International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 2, April 2021, pp. 1485~1497
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i2.pp1485-1497

New approach for Arabic named entity recognition on social media based on feature selection using genetic algorithm

Brahim Ait Benali[1], Soukaina Mihi[2], Ismail El Bazi[3], Nabil Laachfoubi[4]
[1,2,4]IR2M Laboratory Faculty of Sciences and Techniques, Hassan First University of Settat, Settat, Morocco
[3]Sultan Moulay Slimane University, National School of Business and Management, Beni Mellal, Morocco

Article Info

Article history:

Received Mar 13, 2020
Revised Jul 3, 2020
Accepted Nov 2, 2020

Keywords:

Arabic dialect
Genetic algorithm
Named entity recognition
feature selection
NLP
Social media
Support vector machine

ABSTRACT

Many features can be extracted from the massive volume of data in different types that are available nowadays on social media. The growing demand for multimedia applications was an essential factor in this regard, particularly in the case of text data. Often, using the full feature set for each of these activities can be time-consuming and can also negatively impact performance. It is challenging to find a subset of features that are useful for a given task due to a large number of features. In this paper, we employed a feature selection approach using the genetic algorithm to identify the optimized feature set. Afterward, the best combination of the optimal feature set is used to identify and classify the Arabic named entities (NEs) based on support vector. Experimental results show that our system reaches a state-of-the-art performance of the Arab NER on social media and significantly outperforms the previous systems.

Corresponding Author:

Brahim Ait Benali
IR2M Laboratory, Faculty of Sciences and Techniques
Hassan First University of Settat
Casablanca Street, Box 577, Settat, 26000, Morocco
Email: aitbenali.brahim@gmail.com

1. INTRODUCTION

The named entity recognition is a subtask of natural language processing used for identifying and extracting useful information such as names of people, places, and organizations from unstructured text [1]. This approach was first introduced at the *"Sixth Conference on Understanding Messages-MUC6-"* [2]. This focused mainly on extracting information from formal text. A number of papers on these systems were first concerned with the English language, then a series of publications on other languages, namely German, Spanish, Dutch, Japanese and Indian, etc.... Regarding the Arabic language, the application of this concept was initiated only in 2005 [3].

It is stated that most of the research projects around this topic in the different languages have achieved a very high level of performance comparable to that of human subjects, especially in English [4]. Many typical machine learning applications result from the complex relationships between features (also known as input variables or characteristics). A feature is a property or attribute of data that may be employed by algorithms, for instance, in the field of machine learning to extract valuable information from data sets. Each item of data in an application has specific characteristics. For a particular application, all or a subset of the extracted features is used to obtain a meaningful result.

As more and more data becomes available, and dozens or thousands of features are available for individual datasets, system complexity increases not only in terms of understanding the data but also in resource usage and system efficiency [5]. Although dataset size is not controllable, the feature set can be

reduced by including only relevant and unique features, so that overall efficiency improves and resource use decreases [6]. Redundant or irrelevant features may take the form of correlated features in which there is a dependency between them. Dependent features may provide no additional information or have an impact on output. This means that the elimination of such a feature does not influence the total information content. In certain cases, these characteristics can introduce a bias in the system and thus affect performance. Since there can be N possible characteristics for a dataset, there can be 2N combinations of characteristics to be tested to see which features contribute positively to the problem outcome. Evolutionary algorithms, like genetic algorithms (GA) [7], may be used for feature selection, where a subset of features must be found from a very large search space. Through this paper, we study the drawbacks of taking into account a large number of features by implementing a feature selection solution based on the GA and using it in a system to classify named entities according to their categories.

The objective of this work aims to evaluate current works that use evolutionary algorithms for feature selection, as well as to provide a novel GA-based approach for selecting feature subsets, and to apply the proposed solution to classify named entities. We plan to publish this system in Github. The rest of our paper is described as follows: In section 2, we discussed the NER and its different applications. Section 3 is devoted to some challenges of Arabic named recognition, especially on twitter. Afterward, in section 4, we present the literature review related to NER systems in the context of feature selection using GA. Section 5 and 6 are about feature selection techniques and the genetic algorithm used throughout this paper. Furthermore, suitable ways to formulate our problematic are presented in Section 7. In Section 8 and 9, we present our contribution concerning Arabic NER based on feature selection using GA. Then, we end with a concluding part.

2. BACKGROUND

2.1. What is NER?

A named entity is a word or expression that uniquely describes an element among a set of other items with similar attributes. Names of named entities include names of organizations, persons, and places in the global domain, names of genes, proteins, drugs, and diseases in the biomedical area. Named entities recognition (NERs) is defined as the process of identifying and classifying the entities named in the text into predefined entity categories [8]. Formally, considering a series of tokens S = {W1,…,W2,WN}, NER aims to output a list of tuples, each tuple specify the entity referred to a named entity in S here, $I_s \in [1, N]$, and $I_e \in [1, N]$ are the start and end indexes of a named entity mentioning; t is the entity type of a set of predetermined categories.

2.2. The broader role of NER

The general NER implications related to NLP research are too essential to list. Some examples of uses for which NERs are relevant are presented in this section.

- Information retrieval: This consists of identifying and extracting relevant documents from a data set based on an incoming request. A research study by [9], noted that approximately 71% of all search engine queries contain NE. Information extraction can be facilitated by the NER in two phases [10]. First, find the NE in the request; second, identify the NE in the searched documents, and then retrieve the relevant documents with consideration for their classified NEs and how they are linked with the request.
- Question answering: This is quite similar to information extraction, but with over advanced findings. A question-answering approach utilizes questions as an entry and provides short and precise answers in return [11]. Besides, in order to facilitate the recognition of NEs in a query, the NER task can be used during its analysis phase. This will, therefore, enable us to identify and locate the relevant documents and even to provide an appropriate answer.
- Machine translation: This refers to the automatic translation of a text from one natural language into another [12]. NEs require special consideration regarding which parts of a NE need to be translated and which sections need to be transliterated [13]. For instance, people's names tend to be transliterated in the case of a location name. The name portion and the category portion (e. g., mountains) are generally transliterated and translated, each one respectively.
- Text clustering: The clustering of the research results can exploit the NER by sorting the resulting clusters according to the ratio of entities that each cluster contains [10]. This improves the process of analyzing the nature of each cluster and also enhances the clustering approach in terms of selected characteristics.

