Journal of Theoretical and Applied Information Technology
30th September 2015. Vol.79. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PROPER NOUNS RECOGNITION IN ARABIC CRIME TEXT USING MACHINE LEARNING APPROACH

[1]SUHAD AL-SHOUKRY, [2]NAZLIA OMAR

[1a,2] Centre for Artificial Intelligent Technology (CAIT), Faculty of Information Science & Technology (FTSM), Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, MALAYSIA. [1b] Foundation Technical Education, Al-Furat Al-Awsat Technical University (ATU), www.forat.tech.un.edu.iq, IRAQ.

E-mail: [1]suhadalzhra2010@yahoo.com , [2] nazlia@ukm.edu.my

ABSTRACT

Named Entity Recognition (NER) identifies proper nouns in a text and categorizes it as a distinct kind of named entities. This function enables the extraction of peoples name, locations, organizations, and currencies. Several research abound in this area in Arabic NER is concerned. However, recognizing Arabic named entities is challenging due to the complexity in the Arabic language. These complexities are represented by non-existence of capitalization feature which facilitates the process of NER. Furthermore, there is a lack of lexical corpora that may include all the Arabic NEs. On other hand, most of the approaches that have been proposed for Arabic NER were based on handcrafted rule-based methods which can be laborious and time consuming. Therefore, this paper presents our attempt at recognizing and extracting the most important named entities, such as names of persons, locations, organizations, crime types, dates and times in Arabic crime documents using the Decision Tree classifier and feature extraction for crime dataset. The dataset consists of varying data sizes collected from online resources and undergone multiple pre-processing tasks. Additionally, the feature extraction task which includes POS tagging, keyword trigger, definite articles and affixes has been performed. Furthermore the classifier will utilize these features in order to classify the named entities. The results demonstrate that the use of the Decision Tree (DT) yields good results in small size datasets, but failed when the dataset is large in the case of the crime domain. This is due to the difficulty of relevant keywords and features in such fields. The best result for this experiment is 81.35% F-measure.

Keywords: *Natural language processing, Machine Learning, Decision Tree, Named entity recognition, Arabic crime document.*

1. INTRODUCTION

The increasing rate of crime in the Arab world especially in the present day and heightened level of crime cases available online has necessitated for this study. The increasing need to enhance the process of analyzing and determining relevant in-time information such as entities that form the crime cases has become more significant in the current time more than ever. In account of providing detailed information on the increasing online crime, Menara online newspaper published in 2013 stated that a total of 477,759 crime cases have been publicized. Additionally, the recognition and categorization of named entities in crime cases in this study stands out to provide significant information to effectively investigate crime as well as fundamental information required in analyzing different crime situations. These entities may also be valuable to other NLP applications in the field of criminology in processes including text summarization, extraction of relationship, information extraction (IE), and information retrieval (IR) as well as question answering (QA) which could contribute to crime analysis [1]. However, most current studies confined to investigating information extraction (IE) specifically on the analysis of unstructured text and information extraction needed to form structured representation and extended to categorizing information invaluable as entities in the form of individuals, organizations or locations. The inclusion of invaluable information in the text linked to the entities (for instance, the individual's title, or the organizational type; facts regarding the relationship between entities and events where the entities participate in).

Journal of Theoretical and Applied Information Technology
30th September 2015. Vol.79. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Notably, among the widely known sub-tasks of IE is the Named Entity Recognition (NER) process that assists users' to generate a more meaningful corpus by determining proper corpus names and categorizing them into types (i.e. persons, organization, and location) [2]. Moreover, the times and dates are considered as proper names types as illustrated by Traboulsi [3].

The setting of the NER algorithm in the NLP has primarily impacted by various domain studies'. Moreover, the NER application that is characterized as domain-specific may be inapplicable to determine named entities on other distinct domains like restaurant guides. In addition, different languages necessitates for different methods for entity recognition. For instance, the detection of named entity types for articles in the English language could be carried out by detection proper nouns – such nouns often begin with a capital letter and are utilized to represent a distinct entity like people, location and organization. Nevertheless, such a method is suitable for articles in the Arabic language as it does not have distinct symbols for entity detection [4]. The reason lies in the fact that a majority of languages can be distinguished from each other morphologically. In sum, the NER implementation hinges on the studies domain and the kind of languages used.

