The International Journal of Engineering and Information Technology

journal homepage:www.ijeit.misuratau.edu.ly

Using SVM Algorithm to Improve the Extraction of Arabic Composite Names: A Case Study in the Economic Domain

Raweha.D
Computer Science Department
Information Technology College
Misurata, Libya
d.raweha@it.misuratau.edu.ly

Khalil.H
Head of Computer Science Department
Information Technology College
Misurata, Libya
hussein.khalil@misuratau.edu.ly

Ben Sasi.A
College of Industrial
Technology Misurata, Libya
prof_ahmed@cit.edu.ly

Abstract— **The performance of natural languages has been developed by the unique processing applications (Named Entity Recognition) this is why the task of recognition has been of special significance. Therefore, this paper focuses on information extraction from Arabic unstructured text. Named entity recognition is a critical subtask of information extraction; it is the process by which a system can automatically detect and categorize Named Entities (NE). The proposed work in this research is looking for how to extract composite Arabic names in the economic domain by using machine learning approach specifically an SVM algorithm. To implement and test the proposed work the GATE tool was used. Also, for evaluation purposes and measure the efficiency and accuracy of the experiments, several measures were used including Precision, Recall and F- measure rates. Finally, a comparison was made between the results of the proposed system with the results obtained from previous research that depends on the extraction of composite Arabic names using rule-based approach. The application of SVM to extract Arabic composite named was successful. Results show that the proposed model based on machine learning in terms of Precision for extracting composite named is higher than the results obtained from the rule-based method. On the contrary, the obtained results of Recall of rule-based approach were higher than the results obtained from the machine learning SVM model. The general average results were: SVM approach has achieved a 97.1% Precision, 94.25% Recall, and F-measure 95.5%, while the rule-based approach has achieved a 93.4% Precision, 95.5% Recall and 94.3% F-measure.**

Index Terms— **Arabic named entity recognition, natural language processing, Arabic language, SVM approach, rule-based approach, composite Arabic names.**

I. INTRODUCTION

Arabic is one of the richest natural languages. Its morphological inflection and derivation processing has been of great challenge and interest. Interest in Arabic NLP has been gaining momentum in the past decade, and a colossal amount of information is published daily on the Web in the form of articles, documents, reviews, blogs and social media posts. Data is available in the form of unstructured document. Therefore, it makes extracting useful information difficult and takes time. [1]. Unstructured data refers to information that either does not have a predefined data model and/or does not fit well into relational tables; examples include email messages, word-processing documents, web pages, and many other kinds of business documents. Thus, the process of extracting information from an unstructured Arabic text is a particularly difficult task because Arabic is a highly inflective and derived language. The problem is defined by lack of advanced tools and research in Arabic Natural Language Processing (NLP) compared to European languages [2].

Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) that gives machines the capacity to read, understand and derive meaning from human languages. It helps machine to interact with humans in their natural language. The aim is to develop a computer that can "understand" the information included in the documents. NLP consists of many automatic language processing tasks [3]. A fundamental requirement for employing natural language processing (NLP) for information retrieval is called Named Entity Recognition (NER). Which is fundamental requirement to extracting main terms that related to one domain or another for unstructured text [4]. The Named Entity Recognition (NER) task aims to identify and classify named entities mentioned in unstructured text into a set of predefined categories of such as persons, organizations, locations, etc. NER is preprocessing module in several natural language processing (NLP) applications such as syntactic parsing, question answering, machine translation and data mining. Achieving the best performance on NER task requires large amounts of

Received 02 May , 2024; revised 11 May, 2024; accepted 15 Mar 2024. Available online 08 Aug, 2024.

external resources such as gazetteers, plenty of hand-crafted feature engineering and extensive data pre-processing [5].

There are different approaches used to extract named entity recognition: rule-based, machine learning and hybrid approaches. The extension of NER to specialized domains increased the importance of devising solutions that require less human intervention in the annotation of examples or the development of specific rules. Much more adaption is increased for machine learning techniques, therefore, research is being done to make these adaption feasible[6].

