.

This analysis underscores the overall effectiveness of the model in handling a wide range of named entities in Classical Arabic, while also identifying specific areas for potential improvement in future iterations.

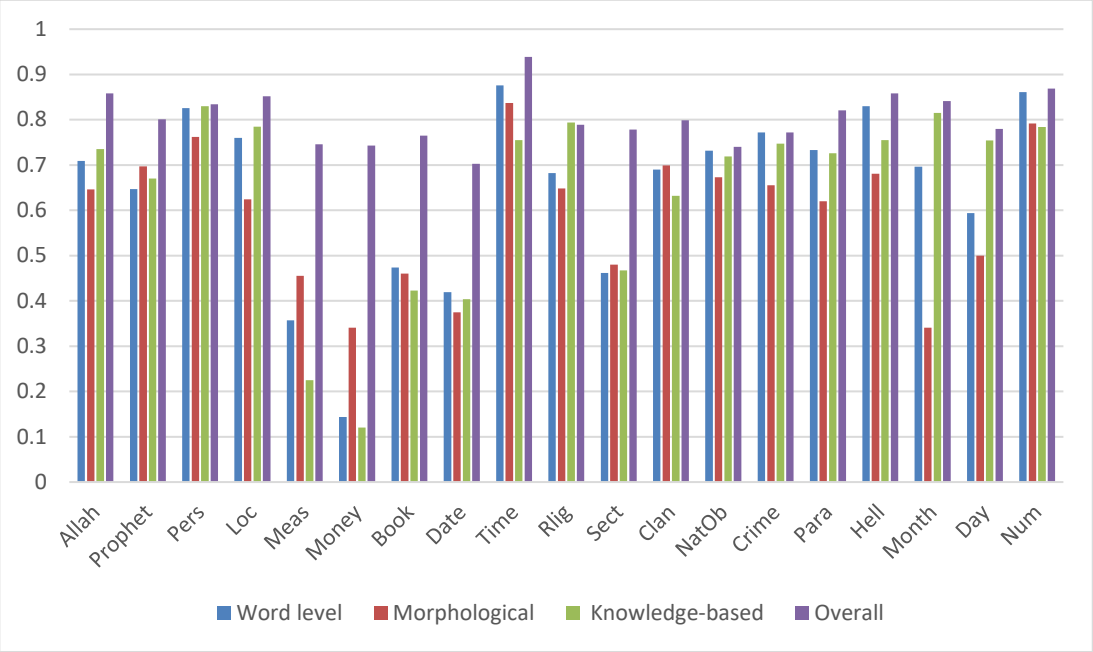

Fig. 8. Compare between features.

5. Conclusion

The characteristics were explained, the method suggested for this study was provided, and the supervised ANER method was explored in this work. For Arabic NER, the supervised method was utilised to assess the impact of various feature combinations. The precision, recall, as well as F-measure for the NB method, were 86%, 75%, and 80%, respectively. Given the complexity of the language, NER of classical Arabic in NLP is a difficult job. However, ML algorithms have been discovered to be useful in obtaining high accuracy. Future studies should focus on a number of topics, including utilising more sophisticated ML methods, using larger and more varied annotated corpora, and enhancing how classical Arabic's complexity is handled. It is anticipated that NER for classical Arabic will continue to perform better as NLP and ML technologies evolve. To sum up, ML has been demonstrated to be a potent tool in NLP for classical Arabic, with a range of methods being used for diverse tasks. Additional study is required to maintain the performance of these algorithms and to investigate new possibilities of ML in NLP for classical Arabic.

One of the key contributions of this study is the identification of effective feature sets for NER in Classical Arabic. Future research can build on these findings by exploring more advanced machine learning models, such as deep learning approaches, which could potentially offer even greater accuracy by capturing more complex patterns within the text. Additionally, expanding the dataset to include a wider variety of Classical Arabic texts, such as historical documents and literary works, could enhance the generalizability of the model. Another promising direction is the development of semi-supervised or unsupervised learning methods, which could leverage unannotated data to further improve the performance of NER systems in resource-scarce languages like Classical Arabic.

