

Towards an approach based on particle swarm optimization for Arabic named entity recognition on social media

Brahim Ait Ben Ali¹, Soukaina Mihi¹, Ismail El Bazi², Nabil Laachfoubi¹

¹IR2M Laboratory, Faculty of Sciences and Techniques, Hassan First University of Settat, Settat, Morocco

²National School of Business and Management, Sultan Moulay Slimane University, Beni Mellal, Morocco

Article Info

Article history:

Received Mar 9, 2022

Revised Jun 10, 2022

Accepted Jul 5, 2022

Keywords:

Dialect arabic language

Feature selection

Named entity recognition

Natural language processing

Particle swarm optimization

Social media

ABSTRACT

Named entity recognition is an essential task for various applications related to natural language processing (NLP). It aims to retrieve a variety of named entities (NEs) from text and categorize them according to predetermined target categories. In many cases, using the entire feature set can be time-consuming and negatively impact the performance. Moreover, it is challenging to find the relevant subsets of features for a particular task due to the high number. The feature selection technique is an unsupervised process for selecting informative features by creating a new subset of informative features. This technique is used to enhance the underlying algorithm's performance. This article implements an effective feature selection algorithm using particle swarm optimization (PSO) to identify and classify the Arabic NEs in the text from social media. PSO is a search algorithm that utilizes a population of particles in a multidimensional space. The proposed method is evaluated using two publicly available Arabic Dialect social media datasets. It is demonstrated through comparisons with both baselines and previous models that the new approach achieves significant accuracy with considerably reduced feature sets in all parameters.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Brahim Ait Ben Ali

IR2M Laboratory, Faculty of Sciences and Techniques, Hassan First University of Settat

Settat, 26000, Morocco

Email: aitbenali.brahim@gmail.com

1. INTRODUCTION

Named entity recognition (NER) identifies and classifies names in a text into predetermined categories like people, places, and organizations [1]. It is an essential task in various natural language processing (NLP) applications like question answering [2], machine translation [3] and information retrieval [4]. Its significance has recently emerged in opinion mining, which uses it as a preprocessing stage to obtain increased performance. However, a NER system's efficiency through machine learning depends on the features used in the training and testing phases.

Nowadays, researchers in machine learning and similar fields recognize the importance of reducing the dimensionality of analyzed data. Such high-dimensional data affects learning models, increasing both search space and computational time, and may also be considered information-poor [5], [6]. Moreover, due to the high-dimensional data and many features, constructing an appropriate machine learning model may be challenging and even unsuccessful in many cases. We are confronted with the "curse of dimensionality" related to a well-known phenomenon when analyzing data in high-dimensional spaces. It states that data in high-dimensional spaces eventually become sparse [7].

Researchers mainly use two approaches to address the problems that arise from the data's high dimensionality. The first is feature extraction, which involves the creation of a new feature space with low

dimensionality. The second is feature selection, which focuses primarily on removing irrelevant and redundant features from the original feature set and selecting a small subset of relevant features. Selecting a set of features out of N features can be solved with 2^N possible solutions (or subsets of features). The complexity is doubled with each additional element. For example, a data set with 1,000 characteristics per instance has 2^{1000} possible characteristic subsets. We apply the feature selection technique to decrease dimensionality and enhance system performance. Our feature selection algorithm relies on a wrapper approach formulated as an optimization problem.

Particle swarm optimization (PSO) represents an evolutionary technique inspired by the social behavior of birds. Some of PSO's inherent features, including avoidance of becoming locked into local optima, parallel selection capability, and the ability to handle noisy environments, have led us to adopt PSO as our method of selection. In addition, compared to other evolutionary algorithms, PSO can be easily implemented and requires few parameters to be adjusted. Motivated by the strength of the evolutionary algorithm, in this paper, we apply PSO [8] as an underlying optimization strategy. Several recent studies demonstrate that PSO converges more rapidly than other widely used optimization techniques [9]. Most of the original benefits of meta-heuristic algorithms, such as the capability to deal with complex non-linearities, discontinuities in the objective function, discrete variable treatment, multi-objective optimization, etc., are still available in the proposed efficient approaches. Motivated by this finding, a PSO is applied in this study. The main contributions of this paper are as: i) proposal of a feature selection technique using PSO in the extraction of named entities, ii) analysis of feature selection on word representation features, and iii) impact of feature selection on our system.

In order to evaluate the impact of the feature selection techniques, we experimented using all feature sets and features minimized by PSO. In addition, we perform experiments on two different data sets, as follows the tweets dataset and the news dataset. The evaluation results demonstrate that we obtain meaningful performance improvements using feature selection. Furthermore, the highest efficiency of the system was achieved when we applied the feature selection technique based on PSO.

The remainder of this paper is organized as follows: section 2 provides an overview of the Arabic language and the challenges related to recognizing Arabic-named entities. Section 3 presents an overview of the PSO algorithm, followed by a description of the proposed approach in section 4. Section 5 presents a dataset to be used and features for entity extraction. Section 6 presents the experimental results. Finally, a conclusion and future direction are provided in section 7.

2. BACKGROUND

2.1. Evolution of NER

During the sixth message understanding conference (MUC-6) [10], the terminology "named entity" (NE) was first used to recognize the names of organizations, persons, and locations in the text, as well as expressions of currency, time, and percentage. The interest in NERs has increased since MUC-6, and numerous scientific events. For example: CoNLL03 [11], ACE [12], IREX [13], and TREC entity track [14] are making a great deal of effort on this issue.

