IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5810~5821
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5810-5821

Arabic data augmentation approaches for enhanced named entity recognition

Brahim Ghazoui[1], Ismail El Bazi[1], Brahim Ait Ben Ali[2], Hicham Moussa[1]
[1]LGS Laboratory, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Beni Mellal, Morocco
[2]Laboratory of Mathematics, Artificial Intelligence and Sustainable Technologies, Cadi Ayyad University, Marrakesh, Morocco

Article Info

Article history:

Received Apr 11, 2025
Revised Aug 8, 2026
Accepted Aug 31, 2026

Keywords:

Arabic named entity recognition
Arabic natural language processing
Data augmentation
Low-resource languages
Transformer-based models

ABSTRACT

Named entity recognition (NER) is a basic task in natural language processing (NLP) that is used in information retrieval, question answering, and text summarization. Although there have been recent developments in NER for high-resource languages, Arabic NER remains a complex problem due to the language's rich morphology, dialect variation, and the lack of large annotated corpora. To address these difficulties, the present paper proposes an innovative Arabic data augmentation model that combines three mutually supportive methods: synonym replacement, deletion of person names, and random selection of person names. The suggested augmentation methods are tested on popular Arabic NER datasets, such as the news and X (formerly Twitter) corpora, and combined with the tree-based pipeline optimization tool (TPOT) to optimize the learning pipeline in an automated way. The experiment's outcomes indicate a steady increase in precision, recall, and F1-score across all entity types, especially in informal and low-resource text. Such results demonstrate the importance of well-conceptualized data augmentation for Arabic NER and help advance Arabic NLP research.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Brahim Ghazoui
LGS Laboratory, Faculty of Sciences and Techniques, Sultan Moulay Slimane University
Beni Mellal 23000, Morocco
Email: brahim.ghazoui@usms.ac.ma

1. INTRODUCTION

Named entity recognition (NER) is a fundamental natural language processing (NLP) task that identifies and categorizes named entities in text, such as people, places, and organizations. It is a key component in many downstream tasks, such as information retrieval, question answering, and text summarization [1]–[4]. Although NER systems have demonstrated exemplary performance across several high-resource languages, they have not been effective in low-resource, morphologically complex languages like Arabic.

Arabic NER is a distinctly different task that poses difficulties unlike those in other languages. Arabic has a rich and complicated morphology and permits words to have numerous surface realizations based on an underlying root, thereby raising lexical scarcity and ambiguity [5], [6]. Moreover, the lack of capitalization, the common omission of diacritics, and the coexistence of modern standard Arabic with many dialects also make entity boundary detection and classification challenging [4], [7]. These language attributes significantly influence the strength and generalization ability of Arabic NER systems, especially when applied to informal or domain-heterogeneous text such as social media posts.

Another significant constraint in Arabic NER is the scarcity of high-quality, large-scale datasets for annotation. Labeling corpora is an expensive process that takes time and limits the usage of training data,

harming model performance [5], [8]. This leads to overfitting and poor generalization in many Arabic NER models, especially when applied in real-world settings where domain knowledge is unknown or the text is noisy [7]. Data sparsity is thus a crucial need for enhancing Arabic NER systems.

Data augmentation has become one of the most successful methods for reducing data sparsity, achieved by generating synthetic training samples from already annotated corpora. Synonym replacement, paraphrasing, and back-translation are data augmentation methods used in NLP to enhance model robustness and generalization by increasing data diversity [9]–[11]. In Arabic NER, augmentation could be used to expose models to richer sources of lexical, morphological, and contextual variation, thereby improving their ability to identify entities across varying linguistic contexts [7], [12].

However, using data augmentation on Arabic text entails other drawbacks. The fact that Arabic is a context-sensitive language and the richness of Arabic morphology increases the likelihood of semantic drift, grammatical inconsistencies, and the production of noisy data, which can bias model training [5], [7]. Strategies such as synonym replacement and paraphrasing should also be applied loosely to maintain sentence meaning, syntactic proficiency, and the labels of entities [12], [13]. These difficulties indicate that Arabic-specific augmentation techniques that account for Arabic's linguistic peculiarities are necessary.

This paper suggests a custom Arabic data augmentation system for NER. The proposed method combines three complementary approaches: semantic validation via synonym replacement, person name deletion, and random person name replacement. These methods are implemented on popular Arabic NER datasets and tested using the tree-based pipeline optimization tool (TPOT), an automated machine learning tool based on predictive modeling that optimizes model pipelines and hyperparameters [14], [15]. This study will improve Arabic NER systems by combining linguistically informed data augmentation and pipeline optimization, thereby enhancing their performance, robustness, and generalization. In general, this study contributes to the development of Arabic NLP by methodically addressing data and linguistic sparsity through efficient augmentation strategies. The findings indicate that carefully crafted Arabic-specific data augmentation could significantly boost NER performance across domains and provide valuable insights for future studies of low-resource, morphologically rich languages.

2. OVERVIEW OF ARABIC DATA AUGMENTATION: TECHNIQUES AND CHALLENGES

Data augmentation is a very important method of machine learning, especially in NLP. The fundamental principle of data augmentation is to create an artificially expanded dataset by generating new data points from an existing dataset [16]. It is done through different techniques; you transform the data to a different level, or the original data, without losing the main structure and meaning. The ultimate goal of data augmentation is to improve the robustness and performance of machine learning models [17] by introducing diversity and variety to the training dataset.