This work explores the scalability problems associated with solving the Named Entity Recognition (NER) problem. This research presents machine learning approach to extract Arabic composite names from Arabic text, namely, Support Vector Machines (SVM). The NER domain chosen as the focus of this work is the economic domain, especially selected due to its importance and inherent challenges.

II. RELATED WORK

Many studies have been done during last few years regarding to Arabic Named Entity Recognition for Arabic and other languages. As mentioned before, there are Three main approaches used to fulfill extraction of NEs in the form of different predefined types: the rule-based approach, the ML-based approach and the hybrid approach.

We will present some related literature review of Named Entity Recognition and Arabic Language. for all three major approaches including rule-based, Statistical Machine Learning based and Hybrid approaches.

A. Rule Based Approach

Elsherif, et al. in 2019 [7] presented a rule-based approach using the (GATE) General Architecture for Text Engineering is adopted by Arabic named entity recognition system. The system was applied on ANERcorp, and the results in terms of F-measure achieved 83% for Person NE, 89% for Organization NE, and 92% for Location NE. Although there are many approaches have been implemented in this rule-based NER track, a considerable amount of time and effort should be spent in order to keep such systems perform well with high recall by continuously adding more rules, lexical resources, grammars etc.

Hussein Khalil, et al. in 2020 [8] have presented a new rule-based approach that uses linguistic grammar-based techniques to extract Arabic composite names from Arabic text. This approach was based on the genitive grammar rules of Arabic language to classify pronouns into definite noun (معرفة) and indefinite noun (نكرة). This research has devised a set of genitive pattern recognition rules to retrieve composite names from unstructured text. The method has shown high recall and accuracy results.

Salah, et al. in 2022 [9] have developed a rule-based NER that can be deal classical Arabic documents. The main step of his approach relied on triggers words, patterns, gazetteers, rules, and blacklists generated by the linguistic information about entities named in Arabic. The method operates in three stages, operational stage, preprocessing stage, and processing the rule application stage. The obtained results were evaluated, and achieved a 90.2% rate of precision, an 89.3% level of recall, and an F-measure of 89.5%.

B. Machine Learning Approach

Ali, et al. in 2018 [10] has used a recurrent neural network (RNN) to solve natural language processing (NLP) tasks. They have proposed a bidirectional (LSTM) long short-term memory that makes the task of entity recognition of Arabic text. The LSTM network can process sequences and relate to each part of it, which makes it useful for the NER task. Moreover, they used pre-trained word embedding to train the inputs that are fed into the LSTM network. The method had been applied on Arabic corpus on a popular dataset called "ANERcor". Experimental results show that the model with word embedding achieves a high F-score measure of approximately 88.01%.

Li, et al. in 2019 [11] proposed a model combining language model conditional random field algorithm (CRF) and bi-directional long short-term memory networks (BiLSTM) the ability to automatically recognize and extract entities is ono-structured medical text. Later quantization was done for the normalized field of drug specification word by a vector as the input to the neural network. Experimentations indicated that Recall, system Precision and F1-value were improved by 5.2%, 6.18% and 4.87%, respectively compared to traditional machine learning models.

Ravikumar and Ramakanth in 2021 [12] have applied machine learning based approach to deal with medical clinical notes and how to extract essential concepts from them. Nowadays, medical publications databases have generated a lot of interest among researchers to make these techniques applicable to medical literature.

In Al-Qurishi, et al. 2021 [13], the authors have proposed a simple but effective model for Arabic named entity recognition The structure of this model is of three layers, a transformer-based language model layer, a fully connected layer, and the last layer is a conditional random field (CRF). Also, the process of summing and concatenating vectors has been shown to be effective in generating additional information that helps improving the tagging accuracy. Since vector representations were made on the words level, the model was able to collect contextual clues related to both syntax and morphology.

C. Hybrid Approach

Meselhi, et al. in 2014 [14] presented a new hybrid approach to NER in Arabic. This system deals with extracting persons, locations, organizations NEs from the ANERcorp corpus extracted from newswires and other web sources. The integration of a rule-based approach with an ML approach was implemented with the selection and correction of tags to identify any false negatives. The extraction of person entities has achieved 96.65% F-measure, while the other entities: locations and organizations has reached 94.8% and 92.9% F-measure, respectively.