2.2. Language aspects and challenges

Arabic has a rich morphology and difficult syntax [15]. There are three significant kinds of Arabic: Classical Arabic, the Holy Koran language, used for over 1,500 years. Modern standard Arabic is one of the six official languages of the United Nations, while most research on NLP Arabic focuses on informal Arabic as a colloquial Arabic language. This language is irregular and differs from country to country and region to region. By comparison with English NER, here are some examples of challenges for Arab NER [16].

- Lack of capitalization: the capitalization in foreign languages is a powerful indicator of the named entity. However, Arabic does not capitalize letters, making identifying NEs more difficult.
- Noun confusion: certain words may be proper nouns, nouns, or adjectives. For example, *jamiyolap* = "جميلة," which signifies "beautiful" can be either a proper noun or an adjective. Another example, *jamAl* = "جمال," which means "beauty," would be a noun but can be a common noun or proper noun.
- Agglutination: named entity (NE) can be attached to different clitics due to Arabic's agglutinating nature. A morphological analysis pre-processing step must be carried out to identify and categorize these entities. This feature makes the task of the Arabic NER more difficult. Examples include conjunctions such as *و* (*wāw*, and) and *ف* (*fā'*, if), prepositions such as *ل* (*lām*, for), *ك* (*kāf*, as), and *ب* (*bā'*, by), or a combination of both such as *ول* (*wāw-lām*, and-for).
- Optional short vowels: short (diacritical) vowels are facultative in Arabic. Most written modern standard Arabic (MSA) texts do not contain diacritics, leading to considerable ambiguity since a single

undiacritized word refers to different terms or meanings. Such ambiguity can often be resolved through contextual information [4].

In addition to the challenges previously mentioned, Arabic NER faces other challenges for dialectal Arabic:

- Insufficient labeled data: for the supervised Dialect Arabic NER.
- Absence of standard spelling or language academics [17]: In contrast to the MSA, many forms of the same word in DA can be rewritten, e.g., mAtEyT\$ = "ما تعيطش," mtEyT\$ = "متعيتش," which means "do not cry," are all acceptable forms since there is no single standard.
- Absence of comprehensive gazetteers: an issue confronting any NER system for any language that deals with NERs in the texts of social media, as those media, by definition, have a ubiquitous occurrence of very public nouns exemplified by the use of pseudonyms; thus, the class PERSON in NERs in social media will always face a coverage issue.

The application of NLP tools developed for MSA to DA yields significantly lower performance, requiring resources and tools specifically addressing DA [18].

2.3. Feature selection

Feature selection is a process that selects a subset of original features based on some criteria. It is essential and commonly utilized as a reduction technique for data mining. Feature selection is an area of research that has been actively pursued for decades in machine learning and data mining.

Typically, a feature selection process has four fundamental steps (shown in Figure 1), including the generation of subsets, the evaluation of subsets, the stopping criterion, and the validation of results. Subset generation is a search process that generates subsets of characteristics that are candidates for evaluation on the basis of a specific search strategy. A candidate subset is evaluated by comparing each candidate subset with the best previous one on the basis of a particular evaluation criterion. If the new subset is found to be better, it replaces the previous one. The process of generating and evaluating the assemblies is repeated until a certain stopping criterion is reached. Then the chosen best subset usually has to be validated by prior knowledge or different tests using synthetic and/or real-world datasets.

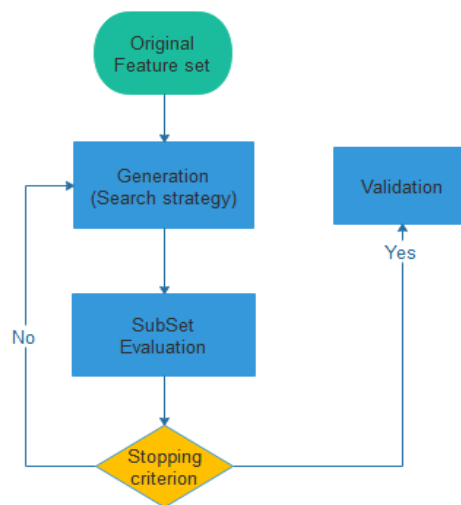


Figure 1. The iterative process of feature selection

3. PARTICLE SWARM OPTIMIZATION

PSO is a smart metaheuristic strategy derived from the social behavior of the swarm for its survival [19]. It is a population-based technique that birds and fish perceive to determine the optimal path. Generally, the PSO comprises a swarm of particles, with each particle having its particular position in the search space and moving at a specific velocity in the search space. The particle chooses the best path in each iteration based on its memory and learning which path the swarm has previously followed.

The new position is selected based on the previous knowledge acquired through its self-control position and the swarm's better position. As a meta-heuristic model, PSO produces limited or no assumptions regarding the problem to be optimized and can generate large areas of candidate solutions. This makes PSO very effective for optimization [20]. Two variables are iterated in the algorithm: the best global position

indicates the highest promising vector identified so far. The best personal position denotes the best particular solution for the particle.

3.1. Algorithm: PSO based feature selection

- At first, we randomly determine the population of swarms. Each swarm particle consists of binary-valued elements of length n (total number of features) with its position and the velocity of its movement in the search space. In math, the position and velocity of a particle are expressed as $\vec{P}(i)$ and $\vec{V}(i)$ respectively:

$$\vec{P}(i) = (p(i, 1), p(i, 2), \dots, p(i, n))$$

$$\vec{V}(i) = (v(i, 1), v(i, 2), \dots, v(i, n))$$

where $P(i, j) \in \{0, 1\}$, $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, n$ where N is no. of particle. The particle retains the best position ($\vec{B}(i)$) it has obtained so far and the best overall position (\vec{G}) i.e., the best position of the particle with the best solution.