The implications of this research extend to various practical applications. For instance, the improved NER system can be integrated into search engines and digital libraries that handle

Classical Arabic texts, enabling more accurate information retrieval. Moreover, this system could be employed in the automatic translation of Classical Arabic texts, enhancing the accuracy of entity recognition in machine translation outputs. The methodology developed in this study could also be adapted for use in other languages with similarly complex morphological structures, broadening its applicability.

This study advances the current state-of-the-art in NER for Classical Arabic by introducing a supervised machine learning approach that effectively handles the complexities of this language. Unlike existing methods that often focus on Modern Standard Arabic, our approach integrates word-level, morphological, and knowledge-based features specifically tailored for Classical Arabic. This has resulted in better performance across various entity types, especially in challenging categories, making our model more robust and accurate than current models.

Based on the results and limitations identified in this study, several specific directions for future work are suggested:

i. Incorporating Deep Learning Models: Future studies could investigate the use of deep learning models, such as Recurrent Neural Networks (RNNs) or Transformers, to capture more intricate dependencies within the text and potentially improve NER performance.
ii. Expanding the Dataset: Increasing the size and diversity of the dataset by including more types of Classical Arabic texts could help in building more robust NER systems. Collaborative efforts to create larger annotated corpora for Classical Arabic would also be beneficial.
iii. Exploring Cross-Lingual Approaches: Considering the similarities between Arabic and other Semitic languages, cross-lingual transfer learning approaches could be explored to leverage annotated data from related languages, further enhancing NER in Classical Arabic.
iv. Real-Time NER Systems: Developing real-time NER systems that can process Classical Arabic in dynamic environments, such as social media or live translations, could open up new avenues for the application of this technology.
v. Addressing Data Sparsity: Future work could focus on techniques to mitigate data sparsity, such as data augmentation strategies or the use of domain adaptation methods, to improve the robustness of NER models in low-resource settings.

Acknowledgement

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/214/45.

References

[1] Nadeau, D., and S. Sekine, “A survey of named entity recognition and classification,” *Lingvisticae Investigationes*, vol.30, no.1, pp.3-26, 2007. Article (CrossRef Link)

[2] Salminen, J. et al., “Developing an online hate classifier for multiple social media platforms,” *Human-centric Computing and Information Sciences*, vol.10, no.1, pp.1-34, 2020. Article (CrossRef Link)

[3] Salah, R.E., and L.Q. binti Zakaria, “A Comparative Review of Machine Learning for Arabic Named Entity Recognition,” *International Journal on Advanced Science, Engineering and Information Technology*, vol.7, no.2, pp.511-518, 2017. Article (CrossRef Link)

[4] Silalahi, S., T. Ahmad, and H. Studiawan, “Transformer-Based Named Entity Recognition on Drone Flight Logs to Support Forensic Investigation,” *IEEE Access*, vol.11, pp.3257-3274, 2023. Article (CrossRef Link)