Gridach, et al. in 2018 [15] have introduced a simple and fast model for Arabic named entity recognition based on Deep Neural Networks (DNNs). have present an Arabic NER system based on DNNs that automatically

learns features from data. The initial experimental results showed that this approach outperforms the model based on Conditional Random Fields by 12.36 points in F-measure. It is worth mentioning that this model outperformed the state-of-the-art by 5.18 points in Precision and very close results in F-measure were obtained.

Ramachandran, et al. in 2021 [16] have built a new dictionary for administration, dosage forms and symptoms to annotate the entities in the medical documents. A new dictionary has been built for route of administration, dosage forms and symptoms to annotate the entities in the medical documents. The annotation entities were trained by the SpaCy machine learning model. The proposed hybrid-based approach has achieved an F1-score of 73.79%.

Different aspects related to how Arabic language and its dialects are processed using different approaches have been targeted extensively as rule base approach, machine learning approach and hybrid approach. The previous studies have adopted the machine learning approach focused on how to extract person name, location, date, time, price, measurement, phone number, … etc. Arabic NEs pipeline has achieved an overall improvement for extracting the Arabic named entity, with the exception problem for the composite name it is need more. Similarly, in work by [17], the authors investigated for extracting the Arabic composite name entity using rule-based approach.

III. METHODOLOGY

The baseline experiment aims to identify composite named entities using a supervised learning approach. The training and testing data used are related to the economic domain. where the names of company, stock, sector, market and index are previously labeled composite named. The proposed algorithm employed in this experiment is the supervised machine learning using Support Vector Machines (SVM), due to their ability to handle high-dimensional features and input space, where the economic and financial field was a case study. It is normal to find numbers and tags related to the economic domain. These unnecessary elements as %, -, + and numbers are needed to be removed. The most challenging feature in machine learning approaches of NLP problems is deciding on the optimal feature sets. Fig. 1 illustrates the proposed algorithm block diagram using SVM learning technique to extract Arabic composite names.

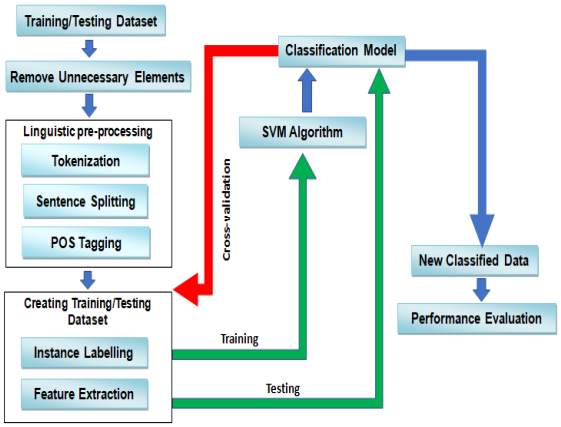

Figure 1. Proposed Research Algorithm

A. Dataset

Several types of unstructured data domain exist throughout the web such as health, historical and financial information etc, and the majority of data in the real world is unstructured data. Extracting non-trivial previously unknown and potentially useful information from published documents is difficult and time-consuming. The dataset used in this research is a set of Arabic documents and related to the economic domain on the web. It contains a lot of named entities, especially Arabic composite names the same corpora that was used in [8].

Removing unnecessary elements Is one of the primary tasks in natural language processing NLP. It is normal for the text data frequently contains several special formats like date formats, number, punctuation, diacritics. In addition, normalizing some writing forms that use different styles such as the optional inclusion of the letter "ء" (Hamza) as in "أحمد" (Ahmed) and "الإمارات" (Emirates). Also, removing non-Arabic words for example "JVC", "ICN", and so on. This process is very important to any successful IE system to make data ready for applying IE techniques tasks.

B. Linguistic Pre-Processing

- **Tokenization**
  Tokenization is the first step in any NLP pipeline. It has an important effect on any pipeline unstructured data and natural language texts that are needed to be broken down into chunks and discrete elements by tokenizer. These tokens enhance the ability to understand the context and develop a model for NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words, which is performed using the GATE tool.