- Position value $\vec{P}(i)$ of the particle is taken to be $\{0, 1\}$ based on the following expression:

$$P(i, j) = \begin{cases} 1, & \text{if } random \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

- The fitness function (value of the F-measure) for each particle $f(\vec{P}(i))$ is used to evaluate the particle. The memory is updated with the best position and the best overall position.
- Initially, the best position value $\vec{B}(i)$ of each particle is taken as 0. In each iteration k , the value of the best position is set to the values:

$$f(\vec{B}(i)) = \max(f(\vec{P}(i))_k, f(\vec{B}(i))_{k-1})$$

- Updating the value of the best overall position is performed when the fitness function $f(\vec{B}(i))$ in the swarm is higher than the current $f(\vec{G})$.
- Initially, the velocity vector is randomly generated. Then, at each iteration, the velocity of the particle is then calculated using the formula:

$$v(i, j) = \omega * v(i, j) + \varphi_1 (b(i, j) - p(i, j)) + \varphi_2 (g(j) - p(i, j))$$

where ω ($0 < \omega < 1$), φ_1 , and φ_2 are also known as inertia weights. These parameters are initialized using a uniformly generated random number across the range (0,1). The $b(i, j)$, $p(i, j)$ and $g(j)$ denote the j -th components of $\vec{B}(i)$, $P(i)$, and G , respectively.

- The particle position is updated using the following mathematical formula:

$$P(i, j) = \begin{cases} 1, & \text{if } (random < S(v(i, j))) \\ 0, & \text{otherwise} \end{cases}$$

where $0 < random < 1$ is a random uniform number,

$$S(v(i, j)) = \frac{1}{1 + \exp(-v(i, j))}$$

in other words, a sigmoid function. Then the value of the position of the particle changes from 0 or 1 according to the value of the velocity.

- Repeat steps (d) to (g) until convergence.

4. PROPOSED FEATURE SELECTION TECHNIQUE

4.1. Proposed approach

The proposed approach consists of five major steps, as shown in Figure 2. The first step concerns the data to be utilized in the experiments. The tweet and news datasets [21] are employed for this purpose. The second step is to normalize and segment the data into an appropriate processing format. The third step is to extract features from the transformed form of the data. As a contribution to this study, the fourth step is where feature selection is carried out on the retrieved features to choose the optimized ones. Lastly, an SVM classifier is employed for classifying the named entities.

4.2. PSO based feature technique

This section proposes our technique for the selection of features. All features are coded within a vector. At first, the value of the vector is randomly initialized, using values to determine if feature is present or not. Then, using a classifier, SVM, the combination of features that are represented as a vector is coached and tested using the validation set. Next, we calculate the value of the F-score, applied for each vector as a fitness function. The F-score value is computed to update the features in the following iterations. Finally, the optimized feature set is obtained. This set of characteristics serves as a final assessment of the test set. Tables 1-4 overview of the proposed feature selection for one iteration. A flowchart of the methodology employed in this work is shown in Figure 2.

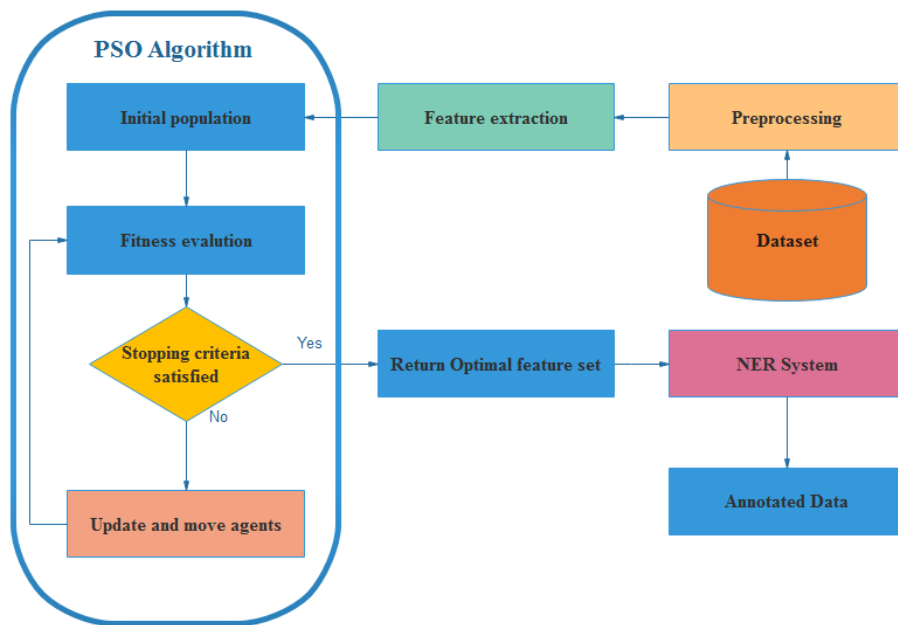


Figure 2. The architecture of the proposed system: PSO-based selection of features