Data augmentation plays an essential role in NLP since, in this case, the model can overfit [18], meaning that a good result may be achieved on the training data but not necessarily on previously unseen data. Models generalize better with a more varied dataset during training and thus are ideal for any practical application. In NLP, the most common data augmentation methods are synonym replacement, back-translation, text shuffle, and paraphrasing [11]. Each of these techniques attempts to perturb the text in some way to generate different lexical versions of the same input so that the model can learn from various language usage scenarios [19].

Arabic NLP adds a new level of complexity with regard to data augmentation because of the unique attributes of the Arabic language [7]. The morphology of Arabic [7] is rich, and it enables the construction of a wide range of words by combining roots and patterns. This richness in the language, although one of the most significant features of the expressiveness of the language, makes it difficult for traditional data augmentation techniques to be applied [7].

2.1. Challenges of data augmentation in Arabic natural language processing

Despite the effectiveness of data augmentation across various NLP tasks, its application to Arabic text poses language-specific challenges. Arabic is a morphologically rich language in which a single root word can produce many surface forms, making augmentation more ambiguous and naive methods unreliable [5], [7]. Fundamental transformations can, in unintended ways, change grammar or distort meaning, thus adding noise that adversely impacts model learning.

A significant problem is semantic drift, in which augmented sentences no longer retain their original meaning. This is especially evident in Arabic because of the heavy reliance on context and word order for semantic interpretation [12], [13]. Synonym replacement and paraphrasing are easy-to-use techniques that can substitute a word with a contextually inappropriate one, resulting in grammatically correct but semantically misguided sentences. These distortions are particularly unfriendly to NER, which requires precise contextual information to classify entities correctly [5].

The second challenge is the need to be morphologically and syntactically consistent. Arabic words represent gender, number, definiteness, and case, and any augmentation method should not violate these grammatical restrictions to ensure the resulting text is linguistically correct [20]. Furthermore, capitals do not exist in Arabic, eliminating one of the helpful hints that NER systems tend to utilize with Latin-script languages [6]. All of these reasons make Arabic NER extremely sensitive to poorly designed augmentation strategies.

Moreover, Arabic NER data augmentation must include entity-specific constraints. The length, structure, and cultural form vary greatly with named entities, especially personal names [21]. Uncontrolled or random changes to entity spans can cross entity boundaries or invalidate annotation integrity, decreasing the accuracy of the augmented data [7]. Thus, the augmentation techniques should be entity-sensitive and annotation-conserving.

2.2. Synonym replacement

As an example in relation to Arabic, data augmentation approaches such as synonym replacement need to be utilized very carefully, as the word is a complex language, and its meaning can change according to context [13]. Arabic is not straightforward for such a use case of word-for-word replacement, as it has complex morphology [22] and, therefore, needs to be treated with semantics [23]. The reason is mostly due to the Arabic language being syntactically different; the meaning of words changes dramatically from context to context and usage to usage. Further, Arabic has a rich morphology, such as gender, number, and case; all these require grammatical parallelism of the synonyms apart from the semantic parallelism in the case of synonyms [24]. Thus, these are not only semantic but also syntactic components, and therefore, the development of appropriate synonym selection must be carried out with respect to the semantics and syntax in that language [7]. Additionally, the polysemy [7] of the Arabic words, where the different meanings of the word change following the context, makes the synonym selection even more complicated to handle. To tackle these issues, advanced algorithms that can understand the context and preserve the text structure to produce augmented data that is both linguistically correct [25] and beneficial for NER. Such adoption, where the application of the replacement is carefully performed, is important in improving the reliability and accuracy of Arabic NER systems.

2.3. Back-translation

This technique involves translating text to another language and then back to the original language [11], [13]. While effective in generating syntactic variations, it struggles with Arabic's complex sentence structures and can result in a loss of meaning or context. Therefore, additional verification steps are often necessary to maintain coherence in the augmented data.

2.4. Paraphrasing

Creating paraphrased [26] versions of Arabic sentences is challenging due to the language's rich idiomatic expressions and nuanced meanings [24]. Automated paraphrasing tools must navigate these subtleties to maintain the original sentence's intent and meaning. As a result, developers often incorporate linguistic knowledge bases or advanced neural models to avoid distorting context.

2.5. Morphological variations

A good example of this can work with Arabic, which has a very rich morphology, allowing the use of words of the same root in different forms for data augmentation. However, this approach requires a high and detailed understanding of Arabic morphology [20] to create grammatically and semantically correct [20] morphological forms. Consequently, implementing morphological variation demands robust linguistic expertise to avoid generating incoherent text.

Existing Arabic NER augmentation techniques suffer from limitations such as the generation of noisy data [7], which may mislead model training, and the lack of linguistic coherence in the augmented data. These challenges necessitate domain-specific augmentation techniques that can handle Arabic-specific morphology and syntax. As a result, researchers must carefully design and validate methods that align with the language's structural nuances.

After the above challenges, the Arabic data augmentation process is improved in three adjusted methods (synonym replacement, person name deletion, and person name random replacement); the newly augmented dataset will be utilized to be evaluated by a TPOT [14], [27]. Nonetheless, parameters such as the n-best candidates adjusted to avoid the multiplier cointegration and the parameters optimized for parallel synthetic means are related to augmentation methods when Arabic NER is the core functions of the model created by TPOT. This study makes a significant contribution to the existing literature on Arabic NER

by proposing effective Arabic data augmentation techniques which are tailored to the challenges posed by processing Arabic text and molecularly, a valuable contribution to general NLP data augmentation techniques.

2.6. How the proposed approach addresses these challenges

To address the above obstacles, this paper presents a linguistically informed entity-conscious data augmentation system designed explicitly for Arabic NER. First, synonym replacement is performed using a large Arabic synonym vocabulary. It is accompanied by semantic validation via a BERT-based similarity model [28], [29]. This move will make augmented sentences semantically equal to the original text, thereby reducing semantic drift.