- **Sentence Splitting**
  The sentence splitter is a main process splitting texts into meaningful chunks.
- **Part of Speech (POS) Tagging**

Part-of-speech (POS) tagging is one of the most important addressed areas in the natural language processing (NLP). The core task of the part of speech tagger (POS) is to categorize each word in the text to a word class (Token). such as noun, verb, and adjective, etc [18]. The Arabic part of speech (POS) tagger is not

supported by the GATE tool. Therefore, the POS which is assigned by the GATE tool that was integrated with Stanford tagger based on the maximum-entropy model [19]. Stanford University developed this model which is now supporting many languages, and Arabic is one of them [20]. The abbreviations are shown in Table 1.

TABLE 1: LIST OF SYMBOLS IN STANFORD TAGGER

Symbol	Description	Symbol	Description
DT	Articles including "a", "an",	DTJJ	an adjective with a definite article attached
IN	preposition	DTJJS	a plural adjective with a definite article attached
JJ	Adjective	NN	noun - singular or mass
DTNN	with a definite article attached	NNP	proper noun
DTNNS	a plural noun with a definite article attached	NNPS	proper noun – plural
NNS	noun – plural	CC	conjunction: "و"(and)
NP	proper noun – singular	NPS	proper noun – plural
		DE	Definiteness

C. Creating Training Dataset

- **Instance Labeling**

In the context of NER, data annotation (or Data labeling) the process of data labeling can be explained briefly as the process of adding target attributes to training data and labeling them. Therefore, a machine learning model is able to expect what prediction to make.

In this research the named target entities (Composite Names) were annotated using two methods: JAPE annotation and Manual annotation. In this paper, it was dependent on manual annotation method. Although, manual data annotation requires a lot of time and effort, However, the authors: Cunningham, et al. argued that they outperformed the search manual annotation method, and more accurate than automatic annotation [21].

- **Feature Extraction**

Feature extraction in natural language processing is defined as the process in which raw data transforming into numerical information in the original data set, by of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. This is for the purpose of decreasing the size of the effective vocabulary. It yields better results than applying machine learning directly to the raw data [22]. In the ML approach, the selection of the features to be taken into account by a classifier is a very critical issue and can significantly affect the performance of a system. The training is preprocessed using the GATE software through the application of ANNEI Pipeline in order to extract morphological and contextual features. The following features related to certain aspects of the Arabic language used in this research:

1) List String (Ls): The list string itself, it refers to the distribution of each composite entity type in the Corpus.

2) The distance of the composite name feature (Dis): denotes the number of words in a composite name. It is important to have a look at the distance of the composite name, the length of (composite name) is essential because short words that are less than two word that very short entity are not composite entities.

3) composite length (N-Ch): It is one of the word-level features described as the characters count which consists of the composite names.

4) Context word feature: The context in which a certain named entity appears is quite useful. There are several measures for context feature; including analysis of number of neighbor words (before/after) up to n numbers. Preceding and following words of a particular word can be used as features. This is based on the observation that the surrounding words are very effective in the identification of NEs. In this research, the value of n ranges from 1 to 5.

5) Trigger words (feature indicator) feature (Indt): is one of the most important features makes NE identification and takes various forms of verb list or noun list. This feature determines whether the words following the indicator are a composite name or not. In order to improve the framework performance and limit the ambiguity related to the NE boundary and to narrow down the search space. such as the word "company" (شركة) that aids to identify the name phrase “Alhelal Oil Company” (شركة نفط الهلال) word locating named entities in an unstructured text relevant to the targeted economic domain. Many Arabic words have been exploited to identify the named entities in natural language [23].

6) Part-Of-Speech tags feature (POS): The POS of the current word and the surrounding words is an important feature for NER, and often used with ML. This feature identifies the word part of speech class (e.g., verbs, nouns, pronouns, etc.).

7) NGRAM: N-grams are continuous sequences of the neighboring sequences of items in composite name.

8) kind (kind): It determines the type of annotation (composite).