[5] Li, J. et al., "A Survey on Deep Learning for Named Entity Recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol.34, no.1, pp.50-70, 2022. Article (CrossRef Link)
[6] Taquini, R., K. R. Finardi, and G. B. Amorim, "English as a Medium of Instruction at Turkish State Universities," *Education and Linguistics Research*, vol.3, no.2, pp.35-53, 2017. Article (CrossRef Link)
[7] McEntee-Atalianis, L., and R. Vessey, "Mapping the language ideologies of organisational members: a corpus linguistic investigation of the United Nations' General Debates (1970–2016)," *Language Policy*, vol.19, no.4, pp.549-573, 2020. Article (CrossRef Link)
[8] Salah, R. E. and L. Q. B. Zakaria, "Arabic Rule-Based Named Entity Recognition Systems Progress and Challenges," *International Journal on Advanced Science, Engineering and Information Technology*, vol.7, no.3, pp.815-821, 2017. Article (CrossRef Link)
[9] Mohd, M. et al., "Quranic Optical Text Recognition Using Deep Learning Models," *IEEE Access*, vol.9, pp.38318-38330, 2021. Article (CrossRef Link)
[10] AbdelRahman, S. et al., "Integrated Machine Learning Techniques for Arabic Named Entity Recognition," *IJCSI International Journal of Computer Science Issues*, vol.7, no.4, pp.27-36, 2010. Article (CrossRef Link)
[11] Benajiba, Y., P. Rosso, and J. M. Benedíruiz, "ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy," in *Proc. of 8th International Conference on Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, vol.4394, pp.143-153, Springer, 2007. Article (CrossRef Link)
[12] Komariah, K. S. et al., "SMPT: A Semi-Supervised Multi-Model Prediction Technique for Food Ingredient Named Entity Recognition (FINER) Dataset Construction," *Informatics*, vol.10, no.1, 2023. Article (CrossRef Link)
[13] Benajiba, Y. and P. Rosso, "Arabic Named Entity Recognition using Conditional Random Fields," in *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, vol.8, pp.143-153, 2008. Article (CrossRef Link)
[14] Al-Twairesh, N., H. Al-Khalifa, and A. Al-Salman, "AraSenTi: Large-Scale Twitter-Specific Arabic Sentiment Lexicons," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, vol.1, pp.697-705, 2016. Article (CrossRef Link)
[15] Abdul-Hamid, A. and K. Darwish, "Simplified Feature Set for Arabic Named Entity Recognition," in *Proc. of the 2010 Named Entities Workshop, Association for Computational Linguistics 2010*, pp.110-115, 2010. Article (CrossRef Link)
[16] AbdelRahman, S. et al., "Integrated Machine Learning Techniques for Arabic Named Entity Recognition," *International Journal of Computer Science Issues (IJCSI)*, vol.7, no.4, pp.27-36, 2010. Article (CrossRef Link)
[17] Bidhend, M. A., B. Minaei-Bidgoli, and H. Jouzi, "Extracting person names from ancient Islamic Arabic texts," in *Proc. of Language Resources and Evaluation for Religious Texts (LRE-Rel) Workshop Programme, Eight International Conference on Language Resources and Evaluation (LREC 2012)*, 2012. Article (CrossRef Link)
[18] Morsi, A. and A. Rafea, "Studying the impact of various features on the performance of Conditional Random Field-based Arabic Named Entity Recognition," in *Proc. of 2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pp.1-5, 2013. Article (CrossRef Link)
[19] Zirikly, A. and M. Diab, "Named Entity Recognition for Dialectal Arabic," in *Proc. of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP 2014)*, pp.78-86, 2014. Article (CrossRef Link)
[20] Alduailaj, A. M. and A. Belghith, "Detecting Arabic Cyberbullying Tweets Using Machine Learning," *Machine Learning and Knowledge Extraction*, vol.5, no.1, pp.29-42, 2023. Article (CrossRef Link)
[21] Alsayadi, H. A. and A. M. ElKorany, "Integrating Semantic Features for Enhancing Arabic Named Entity Recognition," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol.7, no.3, pp.128-136, 2016. Article (CrossRef Link)