Second, the suggested methods of person name deletion and random person name substitution address the problem of variability in named entities directly while maintaining consistency in annotations. The fact that they work directly on B-PERS and I-PERS tags ensures that entities are bound and produce realistic variations of personal names without altering the sentence structure or meaning. The entity-conscious design also enables the model to acquire strong definitions of personal names across various contexts and name forms [21].

Lastly, the approach's effectiveness can be further improved by combining the augmented datasets with TPOT. TPOT is also a method that automatically investigates the best preprocessing step combinations, feature representations, and learning algorithms, thereby making the model more adaptable to the increased diversity induced by augmentation [14], [15]. The combination of these Arabic-specific augmentations and automated pipeline performance improves generalization and performance for both formal and informal Arabic text.

Overall, the suggested framework will address the semantic, linguistic, and structural issues of Arabic data augmentation directly. The method maintains semantic integrity and morphological and entity consistency. It offers a robust, scalable solution for improving Arabic NER in low-resource environments.

3. TREE-BASED PIPELINE OPTIMIZATION TOOL

3.1. Introduction to tree-based pipeline optimization tool

The TPOT [14] is a notable contribution to automatic machine learning (AutoML) research [30]. TPOT has been developed as an open-source tool that can automate and optimize the machine learning pipeline-building process through an intelligent exploration of thousands of possible pipeline configurations to identify the best one on a given dataset. MLGPE genetic pipeline evolution: pipeline evolution of pipeline parts with genetic programming [MLGPE] designs and creates clever optimizations to make the best use of ML algorithms while handling the pipeline structures as a series of operations.

What really makes TPOT shine is that it applies an iterative process to test all possible pipelines, where this terminology means a different configuration of preprocessing techniques, a different hyperparameter configuration, or a different structure on the machine learning models. This kind of iterative testing is similar to the process of natural selection, and eventually, the pipelines that make the best efficiencies will stick around. This leads to a well-tuned pipeline suited to the characteristics and challenges of the dataset being processed.

3.2. Application in Arabic named entity recognition

When the TPOT logic is used for the Arabic NER tasks, this has great benefits since it automates the process of optimizing the machine learning pipeline. One of the main advantages of TPOT for Arabic NER is its ability to automatically explore the intricate domain of machine learning algorithms and preprocessing choices to identify the optimal machine learning pipeline. This is especially beneficial, given that the Arabic language has a very rich morphological structure. Using genetic algorithms, TPOT assesses various models and preprocessing methods to maximize accuracy and performance [15]. This automated method decreases the requirement for manual parameter tuning and experimentation, which can be especially laborious as the challenges Arabic poses; for example, the right-to-left writing system of its script, the existence of diacritics, and the context-sensitive nature of its grammar [6].

Moreover, this ability of TPOT to learn the intricacies of Arabic NER by recalibrating its search for the best pipeline from the cross-validation scores could result in the discovery of new mixing of preprocessing and modeling methods that can be missed by human researchers [21]. These comprise far better-exploited vectorization types, attribute collection, and the combination of language-specific courses of action that are specific to handling Arabic text [5]. Further, the ability of TPOT to optimize models using a genetic algorithm can greatly improve the performance of a model by searching the space of possible machine learning algorithms and hyperparameters to identify the optimal configuration, which could produce higher precision, recall, and F1-scores for Arabic NER tasks than more traditional coded machine learning approaches [31].

3.3. Impact of data augmentation on tree-based pipeline optimization tool's effectiveness in NER

This, in turn, implies that integrating data augmentation adds an extra degree of reliability for the TPOT when applied to the NER domain, more specifically, the Arabic one. However, TPOT is given multiple pipelines to analyze and refine the NER model; thus broadening the training dataset with a substantial and diverse data collection provides TPOT with more information to work with [16]. This increased diversity enables TPOT to evaluate better and optimize, resulting in a more optimized NER pipeline. This helps the model to generalize better since natural languages are naturally very variable and complex. This is a very helpful trait for Arabic [32].

TPOT also enhances its effectiveness by revealing the data augmentation approaches that actually drive substantial improvements in the precision, recall, and F1-score of Arabic NER [13]. Such improvements are essential for using Arabic NER systems in lots of domains, including sentiment analysis [22], content categorization [33], and information extraction [34], [35]. where the reliability of the output of such applications can significantly depend on precise entity recognition.

The experiments place a great deal of emphasis on the utility of data augmentation in improving the performance of TPOT over Arabic NER tasks. TPOT is the First Breakthrough in Arabic NLP [15] that selects data augmentation techniques and model configurations automatically and uses AutoML Techniques. It accelerates the development and enhances the performance, efficiency, and robustness of the Arabic NER models [36]. The implications of the findings call for more experimental setups, but they also demonstrate the utility of the TPOT platform in producing advanced, usable NER models that are suitable for the challenges of the Arabic language [37], [38].

4. METHOD

4.1. Datasets

The proposed method was also tested on popular benchmark datasets in the literature of Arabic NER, i.e., news and tweets corpora. The reason behind choosing these datasets was that they consist of both formal and informal Arabic text, and with it a comprehensive assessment of the generalization capability of the model is achievable. By evaluating the data augmentation strategy with the help of several datasets, it is guaranteed that the proposed strategy is evaluated in a wide range of linguistic and domain conditions.

4.1.1. News dataset

The first dataset, news, is an Arabic NER corpus that was created for annotation purposes. It contains 292 sentences and was taken from Arabic (Egypt) news, at https://news.google.com/home?hl=ar&gl=EG&ceid=EG:ar. October 6, 2012: from the RSS feed on Zilch, I/O. news is a general news dataset that covers diverse topics such as politics, finance, health, sports, entertainment, and technology, sourced from different international and local news outlets [1].