Table 1. Sample tokens with their features

Token	Feature1	Feature2	Feature3	Feature4	Feature5
صباح "morning"	0	1	0	1	0
الخير "Good"	1	0	1	0	1
و "and"	0	0	0	0	1
السعادة "happiness"	0	1	1	1	0
على "to"	1	0	0	0	1
الجميع "everyone"	1	1	0	0	1

Table 2. Particles position initializations

Particles	Feature1	Feature2	Feature3	Feature4	Feature5
P1	0	1	0	1	0
P2	1	0	1	0	1
P3	0	0	0	0	1
P4	0	1	1	1	0
P5	1	0	0	0	1
P6	1	1	0	0	1

Table 3. Optimization process after iteration-(i) global best position value=62.79

Particles	Feature1	Feature2	Feature3	Feature4	Feature5	Fitness value	Best position
P1	0	1	0	1	0	42.05	42.05
P2	1	0	1	0	1	41.54	41.54
P3	0	0	0	0	1	35.80	35.80
P4	0	1	1	1	0	59.45	59.45
P5	1	0	0	0	1	60.48	60.48
P6	1	1	0	0	1	62.59	62.59

Table 4. Optimization process after iteration-(i+1) global best position value=65.74

Particles	Feature1	Feature2	Feature3	Feature4	Feature5	Fitness value	Best position
P1	1	1	1	0	1	45.63	46.20
P2	1	0	0	0	0	47.91	47.91
P3	1	0	1	0	1	65.74	65.74
P4	1	0	0	1	1	62.72	62.72
P5	1	0	0	1	1	59.37	60.14
P6	1	1	0	1	1	61.81	62.61

The PSO optimization process: (Tables 1 and 2) A sample token has specific features to illustrate the feature selection process during each iteration by the PSO algorithm. The value 1 indicates the selection of the feature, and the value 0 indicates the pruning of the feature. (Tables 3 and 4) Random initialization of the position of the particles after iteration i and the following iteration $(i+1)$, the shaded line indicates the most suitable particles in this iteration. After each iteration, the value of the best overall position indicates the best overall solution obtained. All fitness values are hypothetical.

5. DATASETS AND FEATURES

5.1. Evaluation scheme

The standard CoNLL NER evaluation script was used to evaluate the proposed approach. As detailed in [22], The CoNLL evaluation methods are the strongest because they give no sub-credit to a sub-extracted named entity. After running the CoNLL evaluation script, the results are presented in terms of precision, recall, and f-score for each class of NE [23].

- True positive (TP): entities recognized by NER and corresponding to the ground truth.
- False positive (FP): entities recognized by NER do not correspond to ground truth.
- False negative (FN): entities annotated in the ground truth that NER does not recognize.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

F_1 is the harmonic mean of precision and recall, and the balanced F-score is the most used:

$$F - 1 = 2 * \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

5.2. Datasets

The primary dataset evaluated within this model consists of the news dataset, a corpus annotated for the Arabic NER process, developed by [21]. The news dataset contains 292 sentences extracted in the RSS feed for the Arabic (Egypt) language site news.google.com as of October 6, 2012. It includes news from various sources and covers local and international information related to finance, politics, sports, health, technology, and entertainment. The data set has been split into training, testing, and validation datasets of 70%, 20%, and 10%, respectively.

The second dataset we used for evaluation is the tweet dataset. We use the split of the training and test data provided in [21], where the training dataset comprises 3,646 tweets randomly selected and extracted from tweets written between May 3 and May 7, 2012. The tweets were retrieved from Twitter using the lang:ar request. The test data consists of 1,423 tweets randomly extracted from tweets written between November 23, and November 27, 2011. This dataset was similarly tested by [24]. The two datasets follow the ACE tagging guidelines of the linguistics data consortium [25]. Table 5 shows data statistics.

Table 5. Twitter evaluation of data statistics

	Tokens	PER	LOC	ORG
Twitter Train	55k	788	713	449
Twitter Test	26k	464	587	316

5.3. Entities extraction features

The feature extraction of the entities is very important in the NER task. On the basis of the different possible combinations of word and label contexts available, the key features task are determined. To build a PSO-based NER system, we use the following features:

- a. Contextual feature: It consists of the local contextual characteristic related with tokens appearing in the 5-word window, meaning 2 to the left and 2 to the right of the current token.
- b. Word prefixes and suffixes: these features refer to fixed-length character sequences stripped from either the leftmost or rightmost positions of words.
- c. Gazetteer (GAS): A binary characteristic that indicates the word's occurrence in an Individual gazetteer. According to this method, the gazetteer consists of a union of the following elements: i) ANERGaz (<http://users.dsic.upv.es/~ybenajiba/resources/>, s.d.): as provided by Benajiba and Rosso [26], which comprises 2183 LOC, 402 ORG, and 2308 PER; and ii) WikiGaz: Wikipedia: extensive Wikipedia gazetteer [24], which comprises 50141 LOC, 17092 ORG, and 65557 PER.
- d. Affix stemmer: The term "affix" is defined as either a prefix or a suffix. This approach enables removing prefixes and suffixes from words. It removes the longest possible string of characters from a word based on a set of rules.
- e. Morphological characteristics (MORPH): These characteristics are generated by the MADAMIRA tool. Five morphological characteristics have been selected to be used in this work:
 - Aspect: refers to the aspect of an Arabic verb. There are four possible values: command, imperfect, perfective, not applicable. However, since none of the NEs can be verbal, we apply this characteristic as a binary characteristic specifying whether a word is marked for Aspect or not;
 - Gender: the nominal gender. It has three values: female, male, not applicable;
 - Person: it indicates information about the person. Possible values are 1, 2, 3, not applicable. As for the aspect, we use it as a binary feature which indicates if a word is tagged for the person or not;
 - Proclitic2: conjunction proclitic. The Madamira tool creates a total of nine different values related to this characteristic: non-proclitic, not applicable, Conjunction (ف) fa, related particle fa, conditional response fa, Subordinate conjunction fa, Particle (و) wa, Subordinate conjunction wa;
 - Voice: the voice of the verb. Values for this feature are as follows: active, passive, not applicable, undefined.
- f. Part of speech (PoS) information: PoS gives meaningful syntactic evidence to detect named entities (NEs). We utilize the part of speech information of the current token and/or the surrounding token(s) as a feature. The PoS information was retrieved from Madamira tools.
- g. Brown clustering IDs (BC IDs): provided by Brown *et al.* [27], is an approach to hierarchical clustering providing maximum information about each other's bigram words. Word representations, particularly brown clustering, have enhanced the NER system's performance when added as a feature [28].
- h. Word2vec cluster IDs (W2V Cluster ID): is an embedding learning algorithm that uses a model of neural network provided by [29]. The embeddings get mapped onto a set of latent variables, wherein a particular instance of these variables is representing the individual words. The system applies K-means clustering over word vectors and uses the cluster IDs as features.
- i. Word embedding (W2V) (also called "distributed word representations") convinces a latent real-valued semantics or vector syntax of individual words in a large untagged corpus through the use of space-continuous model languages [30]. An improved word can be gained if we have a considerable amount of training data, as the real value obtained from Speech vectors become more representative. We employ the well-known word2vec1 tool provided by [31] to retrieve vector representations of words.

5.4. Pre-processing

Standardization is the process of producing the canonical form of a token and/or a word to maximize the matching between a query token and document collection tokens. In its basic form, normalization pretreats the tokens into a unique form, but in a lightweight way. This is frequently performed in various pre-processing steps in order to make the multiple forms of a specific letter into a single unicode representation, for example, substituting the undotted Arabic letter "ى" with a final dotted letter "ي", when that letter occurs at the ending of an Arabic word.

In its complicated shapes, normalization is employed for dealing with morphological variation and word inflation [32]. This is known as stemming. Stemming is the process of making the various inflected

forms and variants of a given word into a single term, called "stem.". Whereas a root is derived as well from model words. Classifiers and index builders/searchers need to have this derivation process as it reduces the dependence on particular word forms making vocabularies smaller in size and otherwise having to include all available forms. In the present work, the MADAMIRA tool was used [33].

6. EXPERIMENT AND RESULTS

6.1. Parameter's settings

As previously mentioned, particle swarm optimization offers a sensible medium-term between the complexity of the research and the quality of the solution. Therefore, it could be applied to selecting the static classifier [34]. However, several parameters require adjustment to obtain a better convergence by using PSO with constriction factors: population size, Vmax, ϕ and the number of iterations. In order to adjust the parameters mentioned, we exclusively applied SVMs that have been first learned on the training data of the tweet and news datasets and then tested on the corresponding validation data of the same corpus.

Table 6 provides an overview of the parameter value ranges. For example, Vmax is supposed to be $2 \times X_{max}$, wherein X_{max} equals 1. The SVM was applied to calculate each candidate's performance in terms of the F-score as a fitness function. We considered the maximum F-score of the classifier as our decision metric.

Table 6. PSO parameters

Parameters	Range	Step size	Selected value
ϕ	[1.00,2.50]	0.02	1.49
Iteration Size	[50,100]	10	60
Population Size	[1-3] * particle size	1	2* particle size
Vmax	-	-	2

6.2. Results and analysis

Research on NER tasks has shown that the SVM-based NER task needed less time and performed better than the CRF-based NER task to recognize named entities when applying similar features [35]. For that purpose we create the baseline to evaluate our proposed approach for each dataset. The SVM was used as a base classifier for training the model.

Baseline: the baseline is learned utilizing a full set of features detailed in subsection 5.3. Based on the two-stage SVM approach, the effect of feature selection using PSO was investigated. In the first stage, the optimal features were retrieved using PSO, and in the second stage, SVM was performed with all the features. A common information extraction metric was used for the evaluation: precision, recall, and F-measure. We apply our PSO-based feature selection to the baseline models to identify the best feature set for each task. The findings are shown in Tables 7 and 8 for the tweets dataset and Tables 9 and 10 for the news dataset. The results indicate that only a smaller subset could reach higher performance for each dataset; this shows the effectiveness of our proposed approach. Interestingly, models developed utilizing the pruned feature sets reach a higher F_1 across all datasets. The results from feature selection based on the PSO are also significantly better than the baseline.

Table 7. SVM results using all features for the tweet dataset

SVM	Precision (%)	Recall (%)	F-Score
LOC	87.92	63.84	73.97
ORG	66.54	36.57	47.20
PERS	82.17	65.92	73.15
Overall	81.04	57.41	67.21

Table 8. Results with SVM-PSO for tweets dataset

PSO-SVM	Precision (%)	Recall (%)	F-Score
LOC	89.46	70.69	78.97
ORG	79.17	43.93	56.50
PERS	92.59	62.50	74.63
Overall	88.28	55.82	68.40

Table 9. SVM results using all features for the news dataset

SVM	Precision (%)	Recall (%)	F-Score
LOC	91.08	74.13	81.74
ORG	71.38	39.39	50.77
PERS	90.76	71.79	80.17
Overall	87.82	65.67	75.15

Table 10. Results with SVM-PSO for news dataset

PSO-SVM	Precision (%)	Recall (%)	F-Score
LOC	85.64	81.46	83.5
ORG	88.28	55.84	68.6
PERS	95.82	89.20	92.4
Overall	91.67	75.12	82.57

*Significant results are in **bold**.