D. SVM Algorithm

SVM is a known technique in machine learning and a powerful machine learning tool. One learning method is SVM. This method involves other learning techniques that analyze data for classification and regression based on firm statistical and mathematical foundations concerning generalization and optimization theory. Support vector machines have been successfully applied to many real-world problems such as handwriting recognition, information extraction, and others [24]. The SVM mathematical foundation for the binary classification case is a system which is trained to classify input data into one of two categories.

Training and testing data for a classification task typically comprise some of data instances. Every instance in the training set has multiple attributes in addition to one target value. Creating a model that expect the target value of data examples in the testing set that are merely given the attributes is the aim of SVM [25].

In SVMs, a dataset consisting of pairs of input vectors and desired outputs is called the training dataset. This information points to a desired response, validating the accuracy of the system, or helping the system to learn to act correctly. During the classification stage, the vectors are then employed to forecast the class of every upcoming data point. The classification stage provides an estimate for unidentified samples. The training phase's prediction model can be used to classify new data using various critical features.

E. Classification Model

Text classification is one of the tasks that are targeted by PR Batch Learning. For example, the recognition of named entity and learning of relation. It integrates LibSVM for improved speed, along with the PAUM algorithm. For linear classifiers, the perceptron algorithm with margins is a straightforward, quick, and efficient learning algorithm that generates decision hyperplanes within a fixed ratio of the maximal margin. It also offers a Weka interface. The batch learning presents data export functionality, this leads for more flexibility. All the PR configuration parameters are set through one file. Which is external XML file. The Batch Learning PR provides a variety of optional settings, which facilitate different tasks. Every optional setting has a default value; if an optional setting is not specified in the configuration file, the Batch Learning PR will adopt its default value.

F. Finding a good Classification Model

The evaluation Composite name extraction algorithm is documented by this section. The simple SVM structure is applied to evaluate and rank the features performance in order to recognize the named entity, "composite name". The datasets generated from economic domain corpus are used for evaluation of systems. The corpus was split into 70% for the training and the 15% for validation set and the remaining set (15%) is for testing, where training set represents the input values for the classification model of SVM.

The first experience the system based on SVM implementation with different feature combinations to keep only those giving best results for the composite entities detection. In addition, the impact of composite names' length on precision. However, the main part of learning includes the optimal set of features to support the performance of classifier. The k-fold cross-validation experiments were run on all SVM training corpora. This comes after the determining the best feature set to extract the composite name. After completing the experiments, the model will be evaluated in terms of the best results on the training set, in terms of Precision, Recall and F-measure for each case separately. The next step is applying the model on an untrained dataset.

F-NUM	Features	Precision	Recall	F-M
Ft-1	All Feature , except ngram	0.8236241	0.59253	0.68922
Ft-2	Ls, POS_Ls	0.7129834	0.49354	0.58324
Ft-3	Ls	0.4435566	0.48111	0.46151
Ft-4	Ls, Indt, Dis, N-Ch	0.4942594	0.59959	0.54178
Ft-5	Ls, Dis, N-Ch	0.4787440	0.53954	0.50732
Ft-6	Ls, Indt	0.4931464	0.50727	0.50050
Ft-7	Ls , Indt, Str_12345W_B, Str_12345W_A	0.6030843	0.55049	0.57550
Ft-8	Ls , Indt, POS_Ls, Str_12345W_B, Str_12345W_A	0.6921418	0.57088	0.62568
Ft-9	Ls , Str_12345W_B, Str_12345W_A	0.7188345	0.57746	0.64043
Ft-10	Ls , Indt, POS_Ls, N-Ch, Str_12345W_B, Str_12345W_A	0.7054729	0.53112	0.60600
Ft-11	Ls , Indt, POS_Ls, N-Ch, Str_12345W_B, Str_12345W_A, POS_12345W_B	0.6898337	0.51310	0.58848
Ft-12	Ls , Indt, POS_Ls, N-Ch, Str_12345W_B, Str_12345W_A, POS_12345W_ A	0.7542566	0.58623	0.65971
Ft-13	All Feature, except , POS_45W_ B N-gram	0.8420216	0.60231	0.70227
Ft-14	All Feature	0.9160145	0.77806	0.84142

TABLE 2: MEASURES FOR DIFFERENT FEATURE COMBINATIONS

Several experiments were conducted to determine the effective feature set for composite names extraction task. Table 2 shows the results of testing the classification model on different combinations tested with the corresponding Precision, Recall, and F-measure metrics used to evaluate the results.