4.1.2. Tweets dataset

The tweets dataset is the data split used for the final evaluation benchmark, which is taken from [39] consisted of 3,646 randomly sampled tweets collected between May 3 and May 7, 2012, while the test set consisted of 1,423 tweets collected between November 23 and 27, 2011. Using this corpus, a large variety of up-to-date tweets [1] in Arabic was collected through the 'lang' search query on X (formerly Twitter). In both datasets, the automatic content extraction (ACE) tagging scheme defined by the linguistic data consortium (LDC) is followed for entity tagging. The news and tweets datasets offer a wide range of evaluations for the NER model, including semantically ordered pieces of news and informal social media posts that present the challenge of using common language. The diversity of the data sources is useful in validating the generalization capability of the NER approach compared to the state-of-the-art [35]. The motivation for creating these datasets is to obtain a thorough assessment of the NER model on different types of Arabic texts and potentially prepare it for application in real-world scenarios. In both datasets, part-of-speech (POS) [40] tagging was also a necessary step in the preparation of the data.

4.2. Data augmentation

4.2.1. Synonym replacement

Elaboration of an Arabic synonym database as part of our efforts to improve Arabic NER, a large database of Arabic synonyms by combining and collecting 2 of the most used databases together into a single CSV File. In this integration, various modifications and normalization processing were done, including diacritics removal and word forms (both words in variable forms, i.e., singular-plural) to cater to variations of word usage in social media. For example, “ة” and “ه” are used interchangeably in words such as “كرة” and “كره”. The Arabic WordNet (AWN) [41] has only one instance of existence and is a lexical resource based on

the English WordNet. AWN [42] is a lexical resource for Arabic that groups Arabic words into sets of synonyms or synsets and describes the properties of the synsets, semantic relations, and grammatical relations. Such a database is surely very helpful in many NLP tasks, such as semantic analysis, word sense disambiguation, and machine translation, due to the complex nature of Arabic morphology and semantics. AWN is in continuous development through academic collaboration and is designed to address challenges offered by dialectal diversity and changing usage for modern linguistic analysis. Data from Almaany at https://www.almaany.com/, an extensive online resource with a rich monolingual Arabic thesaurus containing Arabic synonyms and semantic relations. This database is also a helpful source for linguists, writers, translators, and language learners, as it allows checking of a wide range of Arabic synonyms in a user-friendly way, thus helping to improve Arabic language skills and vocabulary.

Scripts that merge the two databases to output a single CSV file in order to get the most out of these resources. All words are unique in this script, making parsing easier. This produces a final database, which is a treasure trove of 28,553 lines that can have infinite combinations with which to experiment for Arabic text analysis. Not only does this synthesis of these databases provide a great deal of lexicon variations to the NER tool, but more importantly, it plays a significant role in the ability to acquire Arabic text in different contexts. This integrated database is quite extensive and is essential for the NER system to interpret and identify a wide range of linguistic varieties of the Arabic language. Having this many linguistic resources is extremely valuable in tasks like semantic understanding and text analysis and makes the NER system one of the most advanced Arabic language processing technologies.

Synonym replacement and meaning conservation involve obtaining synonyms for any given word in a sentence from an extensive database as the first step of the process. Every word with a one-word appropriate word without disturbing the NER tags of the sentence. Once the words in the sentence have been carefully swapped out, the validation step is performed. The meaning of the original and modified sentences is still equivalent (to ensure this, several quality assurance mechanisms at this step). This guarantees that the augmented dataset retains the semantics that play an important role in NER performance, something considered an essential part of the proposed method.

A BERT-based model pre-trained for the semantic similarity task, which embeds both a pair of vector sentences and ensures the vectors are close after replacement [28]. Thus, this model is integral to the original and altered sentences. However, the kinds and quality of the augmented dataset, which would be reliable enough to keep its semantic meaning, also imply a good resource for the NER system [29].

As shown in Figure 1, this method is also supported by a large-scale synonyms database in which every word has several synonyms and can be applied multiple times. There are many more synonyms present in the data to repeat this step at the cost of the semantics of the sentence. This cycle has a great impact on dataset augmentation and provides the NER system with an enriched dataset based on diverse languages. This approach could help to produce many forms of sentences that would allow the model to better learn the semantic and syntactic variations of the natural language, especially in the case of Arabic NER data. Using such a process repeatedly adds more depth and diversity to the training set and helps develop a more robust and accurate NER model.

4.2.2. PERS tag replacement

The data augmentation technique is a custom method created to increase the diversity of the AR NER training dataset. This technique is based on replacing names of certain tags (here B-PERS and I-PERS meaning the beginning and interior of person names, respectively) with a wide variety of possible Arabic names. This method makes use of the triplet nature of Arabic personal names.

The script contains an extensive collection of popular Arabic names covering a variety of cultural and regional differences. This list consists of names such as "أحمد", "محمد", and "علي". In the augmentation, the script iterates over the dataset and finds words with the tags 'B-PERS' or 'I-PERS'. After finding those tags, the names are replaced with randomly chosen ones stored in the shared list. Apart from avoiding repetition, this replacement maintains the syntactic and semantic cohesion of the sentences.

The script works on each sentence and makes sure to duplicate and augment each sentence with a word tagged 'B-PERS'. Treating the sentence very carefully maintains much of its original form, but this process gives the dataset more realistic variations on names. The generated sentences are then combined to build a richer dataset, which is ultimately used to train the NER models. This augmentation technique aims to generate a dataset to help the NER system learn a wider range of personal names, which will contribute to its performance when needing to find and classify person entities in other text. The enriched dataset is an essential part of building a strong and portable Arabic NER system that can handle the text found in the wild.