6.3. PSO parameters sensitivity analysis

A number of different experiments are performed with various PSO parameters, and thus, sensitivity analysis is provided to determine the optimal PSO setting. Shi and Eberhart [36] defined the parameters ω as 0.7298 and $\Phi_1=\Phi_2$ as 1.49618. M. Erik *et al.* [37] proposed a single method of obtaining the different parameter values based on scenarios of optimization. On the basis of the detailed studies of previous work [38], [39], and after having carried out various experiments with a different parameterization of the validation set, the various PSO parameter values were finally determined.

The outcomes utilizing the five optimal combinations of parameters from the training dataset are presented in Table 11. Such combinations are referred to as PSO-1 and PSO-2. The swarm size (population) was set to 20 particles with a number of iterations of 60 for all our experiments.

Table 11. PSO-based feature selection results with different parameter settings

PSO-RUN	Parameter settings			Tweet's dataset	News dataset
	Inertia weight	Φ_1	Φ_2		
PSO-1	0.7298	1.49618	1.49618	68.40	80.15
PSO-2	0.3925	1.5586	1.3358	65.48	82.50
PSO-3	-0.4349	-0.6504	2.2073	63.79	79.58
PSO-4	0.4091	2.1304	1.0575	68.12	81.78
PSO-5	-0.3593	-0.7238	2.0289	67.41	80.16

*Significant results are in **bold**.

For the tweet dataset, the optimal outcome is achieved using the combination of inertia weight, learning parameter I, and learning parameter II at "0.7298", "1.49618", and "1.49618", respectively. For the news dataset, the best accuracy was obtained by setting the inertia weight to "0.3925", training parameter I to "2.5586," and training parameter II to "1.3358".

6.4. Comparisons with the existing systems

6.4.1. News dataset

This section compares our system to other state-of-the-art approaches that utilize similar datasets to conduct the experiments. The findings results demonstrate that the use of PSO on the basis of feature selection has considerably improved the system's overall performance, yielding an enhancement of 7.42 points in the F1 score. This approach was compared with three other models. The highest scoring model for this task is [40]. Their system employs a feature selection technique based on a genetic algorithm. They achieved 81.5 points in the F1 score. The second system is provided by [41]. They use a deep co-learning approach using semi-labels and BI-LSTM-CRF on top of the system. They achieved 74.1 points in the F1 score. Finally, the same news dataset was utilized by [21] to test their model, which provides multilingual features and knowledge bases in English via multilingual links. They achieved 63.9 points in the F1 score. Table 12 shows an overview.

Table 12. Comparison of our system with three other models on news dataset

Systems	Loc	Org	Pers	Overall
	F1	F1	F1	Avg. F1
K. Darwish 2013 [21]	73.1	42.1	69.5	63.9
C. Helwe 2019 [41]	81.6	52.7	82.4	74.1
B. Ait benali 2020 [40]	82.7	66.7	90.9	81.5
Our system	83.5	68.6	92.4	82.5

*Significant results are in **bold**.

6.4.2. Tweets dataset

Several different experiments representing the combination of multiple features have been performed on our model to determine their impact on the Arabic NER system on the tweet dataset. The results of these experiments are presented. Compared with the literature results, the system outperforms the state-of-the-art Arab NER systems applied to the tweet dataset, illustrated in Table 13. As a result, the approach outperforms previous approaches in terms of F1 for NE retrieval between locations, organizations, and persons in tweet, with an overall F1 of 68.4%.

We conclude that our approach using PSO based on feature selection significantly improves the Arabic NER task's performance. The Arabic NER systems in Table 13 employ similar corpora and assessment settings to those previously reported for this work, i.e., a Darwish dataset and assessment settings.

This study demonstrates that our new system outperforms the current system by 0.70%. The performance of the proposed method is improved due to the better SVM-based particle swarm optimization technique and its ability to efficiently handle overlapping features for existing systems.

Table 13. Comparison of our system with other models on tweets dataset

Systems	Loc F1	Org F1	Pers F1	Overall Avg. F1
K. Darwish 2014 [24]	76.7	55.6	55.8	65.2
A. Zirikly 2015 [42]	61.0	41.8	68.9	59.5
C. Helwe 2019 [41]	65.3	39.7	61.3	59.2
B. Ait benali 2020 [40]	72.4	46.7	73.7	65.6
B. Ait benali 2020 [43]	76.7	47.5	73.5	67.7
Our system	78.9	56.5	74.6	68.4

*Significant results are in **bold**.

6.5. Error analysis

The results obtained for each data set are analyzed for potential biases. The errors are classified in three ways, as shown:

- Boundary error: it is due to the incorrect recognition of the boundary of the entities; this type of case is mainly observed when entities take long and compound word forms.
- Invalid type of entity: it is issued once the entity has been correctly recognized, but belongs to its wrong class. This error is more significant for the tweet dataset, and this could be explained by the fact that most of the tweets are in dialect Arabic. By applying the PSO, we can see that the classification errors are considerably decreased. For news, the system mostly classified the named entity due to the news dataset where the tweets are composed of 2 languages, Dialect and modern standard Arabic.
- Missed entity: our system is missing a considerable amount of NE instances. We notice that false negatives are 983 and 287 for tweet and news, respectively. However, all these NEs are incorrectly categorized and classified as "other than NEs".