- Navigation systems: Such systems, which make it easier to navigate using digital maps, have now become an essential part of our lives [14]. They give indications, information on neighboring places, which may be related to other online resources, and circulation conditions. In these systems, points of interest (called "waypoints") consist of NEs saved in a database with their geo-coordinates [15].

3. ARABIC NAMED RECOGNITION CHALLENGES

3.1. Arabic challenges

The United Nations recognize Arabic as one of the main languages of the world. It is mostly used around the world by 300 million people in 28 countries. It is part of the Semitic family of languages, which also includes Hebrew and Amharic, the principal language of Ethiopia.

Three different kinds of Arabic language are available:

- Classical Arabic: The language of the Holy Quran;
- Modern standard Arabic (MSA): The language of official documents, newsletters, education. Traditionally, it is the common Arabic language of all Arabs.
- Colloquial Arabic: or dialectal Arabic, it is the informal language that people use for daily communication; it differs from one country to another. In general, five dialects of colloquial Arabic may be distinguished: Egyptian, Levantine, Maghrebian, Gulf, and Iraqi [16].

Arabic is written from right to left. However, unlike English or French, there are no "capital" letters. Arabic morphology is very complex [17]. From a root, we can extract words that are lexically and semantically different. As an example, the words "madrassa" and "modarissa" are taken from the root "d-,r-,s- درس" which is written in the same way in Arabic ٍ) gninaem tnereffid a htiw tub ةسدرمmadrassa = school, modarissa = educator).

3.2. Arabic twitter challenges

Social media networks contain a huge amount of unstructured data. In fact, tweets present multiple challenges for analysis compared to standard text. In the Arab world, the majority of people write social media content informally, sometimes in a mixture of bilingual languages, using Latin words within an Arabic tweet. Besides, there are non-Arabic words written in Arabic letters, more often in Maghreb dialects, people use Latin letters to write familiar Arabic words. Another challenge is to repeat the letters inside or at the end of a word, for example, to call a name. For the named entity, as an example of repeating a letter when calling, for instance, نييمأ == نييييمأ.

4. RELATED WORK

Several studies have focused on finding the best combinations of features of NERs task. Mainly, the authors have focused on one hypothesis: the difficult task of identifying the best features based on NER classes (i.e., PER, ORG, LOC, etc.). In fact, [18-22], employed GA as feature selection approach to identify the optimized feature set. Since there are therefore many classes, the process of determining the appropriate characteristics for these classes is a difficult task. In this way, GA was applied to the CoNLL-2003 reference dataset. The results showed that the application of the GA has a significant impact in terms of identifying the best combination of characteristics that significantly improved the accuracy of the classification.

Therefore, many features are used for named entity recognition. This dimensionality of features requires a selection to determine the optimized feature set. This is due to the variations in strength between these features, some of them can be insignificant in some cases, and others can be powerful. In addition to this, [23] proposed a feature selection approach for reducing the feature dimensionality employed in NERs. Thus, a genetic algorithm (GA) is used to identify the best combination of features. In this, multiple features were used, including POS tagging, word length, affixes, and word frequency. These characteristics were coded as chromosomes for the identification of the features optimized using GA. After that, the best combination is used to classify the entities based on a maximum entropy classifier (ME). The results showed that the performance of the ME with the best combination of characteristics surpassed the performance of the ME with all its characteristics. Regarding [24], They propose a method to identify an optimal feature set for extracting NEs from company web pages using GA. The functionality dimension contains text elements such as POS tags, keywords, and capitalization. Similarly, it provides web features such as fonts, URL, and block position. Therefore, an SVM classifier is applied to classify NEs. The results of the classification showed an improvement in performance when using the optimized feature set.

Similarly Le and Tran [25] Thanh and Tran [26] provides the implementation of a genetic algorithm for the selection of a subset of optimal features that are used in the maximum entropy classifier for the NER task. Various strategies have been proposed to reduce the GA computation time, such as (i) reducing the

population size after a few generations; (ii) parallel computation of the fitness of the individuals in each generation; and (iii) progressive sampling to obtain the optimal sample size of the training data.

The SVM classifier has been used successfully by classifying data instances directly into their actual classes. The evaluation explains that the SVM classification is performed better with the optimized feature selected by GA. For example, [27] introduced a gene selection approach that uses a hybrid combination of genetic algorithms and support vector machines. The primary aim of this hybridization was to exploit their respective merits fully (e.g., robustness to the size of the solution space and the capacity to process very large size of characteristic genes, etc.) to extract essential feature genes (or molecular signatures) for a complex biological phenotype.

In this manner, [28] developed a fine-grained NE corpus for Dutch that consists of 6 types of NE: person, place, organization, product, events, and miscellaneous. They developed the Dutch NE system following the approach of the set of classifiers. They used three classifiers in the development of the system: memory-based learning (MBL), conditional random fields (CRF), and support vector machines (SVM). These classifiers were trained on the features, and the result is a set using various voting mechanisms in the genetic algorithm. Their classifier obtained an F-score of 84.91%. Moreover, [29], they have provided a wide variety of features to identify NEs from biomedical texts. They use two powerful and diverse classification methods, such as the conditional random field (CRF) and the supporting vector machine (SVM), to construct many models based on various representations of the feature set and/or feature models.