Moreover, the NER task in a distinct language is often realized via information collection concerning the terminology. For example, in the English terminology, this type of language may entail well-known titles, capitalizing proper names, suffixes or common prefixes, and noun phrase recognition in documents as well as Part of Speech (POS) tagging. In this regard, [5, 6] asserted that specific method developed for a distinct language may not be appropriate for another language.

Therefore, the NER issue has been examined by several studies in various languages and in different domains. Nevertheless, researchers focusing on the NER in terms of Arabic crime document are very limited. Additionally, when shifting to a new domain, the lexical sources have to be customized requiring modification of the system features which are domain-specific [7].

Therefore, it becomes necessary to examine Arabic language which is the mother tongue of about 323,825,430 people and official language of Arab countries [6]. Arabic has a distinct orthography and a complex morphology that brings to light novel challenges to the NER task requires a thorough study. Although several systems have been created to conduct NER tasks [8], the entire Arabic NER systems are geared towards general domain and only a few studies have been dedicated to an Arabic entity in crime documents. Such entities in a specific domain represent the terms or sentences that emphasize concepts invaluable to a special field. For instance; " وجد المواطن ا نور ابراهيم مقتولا بالقرب من جامعة صلاح الدين في تمام الساعة 10 صباحا يوم السبت 5 آب 2015 أثناء عبوره الشارع حيث انه يعمل بمؤسسة الرياض للطباعة والترجمة".

Hence, the translation for above statement as follows: "Anwer Ibraheem was found killed nearby Salah Aldeen University, at 10:00am on Saturday, August 5, 2015 while he was crossing the street, where he works at Al-Riyad Company for printing and translating".

Table. 1 represent the person name, location, organization, crime type, date and time extracted from the given example.

Table. 1 Extraction of the entities

Entities	Translation	Representation
انور ، ابراهيم	Anwer, Ibraheem	Person name
صلاح الدين	Salah Aldeen	Location
صباحا	Morning	Time
السبت	Saturday	Date
الرياض	Al-Riyad	Organization
قتل	Killed	Crime type

The subsequent part of this paper incorporates related works in the Arabic domain in specific crime domain as well as the proposed method and the experiment performed to evaluate the system and finally concludes and recommended appropriate strategies to improve future work.

2. RELATED WORK

This section deals with several prior studies highlighted on the Arabic language in the crime domain. The majority of NER Arabic authors have focused on the general domain and as a result, the crime texts have been largely ignored.

2.1 Related Work in the Arabic Crime Domain

In crime domain, only a few studies utilized NLP tools and among them, [9] focused on the issue of crime pattern detection by using clustering algorithm to assist in the detection of crime patterns and expedite the solving process. They made use of the k-means clustering method with several

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

enhancements to reinforce the identification process of crime pattern identification and technique application to actual crime data from the crime office. Based on the results, the top attributes in crime patterns found include the crime type and weapon name which are used in the crime.

In a related study, [10] referred to Information Extraction system on the web according to NLP in order to examine the available crime information. The authors used the Semantic Inferential Model that focuses on developing an NLP system with an added layer to clarify the text understanding that also shows an analysis of the text content (explicit and implicit). Moreover, he also used some tools like the Wiki Crimes for the extraction of crime types and crime scenes achieved from online crime news.

Also, [7] generated a rule-based model to classify the entity in Arabic crime text. The system was developed to determine four entity types namely the person name, organization, location, dates and time, they implemented an experiment and found that the performance in terms of accuracy constitutes up to 89% of the F-measure depicting the weakness of their system. However, is the process of entity selection and assigning of weight for terms that are done by hand and required human efforts.