7. CONCLUSION

This paper presents a PSO-based feature selection technique for Arabic named entity recognition on social media. We evaluated our approach on publicly available social media datasets for Arabic NER systems, such as tweets and news datasets. It is concluded that a model with SVM as a classifier outperforms a restricted set of features compared to the model developed using a full set of features. The evaluation results show that using binary PSO to select features improves a classification model's accuracy and reliability on datasets. Furthermore, a thorough sensitivity analysis of the PSO parameters is performed, and their effects on the overall system performance are demonstrated. The effectiveness of our proposed technique is also proven by detailed comparative studies with other existing methods. However, our system suffers from a set of limitations. It can be challenging to specify the initial set of design parameters in certain cases, and it is challenging to address scattering issues. For future work, an optimization multi-objective (MOO) based feature selection technique could be developed to maximize more than a single measure of classification quality simultaneously. Besides, the effectiveness of the MOO for other types of datasets would also be measured.

REFERENCES




- [1] B. A. Ben Ali, S. Mihi, I. El Bazi, and N. Laachfoubi, "A recent survey of Arabic named entity recognition on social media," *Revue d'Intelligence Artificielle*, vol. 34, no. 2, pp. 125–135, May 2020, doi: 10.18280/ria.340202.
- [2] M. Shaheen and A. M. Ezzeldin, "Arabic question answering: systems, resources, tools, and future trends," *Arabian Journal for Science and Engineering*, vol. 39, no. 6, pp. 4541–4564, Jun. 2014, doi: 10.1007/s13369-014-1062-2.
- [3] J. Guo, G. Xu, X. Cheng, and H. Li, "Named entity recognition in query," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, 2009, p. 267, doi: 10.1145/1571941.1571989.
- [4] 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.
- [5] J. Cao, H. Cui, H. Shi, and L. Jiao, "Big data: a parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce," *Plos One*, vol. 11, no. 6, pp. 1–17, Jun. 2016, doi: 10.1371/journal.pone.0157551.
- [6] S. Cheng, Y. Shi, Q. Qin, and R. Bai, "Swarm intelligence in big data analytics," in *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, 2013, pp. 417–426, doi: 10.1007/978-3-642-41278-3_51.
- [7] T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning data mining, inference, and prediction*. New York, NY: Springer New York, 2001.