The CRF and k-nearest classifiers were used in combination with the genetic algorithm for this named entity classification task. For instance, [30] proposes a multi-objective approach for the extraction of biochemical entities based on modified differential evolution, feature selection, and set of classifiers. The algorithm works in two layers. The first layer refers to the identification of adequate features set for the task to be performed within a supervised statistical classifier, i.e., the conditional random field (CRF), it produces a set of solutions, a subset being used to build a set in the next layer. The approach proposed is evaluated for the extraction of entities in chemical texts, which involves identifying the names of IUPAC and its similar and classifying them in some predefined categories. Experiments conducted on a reference data set show recall, accuracy, and F-measure values of 86.15%, 91.29%, and 88.64%, respectively. In [31], The author employed a methodology to carry out a selection of characteristics for the classification task in named entity recognition using a multi-objective genetic algorithm. They have tested this approach with the application of a weak Pareto tournament genetic selection algorithm and a k-nearest neighbors machine learning algorithm. They showed its efficiency on three real-world data sets. They demonstrated that the multi-objective algorithm is well adapted to feature selection and has the benefit of generating different solution options.

This technique also has been used to deal with multi-class problems. Regarding [32], they provide an improved method of feature selection based on a genetic algorithm for unbalanced data from several classes. This approach improves the fitness function by using the EG-mean evaluation criterion instead of the precision of the overall classification to select traits that are suitable for the recognition of minor classes. Thus, [33], they have implemented genetic algorithms (GA) to tackle the problem of multi-class prediction. They describe a GA-based gene selection scheme that automatically determines the members of a predictive gene group, as well as the optimal group size, which maximizes the success of the classification using a maximum likelihood classification method (MLHD).

Recently, [34] proposed the genetic algorithm (GA) to minimize the computational time required to find the relevant and informative features of the Arabic text required for classification. The SVM was utilized as a machine-learning algorithm to evaluate the accuracy of Arabic named entity recognition. At the same time, [35] developed a new heuristic approach based on the genetic algorithm to under-sample the training data that is used for NER. Regarding the fact that the training patterns of the NER are proper sentence forms, the approach developed also takes this issue into account and has applied it to individual sentences from the training data.

5. FEATURE SELECTION

Feature selection is a preprocessing technique that helps to determine the critical features of a particular problem. Traditionally, it has been used across various fields, including biological data processing, finance, intrusion detection systems, and NLP problems. Feature selection represents an important technique applied in reducing dimension reduction; using such a method, relevant features are chosen, and irrelevant and redundant features are eliminated [36]. Reducing the dimensionality of inputs may enhance performance by decreasing the learning speed and complexity of models or by increasing the generalizability and accuracy of classification. Choosing appropriate characteristics can also lower the cost of measurement and enhance the problem understanding. In some cases, the impact of feature selection can be impressive. For instance, in the analysis of microarray data, it is possible to utilize only 2 of the 7129 features to enhance classification [37].

The main reasons for using feature selection are as follows:

- It allows the machine learning algorithm to train faster.
- It reduces the complexity of a model and facilitates its interpretation.
- It enhances the accuracy of a model if the correct subset is selected.
- Avoids overfitting.

Due to the benefits mentioned above, feature selection is actually employed in real-world problems, mainly classification and regression problems. Feature selection has been effectively applied to resolve problems in various areas, including microarray analysis, image classification, facial recognition, and text classification [38]. The feature selection methods are also categorized into filters, embedded, and wrappers methods, according to the relationship with the training method [36]. Filters are totally independent of any training method, as the focus is on the general characteristics of the data. They are not computationally expensive and have an excellent generalization capability due to their independence from the induction algorithm. Both wrapper and embedded methods require a learning method to perform feature selection. For wrappers, an induction method evaluates subsets of best candidate features. Embedded methods lie between filters and wrappers because the selection is part of the induction method's training process.

6. GENETIC ALGORITHM

Genetic algorithms are part of the evolutionary family of algorithms as shown in Table 1. Their purpose is to find an approximate solution to an optimization problem in cases where no exact solution exists (or where the solution is unknown) in order to solve the problem within a suitable period of time. Genetic algorithms utilize the philosophy of natural selection as a concept that can be applied to a population of potential solutions to a specific problem. The solution is approached by successive "jumps", as in a separation and evaluation procedure, except that formulas are searched for instead of direct values. The key GA processes are as follows: population initialization, fitness calculation, selection, crossover, mutation, and termination criteria.

Table 1. Pseudo code of the genetic algorithm

```
1: Define the parameters
2: Selecting encode method
3: Generating the initial population
4: While I < Max-Iter and Best-fitness < Max-Fitness do
5:     Fitness calculation
6:     Selection
7:     Crossover
8:     Mutation
9: end while
10: decode the individual using maximum fitness
11: return the optimal solution
```

6.1. Fitness function

The fitness function is utilized to calculate the degree of fitness related to every chromosome that represents the degree of fitness of the current solution. The evaluation of the chromosome's fitness must take into account two factors, namely the number of selected characteristics and the accuracy of the classification when using the subset of selected features. Thus, the fitness function is presented as follows:

$$f(C_i^j) = J(C_i^j) - w|C_i^j| \quad (1)$$

where C_i^j is the i-th chromosome for the current population at j-th generation; $J(C_i^j)$ is the accuracy of the classification based on the solution represented by C_i^j ; $|C_i^j|$ refers to the number of features selected that could be calculated based on counting the non-zero genes in the chromosome. C_i^j; w is the weight of $|C_i^j|$ Which is an empirical value.

6.2. Selection operation

A new population is generated following each evolution. In order to choose the parent chromosomes to mate, a selection probability needs to be calculated. The probability of selection may be defined as follows:

$$P(C_i^j) = {f(C_i^j)}/{\sum_{k=1}^{M} f(C_i^j)} \quad (2)$$

where f is the fitness value, and M is the size of the current population.