Although several researchers have dedicated their work to the issues surrounding multiple languages of the NER, only a few of them focused on ANER document owing to the limited level of progress in the Arabic NLP and the absence of different resources regarding the Arabic NEs.

Specifically, [11] created an Arabic NER system consisting of a syntactic parser and morphological parser developed within the Nooj linguistic development field. The environment comprises of huge dictionaries, grammars, and parser corpora in actual time. It is utilized to categorize numeric, dated, known proper names and unknown proper names in the Arabic language. Moreover, an evaluation process was conducted on a portion of corpora gathered from a published newspaper, "Le Monde Diplomatique" in the Arabic language. The enhanced system produced the respective scores: precision (92%), recall (79%) and f-measure (85%).

In a related study, [12] provided a description of an NER system with the help of the Support Vector Machine (SVMs) and the mixture of language independent and dependent features of the Arabic NER. The impact of various features was measured independently and in combination throughout various standard data sets and different types.

Moreover, [13] developed an Arabic developed an Arabic NER system namely ARNE. The system was developed based on a gazetteer lookup approach which is used to recognize proper names such as a person, location, and organization. The experimental results gained low performance in terms of Recall, Precision, and F-measurement; 27%, 38%, and 30% respectively.

Meanwhile, [14] examined Arabic NER system developed on the combination of two automatic learning methods namely Conditional Random Fields (CRF) recognize as a supervised method, and bootstrapping semi-supervised method identification. Such a system was utilized to determine the location, organization, person name, and the job as well as other classes.

Also, [15] brought forward a hybrid Arabic known as entity recognition (NER) method that adopted the rule-based and machine learning-based methods, developed using J48 classifiers, wherein the total precision-recall, the J48 classifies was able to reach the following F-measures; 49%, 90% and 88% for person name, location and organization respectively.

3. PROPOSED METHOD

In this paper, the NER system entails modules for linguistic pre-processing, identification of named entity and classification. The detailed architecture of the study's system is depicted in Figure 1.

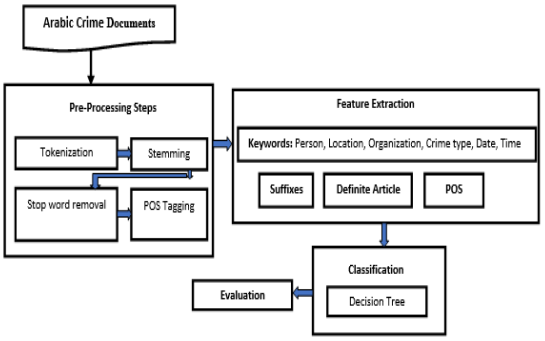


Figure1 The architecture of the proposed method

Journal of Theoretical and Applied Information Technology
30th September 2015. Vol.79. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1 Pre-processing Methods

Four pre-processing modules are included in the system and such modules require usage prior to the NER task. Their use hinges on the input nature when the input is raw text. The modules include sentence splitting, tokenization, removing stop word and stemming that covers POS tagging. These modules are further elaborated in this section.

Sentence splitting and filtering modules are processes where the removal of punctuation marks, commas, end of the line, stop words (non-useful words), duplicated letters, diacritics, prefix, and suffix are conducted. Some of the stop words in the Arabic language are shown in Table. 2.

Figure 2. Specific domain (crime dictionaries)