G. Cross-validation

Cross-validation is the standard way of evaluating Machine Learning systems. In order to evaluate the machine learning system performance while avoiding the over fitting, the k-fold cross-validation is usually used. The dataset is randomly divided into k-folds of equal size. Each fold is employed as a testing set, and the remaining folds are then applied as a training set, and the test results are averaged over the k rounds. The same split must be replicated for training and testing, to make sure that these two steps are used in all the documents, and also used different values for k in order to obtain the best performance for the k-fold cross-validation with the k volume [26].

The standard evaluation measures in the IE community [2] (i.e., Precision, Recall and F-measures), to evaluate and compare the results. These measures depend on contrasting the entities that are extracted by the manually annotated corpus. Precision indicates how many of the extracted entities are correct. Recall indicates how many of the entities that should have been found are effectively extracted, and F-measure is a harmonic mean that gives equal weight for recall and precision, as follows [27]:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True positives+False Positives}} \quad (1)$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives+False negatives}} \quad (2)$$

$$\text{F} - \text{measure} = \frac{2*\text{recall}*\text{precision}}{\text{recall+precision}} \quad (3)$$

H. Model Experiments

To finalize the experiments, the classification model was tested in ANER task. It is clear that the SVM model has given a good performance in extracting Arabic Composite names by using the standard evaluation measures that previously mentioned. Table 3 illustrates the overall obtained experimental results for the model classification experience on an unlabeled dataset.

Composite Name Length	TP	FP	FN	Precision	Recall	F-m
Two words	169	2	15	0.988	0.918	0.952
Three words	103	3	10	0.971	0.911	0.940
Four words	49	2	3	0.960	0.942	0.951
Five words/more	26	1	0	0.962	1	0.981

TABLE 3: PERFORMANCE MEASURES OF THE CLASSIFICATION MODEL

IV. PERFORMANCE COMPARISON OF THE SYSTEMS

The implemented classifier model will be compared against state-of-the-art systems to extract composite name, that used the rule-based approach. In addition, to make a reasonable comparison between these systems on the same dataset in this section.

At First, the implemented system will be symbolized by SVM-ECNO and for Rule-Based system by Rule-ECNO. Fig. 2 and Fig. 3 summarize the performance of both systems in terms of the effect of the length of the compound name. The SVM-ECNO has achieved a precision of 98%, while Rule-ECNO has achieved 95%, and a Recall of 91.8% and 96% for SVM-ECNO, Rule-ECNO, respectively in case of two words.

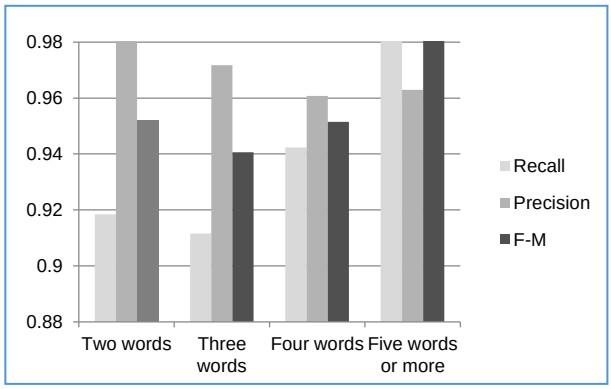

Figure 2. Performance of The SVM Model

Also, for three-word names, the difference was about 4% in terms of precision and recall. The priority was given to Rule-ECNO for Recall that scoring 95%. The precision was the highest that recorded approximately 97% for SVM-ECNO model. The performance of SVM-ECNO model was the highest in case of four words with a difference of 5% for precision, where SVM-ECNO has recorded 96%, and for the Recall was SVM-ECNO 94%. While Rule-ECNO was higher by about 2%. In the last case of five or more words, the performance of the system based on machine learning was the highest in Recall and Precision.