6.3. Crossover operation

The crossing operation generates novel chromosomes through the rearrangement of the genes of the parent chromosomes. The crossing operations generally used are single-point crossing, two-point crossing, multi-point crossing, and so on. During our research, we have chosen a single-point crossing to generate new solutions. Figure 1 shows an overview of the operation carried out.

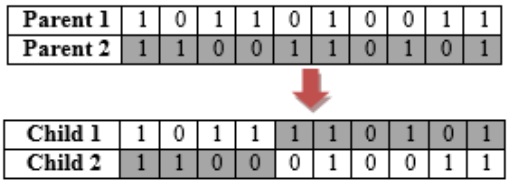

Figure 1. The crossover operation

6.4. Mutation operation

The mutation process is carried out immediately at the end of the crossing operation as shown in Figure 2. Its purpose is to maintain population diversity in each successive generation and to expand the research space. Mutations can be achieved through random inversion of one or more bits of a single parent chromosome to produce a new child. A variety of mutation processes have been initiated by researchers, such as mutation by inversion, mutation by exchanging neighbors, etc.

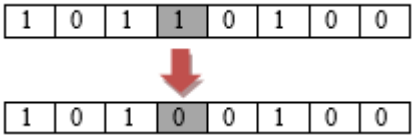

Figure 2. The mutation operation

6.5. Terminate criterion

A GA reiterates the evolution iteration to reach the optimal solution until a predetermined end criterion is reached. Some benefits of the genetic algorithms of this method are as follows:

- They generally outperform traditional feature selection techniques.
- Genetic algorithms can process datasets having numerous features.
- They do not require particular knowledge of the problem being studied.
- Such algorithms may easily be parallelized in computer groups.

And some drawbacks are:

- Genetic algorithms can be costly to compute because the evaluation of each individual requires the construction of a predictive model.
- Genetic algorithms can be expensive to compute because evaluating each individual requires the construction of a predictive model.
- These algorithms can take a long time to converge because they contain stochastic algorithms.

7. FEATURE SELECTION USING GA

Feature selection, or selection of inputs, refers to the process of searching for the most relevant inputs for a given predictive model. These methods can be used to find and eliminate unnecessary, irrelevant, and redundant characteristics that do not add to or reduce the accuracy of the predictive model. Mathematically, input selection is expressed as a problem of combinatorial optimization. Therefore, the

function to be optimized is the predictive model's generalization performance, as represented through the error on the chosen instances of a data set. The input variables are the inclusion (1) or exclusion (0) of the input variables.

A comprehensive selection of characteristics would allow a large number of different combinations to be evaluated (2^N, where N is the number of features). This process requires a lot of computational work, and, if the number of characteristics is large, it becomes impractical. This is why we need intelligent methods for selecting characteristics in practice.

The genetic algorithm represents a very sophisticated algorithm for the selection of features. It is a stochastic approach to function optimization based on the mechanics of natural genetics and biological evolution. In this paper, we demonstrate that genetic algorithms may be employed to optimize the performance of a predictive model for selecting the best features. In nature, organisms' genes change over consecutive generations in order to become better adapted to the environment. The genetic algorithm is a heuristic optimization process based on natural evolutionary procedures

8. EXPERIMENTAL RESULTS

8.1. Evaluation metrics

The standard NER evaluation script of the CoNLL was used to evaluate the proposed approach. As detailed in [39], the valuation methods of the CoNLL are particularly aggressive, as no partial credit is allowed to a partially extracted named entity. The outcomes obtained after the execution of the CoNLL evaluation script are presented in terms of precision, recall, and f-score for each NER class [40].

- True positive (TP): Entities recognized by the NER that match the truth on the ground.
- False positive (FP): Entities recognized by NER that do not match the truth on the ground.
- False negative (FN): Entities marked in the Basic Truth but not recognized by NER.
- Accuracy measures the capacity of a NER system to show only the correct entities, and Recall measures the capacity of a NER system to recognize all entities in a corpus.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

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

$$Fscore = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

8.2. Cross-validation

In this work, cross-validation was used k times to evaluate the quality of the solution achieved using the genetic algorithm, where k is equivalent to five. The data set is divided arbitrarily into five sub-samples with equal size. Only one sub-sample was specified as a validation set to be used in testing performance, and then the k-1 sub-samples were employed as a training set. This procedure was repeated five times. In each fold, each k sub-sample was used exactly once as a validation set. The k results of the folds were then averaged to provide a single score.

8.3. Pre-processing

Normalization is carried out after removing punctuation, diacritics, and stop words from the text. We employed diacritics suppression, punctuation suppression, and Arabic normalization provided by the AraNLP library for the preprocessing of texts. AraNLP developed by [41, 42] described stop words as "words that have no significant semantic relationship with the context in which they exist" [43]. We have established a list of the most frequent stop words that have occurred in corpora. It comprises prepositions, conjunctions, punctuation marks, and numbers. Some Arabic stop word examples are: "(In) "(في)," (who) الذي, and (he/she)" هو\هي.

After we used the word stemming in Arabic, that means the process of deleting all prefixes and suffixes from a word to generate the stem or root. After that, we applied a lemmatization process. This is a process of converting the plural into the singular or deriving a verb from the gerund form. Other possibilities include deriving the root from model words. This process of derivation is important for classifiers and index builders/researchers because it reduces dependency on particular word forms and reduces the potential size of vocabularies, which otherwise might have to contain all possible forms. In our work, we used the MADAMIRA tool [44].