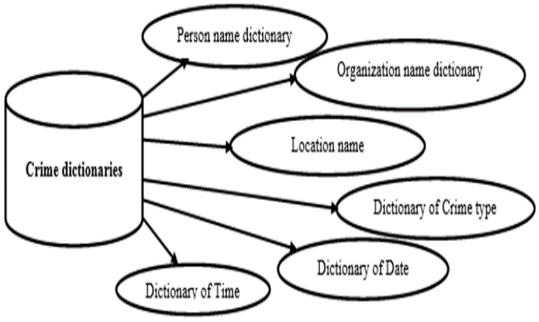


Table. 2 Arabic stop words

Token	Examples
Punctuations	. , ¡ : " ' ÷ × º > < | \ ¿ ' ! @ # $ % ^ & *) (_ -
Duplicate	سد صف سل هم ظل خص ضر جن خط قل
Definite Articles	كال وال تال فال بال وبال لبال فبال
Stop words	بعد ضد يلي الى في من حتى وهو يكون به وليس أحد على وكان
Prefix	ا ب ت س ك ل م ن ه و سن لي سي لن مت من نت وت يت
Suffix	ن ا ت ة ك ه ي وه ين اة تم ته نن ني هن ون يه ان تك تة كم نك نه
Diacritics	ً ٌ ٍ ْ ِ ُ ّ َ

3.2 Crime as NER

The crime domain as shown in Figure, 2 has many civic crimes types. In this paper, all kinds of crimes such as kidnapping, theft, smuggling and homicide are included in order to be recognized by Decision Tree approach concerning the crime domain. The knowledge base is considered as a set of dictionaries initiated from the crime domain, and hence each list is devoted for a specific purpose as illustrated in Table 3.

The dictionaries are derived by using data collection techniques from four online newspapers and the entire process links the words in the lists with relevance to the input text. The dictionaries can be described as a list of crime proper nouns - including the person, name, place name, and organization name as well as date and time. (e.g. مركز الشرطة، جوازات، دوريات ، بيشة ،المواطن، السيد، الجنرال،المصاد□، صباحا، عكاظ). These keywords are listed in Table. 3.

Therefore, the actual names of person come after these keywords in the Arabic language. Hence, these names can be extracted and identified. For example, the word that follows "the citizen/المواطن"is usually a proper name. Table. 3 indicates the listed types with keywords.

3.3 Decision Tree

This section reported a detailed representation of proper nouns in crimes more specifically on Recognition System architecture (RSA) using decision tree (DT). "Decision support tool" utilizes a treelike graph or model of decisions and their potential consequences includes chance event outcomes, resource costs and service program. Though it is a known methods used to display an algorithm, it is widely used in operational research specifically in decision analysis to identify strategy to reach a predetermined goal. In Practice,

Journal of Theoretical and Applied Information Technology
30th September 2015. Vol.79. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table. 2 List of crime proper nouns

Dictionary	Keywords	Translate
Person	المواطن، السيد، القائد، السيد الجنرال	Citizen, Mr., the commander, Gen.
Location	مدينة، منطقة، عاصمة، شارع	City, area, the capital, street
Organizatio	صحيفة، شركة، محكمة، مركز شرطة	Newspaper, company, court, police station
Crime type	السرقة، القتل، الاحتيال، الخطف، الربا، الاغتصاب	Theft, murder, Fraud, kidnap, Usury, Rape
Dates	السبت، يناير، تشرين الاول، رمضان	Saturday, January, October, Ramadan
Time	صبيحة، يوم، ظهرا، صباحا	Morning, day, at noon, in the morning

online decisions are accepted with no recall of incomplete knowledge and decision tree is duplicated by a probability model as an appropriate choice model or online selection model algorithm for computing conditional probabilities [16].

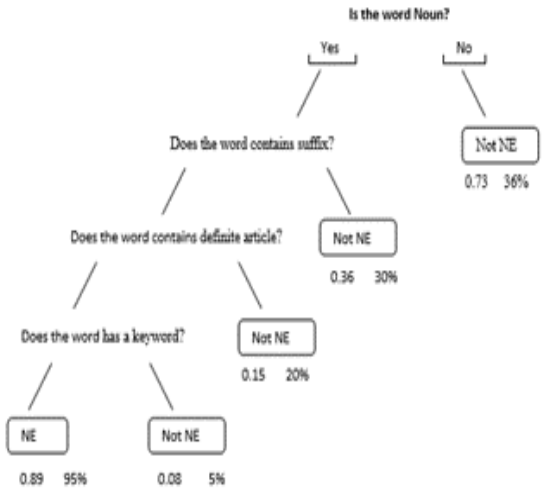
Decision Tree is popular classifiers because they do not require any knowledge base skill or parameter setting [17]. DT are used to predict the value of a dependent class and to handle high-dimensional information. The learning and classification processes of DT training are straightforward and immediate. The DT classifier is a straightforward classification method for humans [18].