The SVM-ECNO system significantly outperformed the rule base system in terms of the precision. On the other hand, Rule-ECNO system slightly outperformed the SVM approach in terms of recall. This result shows that the SVM system achieved high precision with fewer false positive errors, but this leads to a poor recall. A trade-off is usually found between recall and precision. If a wider range of words as labeled by system. The result will be detection of entities (high Recall) as well as false error (lower precision). However, classifying everything in positive category means the classifier is not useful by getting one hundred percent recall and bad precision. Thus, the F-measure is a harmonic mean that gives equal weight to recall and precision. In terms of the F-measure: for the SVM-ECNO system, the names which contain two words demonstrated an F-measures of 95%. While three words recorded (94%), four words (95%), and five or more words recorded (98%). This was the highest value obtained from the system. While the performance of the rule-based system was about 96% for the names, which contain two words, 94% for three words, 93.6% for four words, and 93.5% for five or more words, as shown in Fig. 3. Therefore, in SVM-ECNO system as the number of words of composite names increase, the F-measure scores increase.

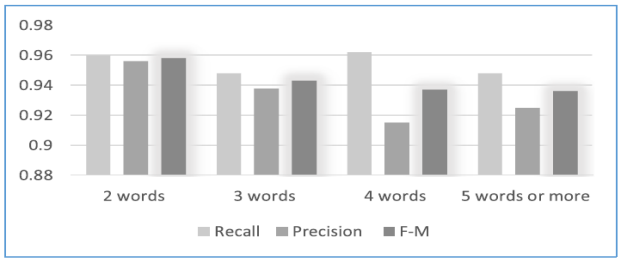

Figure 3. Performance of Rule-based approach [8]

Table 4 shows the summary of the comparison between the average values of the obtained Recall, precision and F-measure on SVM and Rule-based approaches.

TABLE 4: THE SUMMARY OF THE COM PARISON BETWEEN SVM AND RULE-BASED APPROACHES

	Recall	Precision	F-measure
SVM approach	94.25%	97.1%	95.5%
Rule-based approach	95.5%	93.4%	94.3%

V. CONCLUSIONS

In this paper, experiment-sets have been conducted with the aim of extracting composite named. The experiments have been carried using machine learning techniques. Specifically, SVM algorithm on a dataset in the economic domain, and testing the approved model on an unlabeled dataset. Finally, a comparison with the results of a system that used the rule-based approach to extract composite named. The complicated Arabic morphology and the lack of NLP needed tools makes the overall task of Arabic NER challenging for composite names. The conclusions of this research can be summarized in the following points:

1- The proposed SVM-based system to extract composite names has been successfully implemented.

2- The SVM-ECNO implemented system outperformed the Rule-ECNO with an overall average precision of 97.1% and average F-measure of 95.5%, while the Rule-ECNO had a higher average recall of 95.5%.

For further future research, it is suggested to investigate the utilization of advanced deep Learning techniques to extract Arabic composite named.

REFERENCES

[1] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, (2021). Arabic natural language processing: An overview, *Journal of King Saud University-Computer and Information Sciences*, 33(5), 497-507.

[2] B. A. Ali, B. Mihi, S. El Bazi, and I. N. Laachfoubi, (2020). A Recent survey of Arabic named entity recognition on social media, *Rev. d'Intelligence Artif., 34(2), 125-135.*

[3] A. Farghaly, and K. Shaalan, (2009). Arabic natural language processing: Challenges and solutions, *ACM Transactions on Asian Language Information Processing (TALIP), 8(4), 1-22.*

[4] K. Shaalan, (2014). A survey of Arabic named entity recognition and classification, *Computational Linguistics, vol. 40, 469-510.*

[5] H.Shelar, G. Kaur, N. Heda, and P. Agrawal, (2020). Named entity recognition approaches and their comparison for custom NER model, *Science and Technology Libraries, 39(3), 324-337.*