8.4. Datasets

In order to evaluate our model, we use the training and test datasets developed by [45]. These datasets have been labeled using three different kinds of named entities in order to evaluate our model, we use the training and test datasets developed by [45]. These datasets have been labeled using three different kinds of named entities: location, person, and organization as shown in Table 2. The training dataset consists of randomly chosen tweets from May 3-12, 2012. The test data set includes tweets that were randomly selected from November 23, 2011, to November 27, 2011. Note that these two datasets were tagged according to the linguistic data consortium's ACE tagging guidelines. This data set was used for testing in [46-49], as we will see in the experimental results.

Table 2. Twitter evaluation data statistics

	Tokens	Person	Location	Organization
Train set	55k	788	713	449
Test set	26k	464	587	316

8.5. Named entity features

The principal features are identified according to the various combinations of word and tag contexts available. We employ the following features to build NER system based on feature selection using a genetic algorithm. Besides, we defined a semantic-driven feature that has been very successful in improving the system's overall performance:

a. Context words (CTX): These are the words that precede and follow the current word. They are using the observation that surrounding words provide useful information for the identification of NEs. In our case, we take five words in total [-2, 2].
b. Word suffix and prefix (LEX): Suffixes and prefixes of fixed-length words (say, n) are very useful to identify NEs and perform nicely for Arabic languages having strong inflection.
c. Gazetteer (GAZ): A binary feature is used to identify the presence of the word in a single Gazetteer. In our system, the gazetteer employed is the combination of (i) ANERGaz: as proposed by [50], that contains 2183 LOCs, 402 ORGs, and 2308 PERs; and (ii) WikiGaz: Wikipedia extensive gazetteer [46], that contains 50141 LOCs, 17092 ORGs, and 65557 PERs.
d. Morphological features (MORPH): can mostly indicate the absence of named entities. For example, Arabic allows the attachment of pronouns to nouns and verbs. However, pronouns are rarely associated with entities named. In our case, such features are generated by the MADAMIRA tool [44]. Five morphological features have been selected to be used in this work:
   – Aspect: Refers to the aspect of an Arabic verb. It can have four values: command, imperfect, perfective, not applicable. However, as none of the NEs may be verbal, we apply this characteristic as a binary feature specifying whether a word is labeled for aspect or not;
   – Gender: The nominal gender. There are three values: female, male, not applicable;
   – Person: It shows information about the person. Possible values are 1, 2, 3, Not applicable. As for the aspect, we apply it as a binary feature that indicates if a word is marked for the person or not;
   – Proclitic2: The proclitic conjunction. We used a tool that generates nine different values for this feature: no proclitic, not applicable, Conjunction fa, Connective particle fa, Response conditional fa, Subordinating conjunction fa, Conjunction wa, Particle wa, Subordinating conjunction wa;
   – Voice: The verb voice. It has the following values for this feature: active, passive, not applicable, indefinite.
e. Part of speech (POS) tags: POS tags indicate (or counter-indicated) the possible presence of a named entity at word level or at the word sequence level. We used POS tags provided by the MADAMIRA tool.
f. Word2vec cluster IDs (W2V Cluster): Word2vec is an algorithm for learning embedding utilizing a neural network model proposed by [51]. Embeddings are represented as a set of latent variables, with a specific instantiation of these variables representing each word. As part of our system, we employ the K-means method on word vectors and utilize cluster IDs as characteristics.
g. Brown clustering IDs (BC ID): As provided by [52], brown clustering is an approach of hierarchical clustering to maximize the mutual information of bigram words. Word representations, particularly brown clustering, have been shown to enhance the NER system's efficiency by adding them as a feature [53].
h. Word2vec (W2V): Word embedding derived from untagged text has successfully proven valuable for numerous NLP tasks, especially for part-of-speech (POS) tagging [54], named entity recognition [55],

chunking [53], and parsing [56]. In large corpus, names appear in regular contexts that will be profitable for most sequence tagging tasks: such as NER. So that we could initialize our word vectors with pre-trained word embedding, [57] Demonstrate that the use of embedded words may encode morphological information and can add additional information to the embedding of character-based words.

In order to evaluate the performance of pre-trained word embedding, we carried out four experiments using various sets of publicly available word embedding sets and compared the obtained findings with a random sampling method to initialize our model. Table 3 demonstrates the obtained outcomes using the four different word embedding and the random sampling method. Based on the results in Table 3, we obtained a meaningful enhancement using preformed words compared to random words. In our work, we used AraVec, which proposed by [40] and is a pre-trained open-source word embedding project that aims to provide the Arabic NLP research community with powerful and freely usable words embedding models. In the rest of our paper, we used the word-embedding model pre-trained by more than 1,476,715 tokens on Twitter with the Skip-Gram algorithm, and we use an embedding dimension of 300 vectors.

Table 3. Results with a different choice of word embeddings

Embedding	Dimension	F1-Score
Random Initialisation	300	60.97
FastText	300	62.12
Wikipedia2Vec	300	65.45
AraVec	300	**67.70**

8.6. Proposed approach

The proposed method consists of five principal steps, as shown in Figure 3. The first step is associated with the data to be used in the experiment in which a Darwish's dataset is used for this purpose. The second step aims to perform a transformation task to normalize and segment the data into a suitable form for processing. The third step is to extract features from the transformed form of the data. The fourth step consists of the contribution of this study in which a genetic algorithm is performed on the extracted features in order to select the best combination of features. Finally, the fifth step is associated with the classification process in which a support vector machine (SVM) classifier is used to classify the named entities.