[22] Dahan, F., A. Touir, and H. Mathkour, "First Order Hidden Markov Model for Automatic Arabic Name Entity Recognition," *International Journal of Computer Applications*, vol.123, no.7, pp.37-40, 2015. Article (CrossRef Link)
[23] Al-Shoukry, S. and N. Omar, "Proper Nouns Recognition in Arabic Crime Text Using Machine Learning Approach," *Journal of Theoretical and Applied Information Technology*, vol.79, no.3, pp.506-513, 2015. Article (CrossRef Link)
[24] Aoumeur, N. E., Z. Li, and E. M. Alshari, "Improving the Polarity of Text through word2vec Embedding for Primary Classical Arabic Sentiment Analysis," *Neural Processing Letters*, vol.55, pp.2249-2264, 2023. Article (CrossRef Link)
[25] Koulali, R. and A. Meziane, "A contribution to Arabic Named Entity Recognition," in *Proc. of 2012 10th International Conference on ICT and Knowledge Engineering*, pp.46-52, 2012. Article (CrossRef Link)
[26] Mohammed, N. F. and N. Omar, "Arabic Named Entity Recognition Using Artificial Neural Network," *Journal of Computer Science*, vol.8, no.8, pp.1285-1293, 2012. Article (CrossRef Link)
[27] Alanazi, S., "A Named Entity Recognition System Applied to Arabic Text in the Medical Domain," *Doctoral thesis*, Staffordshire University, 2017. Article (CrossRef Link)
[28] Al-Ayyoub, M. et al., "Deep learning for Arabic NLP: A survey," *Journal of Computational Science*, vol.26, pp.522-531, 2018. Article (CrossRef Link)
[29] Kanan, T. et al., "A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media," in *Proc. of 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pp.622-628, 2019. Article (CrossRef Link)
[30] Imene, S. and A. Hassina, "An Unsupervised Semantic Model for Arabic/French Terminology Extraction," in *Proc. of International Conference on Emerging Technologies and Intelligent Systems: ICETIS 2021 Volume 2*, Lecture Notes in Networks and Systems, vol.322, pp.49-59, Springer, 2022. Article (CrossRef Link)
[31] Al-Laith, A. et al., "AraSenCorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus," *Applied Sciences*, vol.11, no.5, 2021. Article (CrossRef Link)
[32] Kartchner, D. et al., "Rule-Enhanced Active Learning for Semi-Automated Weak Supervision," *AI*, vol.3, no.1, pp.211-228, 2022. Article (CrossRef Link)
[33] Shaalan, K. and H. Raza, "Arabic Named Entity Recognition from Diverse Text Types," in *Proc. of 6th International Conference, Advances in Natural Language Processing*, Lecture Notes in Computer Science, vol.5221, pp.440-451, Springer, 2008. Article (CrossRef Link)
[34] Shaalan, K. and H. Raza, "Person Name Entity Recognition for Arabic," in *Proc. of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Association for Computational Linguistics, pp.17-24, 2007. Article (CrossRef Link)
[35] Appelt, D. E. and D. J. Israel, "Introduction to Information Extraction Technology," in *Proc. of Tutorial prepared for the IJCAI Conference*, 1999. Article (CrossRef Link)
[36] Eikvil, L., Information Extraction from World Wide Web-A Survey, 1999. Article (CrossRef Link)
[37] Al-Ayyoub, M. et al., "A comprehensive survey of arabic sentiment analysis," *Information Processing & Management*, vol.56, no.2, pp.320-342, 2019. Article (CrossRef Link)
[38] Salah, R. E. and L. Q. B. Zakaria, "Building the Classical Arabic Named Entity Recognition Corpus (CANERCorpus)," in *Proc. of 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pp.1-8, 2018. Article (CrossRef Link)
[39] Dash, M. and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol.1, no.1-4, pp.131-156, 1997. Article (CrossRef Link)
[40] Farghaly, A. and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol.8, no.4, pp.1-22, 2009. Article (CrossRef Link)
[41] Mohammad, A.-S. et al., "Gated recurrent unit with multilingual universal sentence encoder for Arabic aspect-based sentiment analysis," *Knowledge-Based Systems*, vol.261, 2023. Article (CrossRef Link)