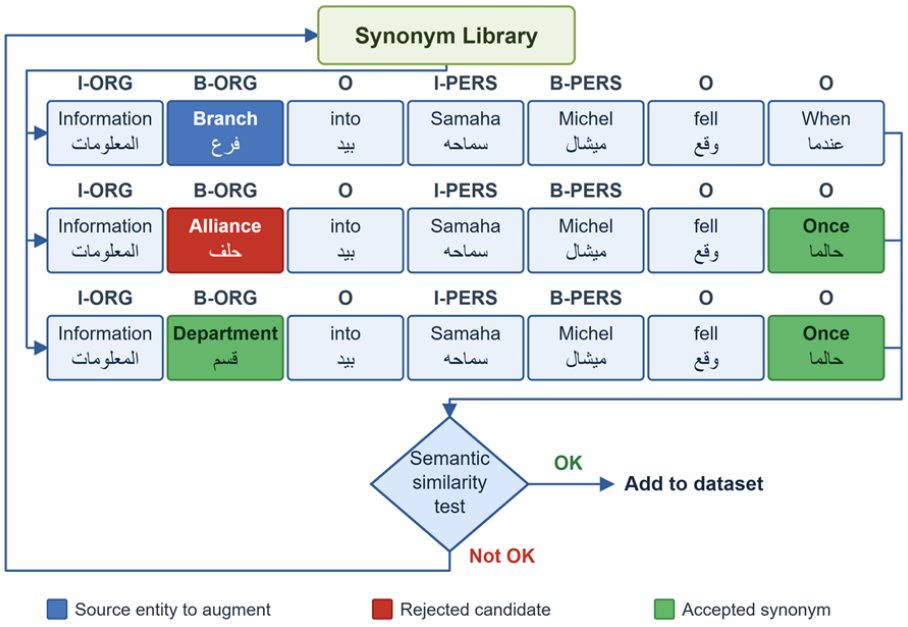

Figure 1. Synonym-based data augmentation with semantic-similarity filtering for Arabic NER

4.2.3. Third PERS tag deletion

Here, a specific data augmentation technique is used to purify the Arabic NER dataset. This approach is for sentences with tokens of person names (other than B-PERS and I-PERS) This technique then is applied to the sentence according to if and how often these tags appear in the preprocessed sentence. The dataset is examined, and the position of every sentence containing any of the tags (here, I-PERS specifically) is recorded. To maintain the information on how many 'I-PERS' tags repeat in one of the sentences, the script is used. If there are less than 2 consecutive 'I-PERS' tags, the sentence stays the same. Otherwise, the script does add the sentence into the augmented dataset if it finds two or more consecutive tags for the same word as 'I-PERS'.

This works especially well for sentences with longer names. This technique adds examples with longer 'I-PERS' sequences in the input sentences to generate more difficult training cases for the NER model that contains certain sub-strings of personal names, which may be ambiguous for NER systems to find the proper parts and classify them correctly. Such contextual embedding is critical for improving the performance of the model, especially in the case where the person names are of greatly varying lengths and structures. When the whole dataset is processed, the script takes the selected sentences and builds a new dataset with the augmented portions. By providing fewer and tailored examples, this dataset specifically encounters a higher complexity in NER to ensure that the NER system is ready for a real-life textual range of person names.

The purpose behind this augmentation method is to create a stronger and more resilient Arabic NER model. The training of the model is done on new data containing difficult examples where there are multiple I-PERS tags. This should make the model more reliable when identifying and counting person entities in the text. This improvement is an important step for significant gains in Arabic NER, where personal names vary in complexity and confusingly. Therefore, the augmented dataset that comes from this way represents one of the most forcing components in producing a highly effective NER model that can better accommodate the complexities of Arabic-named entities.

The three data augmentation techniques have massively increased the amount of training data. The tweets dataset grew significantly, reaching 57,969 lines and leading to a final collection of 155,194 lines (~168% increase). In a similar fashion, the news dataset experienced growth from 10,003 lines to 31,348 lines, resulting in a 213% augmentation. Increasing the dataset size also demonstrates the effectiveness of the augmentation techniques in creating diverse training data, which helps improve the model's performance.

4.3. Experimental setup and results

In this research, the same evaluation metrics as those detailed in [1] were employed. These metrics are pivotal for assessing the NER model's performance. Specifically i) accuracy is calculated as

$\frac{\text{total number of correct predictions}}{\text{total number of predictions}}$, ii) precision is determined by the $\frac{\text{true positives}}{\text{true positives + false positives}}$, representing the proportion of correctly identified entities among all identified entities, iii) recall is computed as $\frac{\text{true positives}}{\text{true positives + false positives}}$, indicating the model's ability to correctly identify all actual entities, and iv) F1-score is the harmonic mean of precision and recall, calculated as $2 \times \frac{precision \times recall}{precision+recall}$, providing a balance between precision and recall, especially useful when the dataset is imbalanced. By adhering to these metrics, a robust and comparable evaluation framework is ensured for the NER model in the context of Arabic language processing [43].

4.3.1. News dataset

A sharp increase in F1-scores from the baseline was demonstrated through data augmentation and the power of TPOT, resulting in a striking improvement across all categories of the news dataset Table 1. In particular, the F1-score for person tags, organization tags, location tags, and the overall F1-score improved by approximately 0.77%, 2.85%, 3.14%, and 1.10%, respectively. These results clearly demonstrate that the approaches are effective in improving the performance of the Arabic NER system.