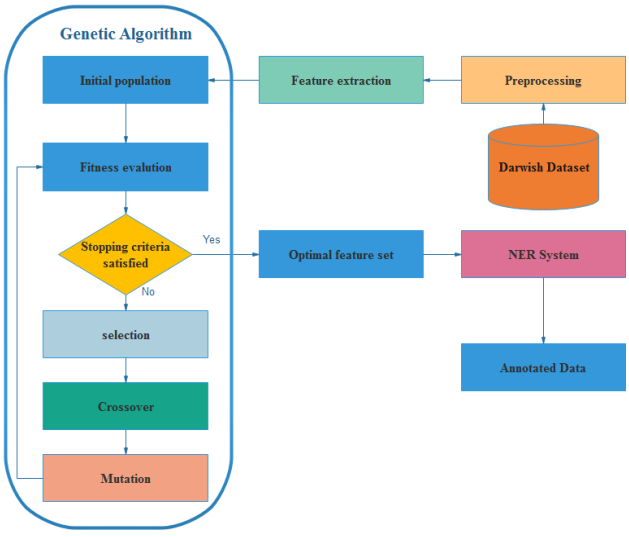

Figure 3. The proposed approach

9. RESULTS AND DISCUSSIONS

9.1. Effects of feature selection

This evaluation aimed to determine the capacity of the proposed approach to address colloquial Arabic text utilized on Twitter. The SVM approach was used in two steps to study the effect of feature selection utilizing GA, firstly, with the set of optimized features obtained from GA, and secondly, the SVM was performed with all features. The evaluation was carried out using common information retrieval

measures accuracy, recall, and F-measure. In Tables 4 and 5, we present the SVM and SVM with GA outcomes using accuracy, recall, and f-measure for all classes, including person, organization, and location.

Notably, as shown in Tables 4 and 5, using the SVM with GA surpassed the SVM with all features for all classes, including person (73.5% vs. 59.11%), organization (47.59% vs. 41.79%) and location (76.79% vs. 64.34%). Finally, in terms of the average percentage of f-measure, the SVM with GA outperformed the SVM with 67.70% vs. 56.21%. The evaluation of the proposed approach has been split into two stages; the first SVM will be implemented using the entire set of features; the second SVM will be implemented using the optimized feature set. The results showed that the application of the SVM with GA as the feature selection approach outperformed the application of the SVM without GA by reaching 67.7% of f-measure. This emphasizes the usability of the structure analysis in terms of identifying a set of optimized characteristics for NERs. In this way, the objective of identifying a set of optimized features for the extraction of NERs from Darwish's data set is achieved.

Table 4. Results with SVM and all features

SVM	Precision (%)	Recall (%)	F-Score
LOC	81.63	53.10	64.34
ORG	50.28	35.76	41.79
PERS	78.11	47.55	59.11
Overall	71.30	46.39	56.21

Table 5. Results with SVM-GA

GA-SVM	Precision (%)	Recall (%)	F-Score
LOC	88.66	65.80	**76.79**
ORG	63.73	37.98	**47.59**
PERS	82.40	66.34	**73.50**
Overall	80.58	58.37	**67.70**

Significant results are in **bold.**

9.2. Comparisons with existing systems

We carried out numerous experiments that combine various models and architectures to understand their impact on the Arab NER system on social media. Table 6 illustrates the outcomes of these experiments. Experiments show that the tremendous increase in overall system performance was achieved using the combination of GA-based feature selection, yielding an improvement of 2.5 points in the F1 score. We compare our system with four other models. The highest score reported on this task was achieved by [46]. Their system uses simple effective language-independent approaches based on using extensive gazetteers, domain adaptation, and a two-pass semi-supervised method. They scored 65.2 points in the F1 score. The same Twitter dataset was used by [47] to test their model, which adopted a supervised machine learning approach by using the Conditional Random Fields sequence labeling, word embedding, and word representations. They scored 59.59 points in F1. The third system [48], implemented a hybrid approach to extract Arab person names from tweets and resolve their ambiguity by utilizing contextual bigram models. They scored 66.75 points in the F1 score in the extraction of a person's names. The fourth system [49] used a deep co-learning approach using semi-labeled and BI-LSTM-CRF on the top of the system; they scored 59.2 points in the F1 score. Our model outperformed these four models. Table 6 presents our findings on the Arab NER for social media compared to these systems.

Table 6. Comparative evaluation results

	Entity	Precision	Recall	F-score
Darwish Kareem	LOC	83.60	70.80	76.70
2014 [46]	ORG	76.40	43.70	**55.60**
	PERS	67.10	47.80	55.80
	Overall	76.8	56.6	65.20
Ayah Zirikly	LOC	*	*	61.03
2015 [47]	ORG	*	*	41.28
	PERS	*	*	68.92
	Overall	81.70	46.90	59.59
Omnia Zayed	LOC	*	*	*
2015 [48]	ORG	*	*	*
	PERS	81.90	56.32	66.75
	Overall	*	*	*
Chadi Helwe	LOC	*	*	65.30
2019 [49]	ORG	*	*	39.70
	PERS	*	*	61.30
	Overall	*	*	59.20
Our system	LOC	88.66	65.80	**76.79**
	ORG	63.73	37.98	47.59
	PERS	82.40	66.34	**73.50**
	Overall	80.58	58.37	**67.70**

Significant results are in **bold**.

The existing Arab NER systems presented in Table 6 utilize the same corpus and evaluation parameters as those described in this work, i.e., the Darwish's dataset and evaluation parameters. The results obtained confirm that the proposed system outperforms by 2.50% of the existing models. The reasons for the higher performance of the proposed system are the better optimization technique of the genetic algorithm based on SVM and its ability to deal with overlapping features compared to existing systems effectively. As shown in Figure 4, it summarizes the results for the overall state of the art system that exists in the literature.