[42] Wang, X. and J. Liu, "A novel feature integration and entity boundary detection for named entity recognition in cybersecurity," *Knowledge-Based Systems*, vol.260, 2023. Article (CrossRef Link)
[43] Guo, X. et al., "CG-ANER: Enhanced contextual embeddings and glyph features-based agricultural named entity recognition," *Computers and Electronics in Agriculture*, vol.194, 2022. Article (CrossRef Link)
[44] Sun, M. et al., "Learning the Morphological and Syntactic Grammars for Named Entity Recognition," *Information*, vol.13, no.2, 2022. Article (CrossRef Link)
[45] Alotaibi, F. S. et al., "Keyphrase Extraction Using Enhanced Word and Document Embedding," *IETE Journal of Research*, vol.69, no.12, pp.8876-8888, 2022. Article (CrossRef Link)
[46] Wei, H. et al., "Named Entity Recognition From Biomedical Texts Using a Fusion Attention-Based BiLSTM-CRF," *IEEE Access*, vol.7, pp.73627-73636, 2019. Article (CrossRef Link)
[47] Meselhi, M. A. et al., "A Novel Hybrid Approach to Arabic Named Entity Recognition," in *Proc. of 10th China Workshop on Machine Translation, Communications in Computer and Information Science*, vol.493, pp.93-103, Springer, 2014. Article (CrossRef Link)
[48] Salah, R. E. and L. Q. binti Zakaria, "Arabic Rule-Based Named Entity Recognition Systems Progress and Challenges," *International Journal on Advanced Science Engineering and Information Technology*, vol.7, no.3, pp.815-821, 2017. Article (CrossRef Link)
[49] Shaalan, K. and M. Oudah, "A hybrid approach to Arabic named entity recognition," *Journal of Information Science*, vol.40, no.1, pp.67-87, 2014. Article (CrossRef Link)
[50] Abdallah, S., K. Shaalan, and M. Shoaib, "Integrating Rule-Based System with Classification for Arabic Named Entity Recognition," in *Proc. of 13th International Conference on Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, vol.7181, pp.311-322, Springer, 2012. Article (CrossRef Link)
[51] Boujelben, I., S. Jamoussi, and A. Ben Hamadou, "A hybrid method for extracting relations between Arabic named entities," *Journal of King Saud University - Computer and Information Sciences*, vol.26, no.4, pp.425-440, 2014. Article (CrossRef Link)
[52] Oudah, M. M., "Integrating Rule-based Approach and Machine learning Approach for Arabic Named Entity Recognition," *The British University in Dubai (BUiD)*, 2012. Article (CrossRef Link)
[53] Pasha, A. et al., "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," in *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp.1094-1101, 2014. Article (CrossRef Link)
[54] Habash, N., O. Rambow, and R. Roth, "MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in *Proc. of the 2nd international conference on Arabic language resources and tools*, Cairo, Egypt, 2009. Article (CrossRef Link)
[55] Farber, B. et al., "Improving NER in Arabic Using a Morphological Tagger," in *Proc. of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008. Article (CrossRef Link)
[56] Shaalan, K., "A survey of arabic named entity recognition and classification," *Computational Linguistics*, vol.40, no.2, pp.469-510, 2014. Article (CrossRef Link)
[57] Saif, A., M. J. Ab Aziz, and N. Omar, "Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features," *Natural Language Engineering*, vol.23, no.1, pp.53-91, 2017. Article (CrossRef Link)
[58] Zirikly, A. and M. Diab, "Named entity recognition for arabic social media," in *Proc. of NAACL-HLT 2015*, pp.176-185, 2015. Article (CrossRef Link)
[59] Althobaiti, M., U. Kruschwitz, and M. Poesio, "Combining Minimally-supervised Methods for Arabic Named Entity Recognition," *Transactions of the Association for Computational Linguistics*, vol.3, pp.243-255, 2015. Article (CrossRef Link)
[60] Benajiba, Y. and P. Rosso, "ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information," in *Proc. of 3rd Indian International Conference on Artificial Intelligence (IICAI-07)*, pp.1814-1823, 2007. Article (CrossRef Link)