Table 1. NER results for the augmented news dataset

Datasets	Person F1	Organization F1	Location F1	Overall Avg.F1
News dataset	90.9	66.7	82.7	81.5
Augmented news dataset	91.6	68.6	85.3	82.4

The numerical improvements in F1-scores are further illustrated in Figure 2. It shows a comparative visualization of F1-scores on the news test set in terms of named entity. The horizontal axis represents the entity types, and the vertical axis represents the corresponding F1-scores. As the figure illustrates, the proposed data augmentation approach and TPOT optimization steadily outperform the baseline model across all entity types.

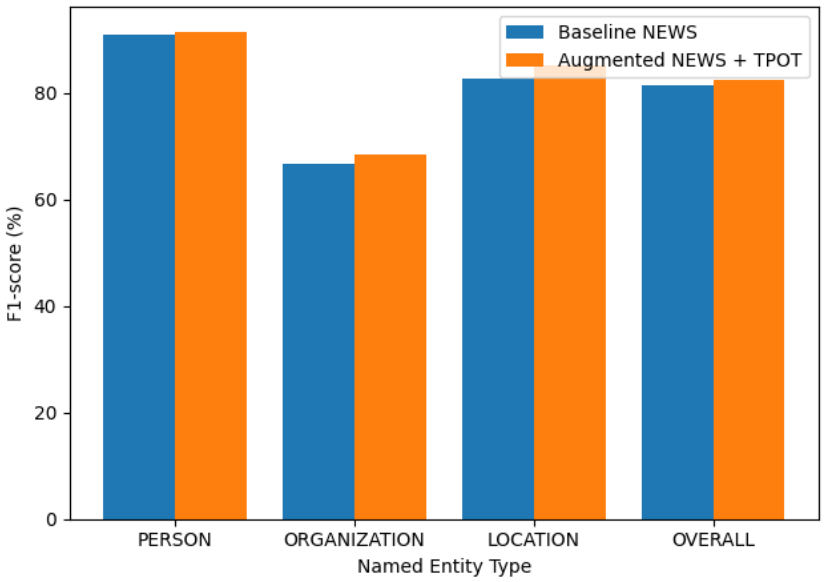

Figure 2. Comparison of F1-scores on the news test set before and after data augmentation

4.3.2. Tweets dataset

After applying data enhancement and then the TPOT algorithms, notable improvements in the F1-score of each entity type in the tweets dataset are presented in Table 2. The F1-score of person tags increased from 73.7 to 75.5, about a 2.44% growth. Then, for organization tags, it went from 46.7 to 48.0, a 2.78% increase. Location tags improved from 72.4 to 74.9 (3.45% improvement). Notably, the overall

F1-score jumped from 65.6 to 69.1, an increase of +5.34%. This translates into significant improvements in the performance of the Arabic NER system, particularly with online text, and demonstrates the impact of these methods.

Table 2. NER results for the augmented tweets dataset

Datasets	Person F1	Organization F1	Location F1	Overall Avg.F1
Tweets dataset	73.7	46.7	72.4	65.6
Augmented tweets dataset	75.5	48.0	74.9	69.1

The impact that the proposed data augmentation strategy has on the tweets test set, based on the various named entity types, is presented in Figure 3. The horizontal line is the type of entities, and the vertical one is the values of F1-score. The findings demonstrate that all the types of entities exhibit significant performance gains, which proves that data augmentation is especially successful in informal and noisy Arabic text.

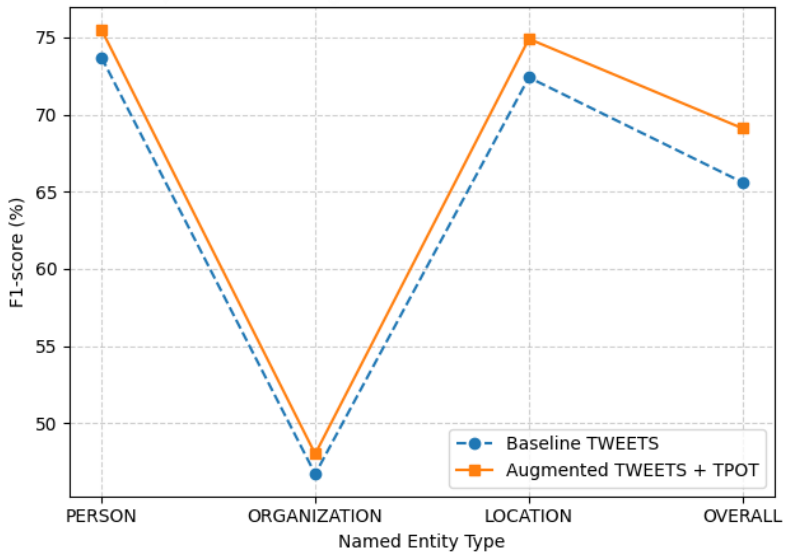

Figure 3. Impact of data augmentation on F1-scores across entity types in the tweets test set

5. FUTURE WORK

In future research, cover how deep learning algorithms can be used on news and tweets Arabic NER datasets. In recent times, the deep learning process expects richer linguistic information from the data points of these datasets [44]. Due to their advanced capability of pattern recognition, deep learning models are anticipated to be one of the most powerful tools for modeling the semantics and high-level structures of Arabic [45]. This not only gives the right and appropriate English and Turkish modalities to describe the non-romanized Arabic text, but it can also provide a significant step in Arabic NLP, as this technique sets the ground for future and more sophisticated and more helpful techniques regarding NER in many other aspects of the Arabic text [46].