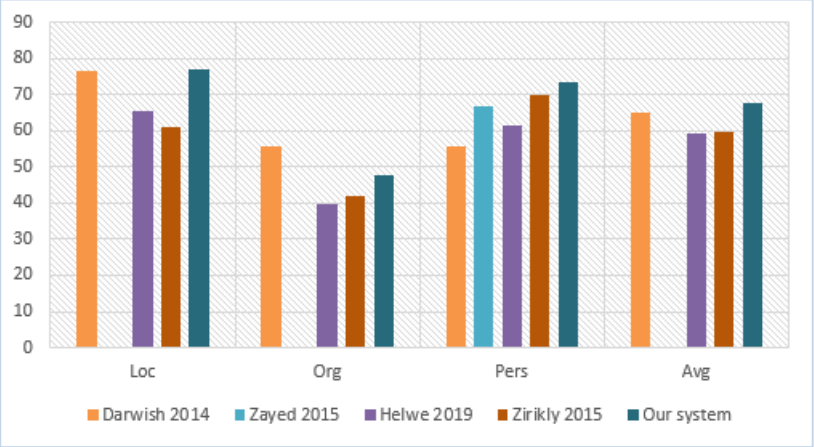

Figure 4. TWEETS test set results

10. CONCLUSION

In this paper, we have shown that our system, which uses feature selection and genetic algorithm, achieves state-of-the-art results by creating an Arabic named entity recognition system on social media and significantly outperforming the previous state-of-the-art system. The experimental results achieved demonstrate the benefits of using feature selection techniques to enhance the performance of our system. The reported results demonstrate that a strategy of adaptive trait selection based on genetic algorithms can considerably decrease the number of features needed to train our system and, at the same time, improve the performance in recognition of the named entity. This is one step towards applying machine learning techniques to automate the construction of classification systems for severe text processing problems. Our future work is to find some strategies to minimize GA calculation time, like improving the method of calculating the ability of individuals to calculate GA. Rather than calculating individuals' actual ability, these values can be predicted according to the ability of individuals in previous generations of GA.

REFERENCES

[1] B. Ait 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, 2020.
[2] B. S. Grishman, Ralph, "Message Understanding Conference -6: A Brief History," *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, vol. 1, 1996, pp. 466-471.
[3] F. Huang, "Multilingual Named Entity extraction and translation from text and speech," PhD thesis, Lang. Technol. Institute, Sch. Comput. Sci. Carnegie Mellon Univ., 2005.
[4] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 363-370.
[5] E. M. Mashhour, E. M. F. El Houby, K. T. Wassif, and A. I. Salah, "Feature selection approach based on firefly algorithm and chi-square," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 4, pp. 2338-2350, 2018.
[6] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014.
[7] Mitchell Melanie, "Introduction to genetic algorithms," *Proceedings of GECCO 2007: Genetic and Evolutionary Computation Conference, Companion Material*, 2007, pp. 3205-3224.
[8] R. Sharnagat, "Named Entity Recognition Literature Survey," *Cent. Indian Lang. Technol.*, pp. 1-47, 2014.
[9] J. Guo, G. Xu, X. Cheng, and H. Li, "Named entity recognition in query," *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 267-274.

[10] Y. Benajiba, P. Rosso, and M. Diab, "Arabic Named Entity Recognition: A Feature-Driven Study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 926-934, 2009.
[11] A. C. O. Reddy and K. Madhavi, "Convolutional recurrent neural network with template based representation for complex question answering," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 2710-2718, 2019.
[12] H. Sujaini, "Improving the role of language model in statistical machine translation (Indonesian-Javanese *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 2, pp. 2102-2109, 2020.
[13] Yasser Al-Onaizan, H. Hassan, and J. Sorensen, "An integrated approach for Arabic-English named entity translation," *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 2005, pp. 87-93.
[14] S. Nurmaini and B. Tutuko, "Intelligent robotics navigation system: Problems, methods, and algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 6, pp. 3711-3726, 2017.
[15] S. Y. Kim, S. H. Kim, and H. G. Cho, "Developing a system for searching a shop name on a mobile device using voice recognition and GPS information," *ICUIMC '12: Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, 2012, pp. 1-8.
[16] K. A. Kwaik, M. Saad, S. Chatzikyriakidis, and S. Dobnik, "Shami: A corpus of levantine Arabic dialects," *European Language Resources Association (ELRA)*, pp. 3645-3652, 2019.
[17] A. I. Al Huneety, "The Phonology and Morphology of Wadi Mousa Arabic," Phd thesis, Sch. Humanit. Lang. Soc. Sci. Univ. Salford, UK, 2015.
[18] A. Ekbal and S. Saha, "Classifier ensemble using multi-objective optimization for named entity recognition," *Frontiers in Artificial Intelligence and Applications*, vol. 215, pp. 783-788, 2010.
[19] A. Ekbal and S. Saha, "Classifier Ensemble Selection Using Genetic Algorithm for Named Entity Recognition," *Research on Language and Computation*, vol. 8, no. 1, pp. 73-99, 2010.
[20] A. Ekbal and S. Saha, "A multi-objective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14760-14772, 2011.
[21] A. Ekbal, S. Saha, U. K. Sikdar, and M. Hasanuzzaman, "A genetic approach for biomedical named entity recognition," *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, Arras, 2010, pp. 354-355.
[22] A. Ekbal and S. Saha, "Multi-objective optimization for classifier ensemble and feature selection: An application to named entity recognition," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 15, no. 2, pp. 143-166, 2012.
[23] M. Hasanuzzaman, S. Saha, and A. Ekbal, "Feature subset selection using genetic algorithm for named entity recognition," *PACLIC 24 Proceedings*, 2010, pp. 153-162.
[24] M. M. Abdulghani and S. Tiun, "Feature selection in web NER using genetic algorithm approach," *Journal of Theoretical and Applied Information Technology*, vol. 93, no. 2, pp. 552-560, 2016.
[25] H. T. Le and L. Van Tran, "Automatic feature selection for named entity recognition using genetic algorithm," *SoICT '13: Proceedings of the Fourth Symposium on Information and Communication Technology*, 2013, pp. 81-87.
[26] H. L. E. Thanh, L. Van Tran, T. H. Nguyen, and X. H. Nguyen, "Optimizing genetic algorithm in feature selection for named entity recognition," *ACM International Conference Proceeding Series*, 2015, pp. 11-16.
[27] L. Li et al., "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset," *Genomics*, vol. 85, no. 1, pp. 16-23, 2005.
[28] B. Desmet and V. Hoste, "Dutch named entity recognition using classifier ensembles," *Proceedings of the 20th Meeting of Computational Linguistics in the Netherlands*, 2010, pp. 29-40.
[29] A. Ekbal, S. Saha, and U. K. Sikdar, "Multi-objective Optimization for Biomedical Named Entity Recognition and Classification," *Procedia Technology*, vol. 6, pp. 206-213, 2012.
[30] U. K. Sikdar, A. Ekbal, and S. Saha, "Entity extraction in biochemical text using multi-objective optimization," *Computacion y Sistemas*, vol. 18, no. 3, pp. 591-602, 2014.
[31] C. N. Fredrick Edward Kitoogo, dos Santos and R. L. Milidiú, "Named entity recognition," *SpringerBriefs in Computer Science*, no. 9781447129776, pp. 51-58, 2012.
[32] L. Du, Y. Xu, and H. Zhu, "Feature Selection for Multi-Class Imbalanced Data Sets Based on Genetic Algorithm," *Annals of Data Science*, vol. 2, no. 3, pp. 293-300, 2015.
[33] C. H. Ooi and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," *Bioinformatics*, vol. 19, no. 1, pp. 37-44, 2003.
[34] A. Mirane Shahine, Gelbkuh, "Hybrid Feature Selection Approach for Arabic Named Entity Recognition," *International Conference on Intelligent Text Processing and Computational Linguistics*, vol. 9623, 2018, pp. 452-464.
[35] A. Akkasi, "Sentence-based undersampling for named entity recognition using genetic algorithm," *Iran Journal of Computer Science*, vol. 1, no. 3, pp. 165-174, 2018.
[36] K. Wu and K.-H. Yap, "A Perceptual Subjectivity Notion in Interactive Content-Based Image Retrieval Systems," *Intelligent Multimedia Processing with Soft Computing*, vol. 168, pp. 55-73, 2006.
[37] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "On the effectiveness of discretization on gene selection of microarray data," *The 2010 International Joint Conference on Neural Networks (IJCNN)*, Barcelona, 2010, pp. 1-8.
[38] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," *Knowledge-Based Systems*, vol. 86, pp. 33-45, 2015.
[39] Khaled Shaalan, "A Survey of Arabic Named Entity Recognition and Classificatio," *International Affairs, History & Political Science*, vol. 40, no. 2, pp. 469-510, 2013.