[61] Benajiba, Y., P. Rosso, and J. M. Benedíruiz, "Anersys: An arabic named entity recognition system based on maximum entropy," in *Proc. of 8th International Conference on Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, vol.4394, pp.143-153, Springer, 2007. Article (CrossRef Link)

[62] Chi, W. W., T. Y. Tang, N. M. Salleh, M. Mukred, H. AlSalman, and M. Zohaib, "Data Augmentation With Semantic Enrichment for Deep Learning Invoice Text Classification," *IEEE Access*, vol.12, pp.57326-57344, 2024. Article (CrossRef Link)

[63] Moussaoui, T. E., C. Loqman, and J. Boumhidi, "Flat and Nested Named Entity Recognition in Arabic Language," in *Proc. of 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pp.1-7, May 2024. Article (CrossRef Link)

[64] Mekki, A., I. Zribi, M. Ellouze, and L. H. Belguith, "Named Entity Recognition of Tunisian Arabic Using the Bi-LSTM-CRF Model," *International Journal on Artificial Intelligence Tools*, vol.33, no.02, 2024. Article (CrossRef Link)

[65] Qarah, F., and T. Alsanoosy, "A Comprehensive Analysis of Various Tokenizers for Arabic Large Language Models," *Applied Sciences*, vol.14, no.13, 2024. Article (CrossRef Link)

Dr. Ramzi Salah received the PHD degree from Universiti Kebangsaan Malaysia (UKM). He is currently a project manager with RMZTECH. His current research interests include named entity recognition and speech recognition.

Dr. Muaadh Mukred currently works as a lecturer at the Department of Business Analytics at Sunway Business School; he is also an associated fellow at the Cyber Security Center at the Information Science and Technology faculty, University Kebangsaan Malaysia (UKM), Malaysia. Muaadh has been a post-doctoral researcher at the Cyber Security Research Center in the Information Science and Technology Faculty. Muaadh has worked for over four years as a software developer at the Al-Noor Foundation, Putrajaya. Previously, he was a lecturer at the Computer Science Department at Sana'a Community College for over eight years. Muaadh completed his first degree in Computer Science at AL-Mustansiriyah University in 2002. Besides his permanent position at Sana'a Community College, Muaadh has joined the industry for five years, working in Computer Science, data analysis, and programming. Afterwards, Muaadh moved to Malaysia after getting a scholarship, and he earned his MSc in Computer Science from the University of Technology of Malaysia in 2009. Muaadh joined SCC again in 2011, where he started a new position as a head of the Higher Professional Education Division and served as a lecturer. In 2013, Muaadh got another scholarship and moved to Malaysia, where he was awarded a PhD from the UKM and Outstanding Researcher award as one of the best students in his batch. Muaadh has also been awarded an outstanding publication. He has published several scientific/research papers in well-known international journals and conferences. Muaadh was involved in some research grant projects, some in artificial intelligence and machine learning, and others in analyzing and architecting big data. He has recently been involved as an academic partner with some research projects funded by some universities in Malaysia and the Middle East. He has supervised some PhD and master students and some other undergraduate final-year projects. Muaadh has also been invited as an external examiner for many PhD and Master's students and served as a reviewer for many high-impact journals. His research interests include big data analytics, artificial intelligence, natural language processing, technology adoption, and database architecture.

Dr. Lailatul Qadri binti Zakaria received the bachelor's and master's degrees in information science from the Universiti Kebangsaan Malaysia, and the Ph.D. degree from the University of Southampton, U.K. She is currently a Lecturer in natural language processing with the Faculty of Information Science and Technology, University Kebangsaan Malaysia. Her research interests include natural language processing, ontology development, and semantic web technologies.

Dr. Fuad A. M. Al-Yarimi received his Ph.D. degree in Computer Science from Jawaharlal Nehru University, New Delhi, India, 2014. He is currently an associated professor in Computer Science at King Khalid University, KSA. He is the author of several articles. His research interests include Data Mining, Machine learning, Artificial intelligence, and Information privacy and security, and more recently cloud computing.