[6] N. F. Mohammed, and N. Omar, (2012). Arabic named entity recognition using artificial neural network, *Journal of Computer Science, 8(8), 1285.*

[7] H. Elsherif, M. Alomari, K. M. AlHamad, A. Q. M., and K. Shaalan, (2019). Arabic rule-based named entity recognition system using GATE. *In MLDM (1), 1-15.*

[8] H. Khalil, T. Osman, and M. Miltan, (2020). Extracting Arabic composite names using genitive principles of Arabic grammar. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(4), 1-16.*

[9] R. Salah, M. Mukred, L. Qadri binti Zakaria, R. Ahmed, and H. Sari, (2022). A new rule-based approach for classical Arabic in natural language processing. *Journal of Mathematics, 1-20.*

[10] M. N. Ali, G. Tan, and A. Hussain, (2018). Bidirectional recurrent neural network approach for Arabic named entity recognition. *Future Internet, 10(12), 123.*

[11] W. Li, et al., (2019). Drug Specification Named entity recognition base on BiLSTM-CRF model, *IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 429-433.*

[12] J. Ravikumar, and K. P. Ramakanth, (2021). Machine learning model for clinical named entity recognition. *International Journal of Electrical and Computer Engineering, 11(2), 1689.*

[13] M. S. Al-Qurishi, and R. Souissi, (2021). Arabic named entity recognition using transformer-based-CRF model. *In Proceedings of the Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021), 262-271.*

[14] M. A. Meselhi, H. M. A. Bakr, I. Ziedan, and K. Shaalan, (2014). A novel hybrid approach to Arabic named entity recognition. *In China Workshop on Machine Translation (pp. 93-103). Springer, Berlin, Heidelberg.*

[15] M. Gridach, (2018). Deep learning approach for Arabic named entity recognition. *In Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing, Konya, Turkey, Revised Selected Papers, Springer International Publishing, Part I 17, 439-451.*

[16] R. Ramachandran, and K. Arutchelvan, (2021). Named entity recognition on bio-medical literature documents using hybrid-based approach. *Journal of Ambient Intelligence and Humanized Computing, 1-10.*

[17] K. Shaalan, and H. Raza, (2008). Arabic named entity recognition from diverse text types. *In International Conference on Natural Language Processing (pp. 440-451). Springer, Berlin, Heidelberg.*

[18] M. Konkol, (2012). Named entity recognition: technical report no. *DCSE/TR-2012-04, Doctoral dissertation, University of West Bohemia in Pilsen, Czech Republic.*

[19] Á. Rodrigo, J. Pérez-Iglesias, A. Peñas, G. Garrido, and Araujo, L. (2013). Answering questions about European legislation. *Expert systems with applications, 40(15), 5811-5816.*

[20] S. Green, and C. D. Manning, (2010). Better Arabic parsing: Baselines, evaluations, and analysis. *In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 394-402.*

[21] H. Cunningham et al. (2009) Developing language processing components with Gate Version 8: A user guide. *University of Sheffield.*

[22] S. Alanazi, (2017). A named entity recognition system applied to Arabic text in the medical domain (Doctoral dissertation, *Staffordshire University).*

[23] A. Saif, M. J. Ab Aziz, and N. Omar (2013). Measuring the compositionality of Arabic multiword expressions. *In Soft Computing Applications and Intelligent Systems, ed: Springer, 245-256.*

[24] M. S. Habib, (2008). Improving scalability of support vector machines for biomedical named entity recognition. *University of Colorado Colorado Springs.*

[25] X. Song, H. Wang, and L. Wang, (2014). FPGA implementation of a support vector machine-based classification system and its potential application in smart grid. *In 2014 11th IEEE International Conference on Information Technology: New Generations, 397-402.*

[26] Rabiee, (2011). Adapting standard open-source resources to tagging a morphologically rich language: A case study with Arabic. *In Proceedings of the Second Student Research Workshop associated with RANLP, 127-132.*

[27] S. Alanazi, (2017). A named entity recognition system applied to Arabic text in the medical domain. *Doctoral dissertation, Staffordshire University.*