6. CONCLUSION

This work highlights the remarkable effect of data augmentation methods in Arabic NER, with special reference to new techniques such as cross-resource synonym replacement and tagger-based augmentation tours. The TPOT has definitely been a facilitating factor. Still, the enhancements observed are fundamentally the product of strategy and creativity in the application of data augmentation. These techniques have crossed the barriers of the data from which they originated, forming the basis of much of the progress in Arabic NER observed, which represents not just an incredible feat for handling the complexities

of the language but also highlights the generalizability and possibilities that these methodologies may afford in other linguistically complex situations.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Brahim Ghazoui	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Ismail El Bazi	✓	✓				✓	✓	✓		✓	✓	✓		
Brahim Ait Ben Ali	✓		✓	✓			✓			✓	✓		✓	
Hicham Moussa					✓					✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [BG], upon reasonable request.

REFERENCES

[1] B. A. B. Ali, S. Mihi, N. Laachfoubi, and A. A. Mlouk, "Arabic named entity recognition in Arabic tweets using BERT-based models," *Procedia Computer Science*, vol. 203, pp. 733–738, 2022, doi: 10.1016/j.procs.2022.07.109.

[2] H. Dai *et al.*, "AugGPT: leveraging ChatGPT for text data augmentation," *IEEE Transactions on Big Data*, vol. 11, no. 3, pp. 907–918, 2025, doi: 10.1109/TBDATA.2025.3536934.

[3] W. Antoun, F. Baly, and H. Hajj, "AraBERT: transformer-based model for Arabic language understanding," in *4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, May. 2020, pp. 9–15.

[4] L. Liu, J. Shang, and J. Han, "Arabic named entity recognition: what works and what's next," in *Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 60–67, doi: 10.18653/v1/W19-4607.

[5] 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 Transactions on Knowledge and Data Engineering*, vol. 36, no. 3, pp. 943–959, Mar. 2024, doi: 10.1109/TKDE.2023.3303136.

[6] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition: a feature-driven study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 926–934, Jul. 2009, doi: 10.1109/TASL.2009.2019927.

[7] W. Liu and X. Cui, "Improving named entity recognition for social media with data augmentation," *Applied Sciences*, vol. 13, no. 9, Apr. 2023, doi: 10.3390/app13095360.

[8] S. Feng *et al.*, "A survey of data augmentation approaches for NLP," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 968–988, doi: 10.18653/v1/2021.findings-acl.84.

[9] S. Kobayashi, "Contextual augmentation: data augmentation by words with paradigmatic relations," in *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 452–457, doi: 10.18653/v1/N18-2072.

[10] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 86–96, doi: 10.18653/v1/P16-1009.

[11] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020, doi: 10.5555/3495724.3496249.

[12] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in Neural Information Processing Systems*, vol. 28, pp. 649–657, 2015, doi: 10.5555/2969239.2969312.

[13] J. Wei and K. Zou, "EDA: easy data augmentation techniques for boosting performance on text classification tasks," in *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6381–6387, doi: 10.18653/v1/D19-1670.

[14] J. Wan, X. Yu, and Q. Guo, "LPI radar waveform recognition based on CNN and TPOT," *Symmetry*, vol. 11, no. 5, May 2019, doi: 10.3390/sym11050725.

[15] R. S. Olson and J. H. Moore, "TPOT: a tree-based pipeline optimization tool for automating machine learning," in *Workshop on automatic machine learning*, 2019, pp. 151–160, doi: 10.1007/978-3-030-05318-5_8.

[16] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[17] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, "Data augmentation techniques in natural language processing," *Applied Soft Computing*, vol. 132, Jan. 2023, doi: 10.1016/j.asoc.2022.109803.

[18] A. J. Ratner, H. R. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, "Learning to compose domain-specific transformations for data augmentation," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3239–3249, 2017, doi: 10.5555/3294996.3295083.

[19] F. M. Luque, "Atalaya at TASS 2019: Data augmentation and robust embeddings for sentiment analysis," in *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, vol. 2421, 2019, pp. 561–570.

[20] C. Jacquemin and E. Tzoukermann, "NLP for term variant extraction: synergy between morphology, lexicon, and syntax," *Natural language information retrieval*, Springer, 1999, pp. 25–74, doi: 10.1007/978-94-017-2388-6_2.

[21] S. A. Salloum, A. Q. AlHamad, M. Al-Emran, and K. Shaalan, "A survey of Arabic text mining," *Intelligent natural language processing: Trends and applications*, Springer, 2018, pp. 417–431, doi: 10.1007/978-3-319-67056-0_20.

[22] A. Oussous, F. -Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "ASA: A framework for Arabic sentiment analysis," *Journal of Information Science*, vol. 46, no. 4, May 2022, doi: 10.1177/0165551519849516.

[23] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: a comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, Apr. 2021, doi: 10.1145/3439726.

[24] S. T. Aroyehun and A. Gelbukh, "Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling," *COLING 2018 - 1st Workshop on Trolling, Aggression and Cyberbullying, TRAC 2018 - Proceedings of the Workshop*, 2018, pp. 90–97.

[25] J. Chen, Z. Yang, and D. Yang, "MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2147–2157, doi: 10.18653/v1/2020.acl-main.194.

[26] I. Androutsopoulos and P. Malakasiotis, "A survey of paraphrasing and textual entailment methods," *Journal of Artificial Intelligence Research*, vol. 38, pp. 135–187, 2010, doi: 10.1613/jair.2985.