[40] A. De Sitter, T. Calders, and W. Daelemans, "A Formal Framework for Evaluation of Information Extraction," *Faculteit Communicatie en Cultuur*, vol. 2004, no. 12, 2004.
[41] M. Althobaiti, U. Kruschwitz, and M. Poesio, "AraNLP: A Java-based library for the processing of Arabic text," In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014, pp. 4134-4138.
[42] M. Khosrow, "Encyclopedia of Information Science and Technology," *IGI Global; 1st Edition*, 2017.
[43] A. Alajmi and E. mostafa Saad, "Toward an ARABIC Stop-Words List Generation," *International Journal of Computer Applications*, vol. 46, no. 8, pp. 8-13, 2012.
[44] A. Pasha et al., "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 1094-1101.
[45] K. Darwish, "Named entity recognition using cross-lingual resources: Arabic as an example," *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2013, pp. 1558-1567.
[46] K. Darwish and W. Gao, "Simple effective microblog named entity recognition: Arabic as an example," *European Language Resources Association (ELRA)*, pp. 2513-2517, 2014.
[47] A. Zirikly and M. Diab, "Named Entity Recognition for Arabic Social Media," *Proceedings of NAACL-HLT 2015*, 2015, pp. 176-185.
[48] O. H. Zayed and S. R. El-Beltagy, "Named entity recognition of persons' names in Arabic tweets," *Proceedings of Recent Advances in Natural Language Processing*, 2015, pp. 731-738.
[49] C. Helwe and S. Elbassuoni, "Arabic named entity recognition via deep co-learning," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 197-215, 2019.
[50] P. R. Yassine Benajiba, "Arabic named entity recognition using conditional random fields," *2008 5th Int. Conf. Inf. Commun. Technol. ICoIC7 2008*, 2008.
[51] J. D. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, "Distributed Representations of Words and Phrases and their Compositionality," *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, vol. 2, 2013, pp. 3111-3119.
[52] P. F. Brown, V. J. Della Pietra, P. de Souza, J.C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 1950, pp. 467-479, 1992.
[53] Y. B. Joseph Turian and Lev Ratinov, "Word representations: A simple and general method for semi-supervised learning Joseph," *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 384-394.
[54] F. Huang, A. Ahuja, D. Downey, Y. Yang, Y. Guo, and A. Yates, "Learning Representations for Weakly Supervised Natural Language Processing Tasks," *Diss. Abstr. Int. B Sci. Eng.*, vol. 70, no. 8, p. 4943, 2010.
[55] K. M. Ronan Collobert, Jason Weston, Léon Bottou and P. Koray Kavukcuoglu, "Natural Language Processing (Almost) from Scratch Ronan," *Journal of Machine Learning Research*, vol. 12, pp. 328-338, 2011.
[56] M. Bansal, K. Gimpel, and K. Livescu, "Tailoring continuous word representations for dependency parsing," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, 2014, pp. 809-815.
[57] R. Soricut and F. Och, "Unsupervised morphology induction usingword embeddings," *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 1627-1637.