- [8] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, 1997, vol. 5, pp. 4104–4108, doi: 10.1109/ICSMC.1997.637339.
- [9] J. C. Bansal, P. K. Singh, M. Saraswat, A. Verma, S. S. Jadon, and A. Abraham, "Inertia weight strategies in particle swarm optimization," in *2011 Third World Congress on Nature and Biologically Inspired Computing*, Oct. 2011, pp. 633–640, doi: 10.1109/NaBIC.2011.6089659.
- [10] N. Chinchor and B. Sundheim, "Message understanding conference (MUC) 6: a brief history," in *COLING 1996: The 16th International Conference on Computational Linguistics*, 2003, vol. 1.
- [11] R. Ljung, "Hemophilia and prophylaxis," *Pediatric Blood & Cancer*, vol. 60, no. S1, pp. S23–S26, 2013, doi: 10.1002/pbc.24340.
- [12] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, "The automatic content extraction (ACE) program tasks, data, and evaluation," in *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, 2004, pp. 837–840.
- [13] G. Demartini, T. Iofciu, and A. P. de Vries, "Overview of the INEX 2009 entity ranking track," in *INEX 2009: Focused Retrieval and Evaluation*, 2010, pp. 254–264.
- [14] K. Balog, P. Serdyukov, and A. P. De Vries, "Overview of the TREC 2011 entity track," in *TREC*, 2011.
- [15] I. A. Al-Sughayer and I. A. Al-Kharashi, "Arabic morphological analysis techniques: a comprehensive survey," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 3, pp. 189–213, Feb. 2004, doi: 10.1002/asi.10368.
- [16] A. Abdul-hamid and K. Darwish, "Simplified feature set for Arabic named entity recognition," in *Proceedings of the 2010 Named Entities Workshop*, 2010, pp. 110–115.
- [17] N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh, "Morphological analysis and disambiguation for dialectal Arabic," in *Proceedings of NAACL-HLT*, pp. 2013, 426–432.
- [18] N. Habash, M. Diab, and O. Rambow, "Conventional orthography for dialectal Arabic," in *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 2012, pp. 711–718.
- [19] R. C. Eberhart and Y. Shi, "Comparison between genetic algorithms and particle swarm optimization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1447, 1998, pp. 611–616.
- [20] X. Yan, Q. Wu, H. Liu, and W. Huang, "An improved particle swarm optimization algorithm and its application," *International Journal of Computer Science Issues (IJCSI)*, vol. 10, no. 1, pp. 316–324, 2013.
- [21] K. Darwish, "Named entity recognition using cross-lingual resources: Arabic as an example," in *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference*, 2013, vol. 1, pp. 1558–1567.
- [22] K. Shaalan, "A survey of Arabic named entity recognition and classification," *Computational Linguistics*, vol. 40, no. 2, pp. 469–510, Jun. 2014, doi: 10.1162/COLI_a_00178.
- [23] A. De Sitter, T. Calders, and W. Daelemans, "A formal framework for evaluation of information extraction," *CiteSeerX*, 2004.
- [24] K. Darwish and W. Gao, "Simple effective microblog named entity recognition: Arabic as an example," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 2513–2517.
- [25] N. Alshammari and S. Alanazi, "An Arabic dataset for disease named entity recognition with multi-annotation schemes," *Data*, vol. 5, no. 3, p. 60, Jul. 2020, doi: 10.3390/data5030060.
- [26] Y. Benajiba and P. Rosso, "Arabic named entity recognition using conditional random fields," in *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, 2008, pp. 143–153.
- [27] P. F. Brown, V. J. D. Pietra, P. V. Desouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.
- [28] Y. B. Joseph Turian, Lev Ratinov, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 384–394.
- [29] J. D. T. Mikolov, I. Sutskever, K. Chen, and G. Corrado, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1389–1399.
- [30] B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu, "Evaluating word representation features in biomedical named entity recognition tasks," *BioMed Research International*, pp. 1–6, 2014, doi: 10.1155/2014/240403.
- [31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013, pp. 1–12.
- [32] G.-A. Levow, D. W. Oard, and P. Resnik, "Dictionary-based techniques for cross-language information retrieval," *Information Processing & Management*, vol. 41, no. 3, pp. 523–547, May 2005, doi: 10.1016/j.ipm.2004.06.012.
- [33] A. Pasha *et al.*, "MADAMIRA: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic," in *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 2014, pp. 1094–1101.
- [34] L. Ying Yang, J. Ying Zhang, and W. Jun Wang, "Selecting and combining classifiers simultaneously with particle swarm optimization," *Information Technology Journal*, vol. 8, no. 2, pp. 241–245, Feb. 2009, doi: 10.3923/itj.2009.241.245.
- [35] M. Muhammad, M. Rohaim, A. Hamouda, and S. Abdel-Mageid, "A comparison between conditional random field and structured support vector machine for Arabic named entity recognition," *Journal of Computer Science*, vol. 16, no. 1, pp. 117–125, Jan. 2020, doi: 10.3844/jcssp.2020.117.125.
- [36] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*, 1998, pp. 69–73, doi: 10.1109/ICEC.1998.699146.
- [37] M. Erik, H. Pedersen, and M. E. H. Pedersen, "Good parameters for particle swarm optimization," 2010. [Online]. Available: <http://www.hvass-labs.org/people/magnus/publications/pedersen10good-ppo.pdf>.
- [38] Y. Shi and R. C. Eberhart, "Fuzzy adaptive particle swarm optimization," in *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*, 2001, vol. 1, pp. 101–106, doi: 10.1109/CEC.2001.934377.
- [39] B. Alatas and E. Akin, "Rough particle swarm optimization and its applications in data mining," *Soft Computing*, vol. 12, no. 12, pp. 1205–1218, Oct. 2008, doi: 10.1007/s00500-008-0284-1.
- [40] B. A. B. Ali, S. Mihi, I. E. Bazi, and N. Laachfoubi, "Arabic named entity recognition based on treebased pipeline optimization tool," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 15, pp. 2963–2976, 2020.
- [41] C. Helwe and S. Elbassuoni, "Arabic named entity recognition via deep co-learning," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 197–215, Jun. 2019, doi: 10.1007/s10462-019-09688-6.
- [42] A. Zirikly and M. Diab, "Named entity recognition for arabic social media," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015, pp. 176–185, doi: 10.3115/v1/W15-1524.




- [43] B. A. Benali, S. Mihi, I. El Bazi, and N. Laachfoubi, "New approach for Arabic named entity recognition on social media based on feature selection using genetic algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 2, pp. 1485–1497, Apr. 2021, doi: 10.11591/ijece.v11i2.pp1485-1497.

BIOGRAPHIES OF AUTHORS






Brahim Ait Ben Ali    is a computer science Engineer, graduated from National School of Applied Sciences (ENSA at Cadi Ayyad University of Marrakesh, Morocco). Since 2019, He is preparing his Ph.D. in the IR2M Laboratory, Department of Computer Science, Faculty of Sciences and Techniques, Settat, Morocco, at Hassan first University of Settat. He has published several papers in reputed journals and international conferences. His research interest is Machine Learning and Deep Learning for Natural Language Processing and its application. He can be contacted at email: b.aitbenali@uhp.ac.ma.






Soukaina Mihi    holds an engineer degree from Cadi Ayad University and a Masters degree from INSA Lyon in Artificial Intelligence. She is a PhD student at the IR2M Laboratory, which stands for Informatics, networks, Mobility and Modeling in Faculty of Sciences and Technologies Hassan 1st University, Settat, Morocco. Her research interests are Deep Learning and Machine Learning. Her current research focus is on NLP and Sentiment Analysis especially in Arabic. She can be contacted at email: soukaina.mihi@uhp.ac.ma.



Ismail El Bazi    holds a Doctorate in Computer Science from Hassan 1er 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: ismail.elbazi@umsba.ac.ma.



Nabil Laachfoubi    is a computer science professor at Hassan 1st University of Settat, Morocco. He defended his doctoral thesis in 2000 and continues research in various areas notably machine learning and computer vision. He published several papers in reputed journals. He can be contacted at email: nabil.laachfoubi@uhp.ac.ma.