[27] W. Zhang, P. Ge, W. Jin, and J. Guo, "Radar signal recognition based on TPOT and LIME," in *2018 37th Chinese Control Conference (CCC)*, Jul. 2018, pp. 4158–4163, doi: 10.23919/ChiCC.2018.8483165.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[29] N. Reimers and I. Gurevych, "Sentence-BERT: sentence embeddings using Siamese BERT networks," in *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3980–3990, doi: 10.18653/v1/D19-1410.

[30] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn," *13th Python in Science Conference*, pp. 32–37, 2014, doi: 10.25080/majora-14bd3278-006.

[31] D. E. Goldberg, "Genetic algorithms in search, optimization, and machine learning," *Choice Reviews Online*, vol. 27, no. 2, pp. 27-0936-27–0936, 1989, doi: 10.5860/choice.27-0936.

[32] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics,* vol. 2, pp. 567–573, 2017, doi: 10.18653/v1/P17-2090.

[33] S. F. Abdhood, N. Omar, and S. Tiun, "Data augmentation for Arabic text classification: a review of current methods, challenges and prospective directions," *PeerJ Computer Science*, vol. 11, 2025, doi: 10.7717/peerj-cs.2685.

[34] H. Huang *et al.*, "Unicoder: a universal language encoder by pre-training with multiple cross-lingual tasks," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 2485–2494, 2019, doi: 10.18653/v1/D19-1252.

[35] K. Darwish and W. Gao, "Simple effective microblog named entity recognition: Arabic as an example," *9th International Conference on Language Resources and Evaluation, LREC 2014*, pp. 2513–2517, 2014, doi: 10.63317/5imzhbt74xh4.

[36] K. Shaalan, "Introduction to Arabic natural language processing (synthesis lectures on human language technologies)," *Machine Translation*, vol. 24, no. 3–4, pp. 285–289, Dec. 2010, doi: 10.1007/s10590-011-9087-8.

[37] R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd, and J. H. Moore, "Automating biomedical data science through tree-based pipeline optimization," *Applications of Evolutionary Computation (EvoApplications 2016)*, Cham: Switzerland: Springer, 2016, pp. 123–137, doi: 10.1007/978-3-319-31204-0_9.

[38] O. F. Zaidan and C. C.-Burch, "Arabic dialect identification," *Computational Linguistics*, vol. 40, no. 1, pp. 171–202, 2014, doi: 10.1162/COLI_a_00169.

[39] M. Jarrar, M. Khalilia, and S. Ghanem, "Wojood: nested Arabic named entity corpus and recognition using BERT," *2022 Language Resources and Evaluation Conference, LREC 2022*, pp. 3626–3636, 2022, doi: 10.63317/57rz4w3or4jm.

[40] R. Mitkov, Ed., *The oxford handbook of computational linguistics*, 2nd ed. Oxford: Oxford University Press, 2022, doi: 10.1093/oxfordhb/9780199573691.001.0001.

[41] S. Elkateb *et al.*, "Building a WordNet for Arabic," in *Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, 2006, pp. 29–34, doi: 10.63317/26v89mej5wrk.

[42] A. A. Freihat, H. Khalilia, G. Bella, and F. Giunchiglia, "Advancing the Arabic WordNet: elevating content quality," *6th Workshop on Open-Source Arabic Corpora and Processing Tools, OSACT 2024 with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation at LREC-COLING 2024 - Workshop Proceedings*, pp. 74–83, 2024, doi: 10.63317/3sodtwuwrss4.

[43] D. Jurafsky and J. H. Martin, "Speech and language processing," web.stanford.edu. Accessed: Oct. 15, 2025. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[44] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2022, doi: 10.1109/TKDE.2020.2981314.

[45] Y. Kim, Y. Jernite, D. Sontag, and A. Rush, "Character-aware neural language models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, Mar. 2016, doi: 10.1609/aaai.v30i1.10362.

[46] O. F. Zaidan and C. C.-Burch, "The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content," *ACL-HLT 2011 - 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, pp. 37–41, 2011.

BIOGRAPHIES OF AUTHORS



Brahim Ghazoui is a computer science engineer, graduated from the National Institute of Posts and Telecommunications (INPT), Morocco. Since 2022, he has been pursuing his Ph.D. at the LGS Laboratory, Department of Computer Science, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Morocco. His research interests include machine learning and deep learning for natural language processing and its applications. He can be contacted at email: brahim.ghazoui@usms.ac.ma.



Ismail El Bazi holds a doctorate in Computer Science from Hassan 1st University and an Engineering degree in Computer Engineering from Cadi Ayyad University. He is also certified in project management (PMP) and in Agile methods (PMI-ACP) since 2013. After 10 years of professional experience in the field of software engineering with international IT companies, he joined the Sultan Moulay Slimane University in 2019 as assistant professor. His research focuses are artificial intelligence, Arabic natural language processing, and data science. He can be contacted at email: i.elbazi@usms.ac.ma.



Brahim Ait Ben Ali holds a doctorate in Computer Science from the Faculty of Sciences and Techniques, Hassan First University of Settat, Morocco, where he conducted his research at the IR2M Laboratory. He is affiliated with the Laboratory of Mathematics, Artificial Intelligence and Sustainable Technologies, Cadi Ayyad University, Marrakesh, Morocco. He has published several papers in reputed journals and international conferences. His research interests include machine learning and deep learning for natural language processing and its applications. He can be contacted at email: b.aitbenali@uca.ac.ma.



Hicham Moussa holds a doctorate in Applied Mathematics, currently works at the Department of Mathematical Sciences, FST Béni Mellal, Université Sultan Moulay Slimane. He does research in analysis and applied mathematics. Their current project is 'Coupled system in modular spaces'. He can be contacted at email: hichammoussa23@gmail.com.