However, DTs have many types depending on the situation and the desired outcome. The classification tree is considered the most important of these types. Classification trees are used when different pieces of information are calculated to determine the most predictable outcome. With DT, the user applies a binary process of categories and subcategories to layout the different variables that constitutes an outcome. Owing to the versatility of DT, its application has been used in probability and statistical measurements.

Basically, DT work by splitting the data into Training and Testing portions. Training data are a set $S = s_1, s_2,.., s_n$ of already classified samples.

Each sample $s_i = x_1, x_2,.., x_n$ is a vector where $x_1, x_2,...$ represent the attributes or features of the sample. Training data are augmented with a vector $C = c_1, c_2,.., c_n$ where $c_1, c_2, ...$ represent the class to which each sample belongs. At each node, DT selects one attribute of the data that most effectively splits its set of samples into subsets enriched in one category or the other.

The simplest illustration of this lower bound technique is for the problem of finding the smallest number among n numbers using only comparisons. In this paper, the decision tree model is a binary tree. Algorithms for this searching problem may result in n different outcomes (since any of the n given numbers may turn out to be the smallest one). It is recognized that the profundity of a binary tree with no leaves is at least \log n, which gives a lower bound of \Ω (\log n) for the searching problem. Even so this lower bound is known to be slack, since the following simple statement shows that at least n-1 comparisons are needed: Before the smallest amount can be determined, every figure except the smallest must "lose" (compare greater) in at least one comparison. Similarly, lower bound of \ Ω (n\log n) for sorting may be proved. In this case, the existence of numerous comparison-sorting algorithms having this time complexity, such as merge sort and heap sort, demonstrates tight bound. Illustration below predicted the probability for each node depicting features used in this experiment.



Then, Cx(c) averages over probabilities assigned by the different trees to a test example by the equation below:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$DTp = \sum fi\ log2fimi = 1$$

fi is the feature that has been used, so it each feature will be calculated using this equation in order to estimate the probability of such feature.

3.4 Evaluation and results

The standard evaluation measures in the information extraction, precision, recall, and F-measure are computed in this section to evaluate the performance of our approach. NER were the evaluated using the following metrics:

$$Precision = \frac{\#\ of\ correctly\ classified\ entities}{\#\ of\ entities\ found\ by\ algorithm} = \frac{tp}{tp + fp}$$

$$Recall = \frac{\#\ of\ correcily\ classified\ entities}{\#\ of\ entities\ in\ the\ corpus} = \frac{tp}{tp + fn}$$

$$F_\beta - \frac{(1 + \beta^2) * (Precision * Recall)}{(\beta^2 * Precision + recall)}$$

The precision refers to the percentage of NEs found by the system that are correct. Recall refers to the percentage of NEs existing in the corpus that were found by the system, and the F-measure refers to the average of the recall and precision.

3.5 Experimental dataset

In this study, a dataset was collected from Arabic online newspapers such as Al-Bayan, Al-Mutawaset, Al-Jazeera and Okad applied by Asharef et al. [7] has been adopted for this paper. These newspapers were used to show that the system is capable of handling writing styles and Arabic names.

4. RESULTS

After testing the Arabic crime data set, the results obtained are listed in Table. 4.

this experiment presented in this study was successfully applied on datasets with different sizes. The first dataset contains 30 news of crime and the results represents the Precision, Recall, and F-measure as shown in Table. 4.

Table. 4 Results of the system with limited dataset

Class	Precision	Recall	F-measure
PERSON	84.15%	76.87%	80.34%
LOCATION	85.38%	77.56%	81.28%
ORGANIZATION	84.93%	79.66%	82.21%
CRIME TYPE	84.98%	80.38%	82.61%
DATE	82.67%	79.04%	80.81%
TIME	81.45%	80.01%	80.72%
OVERALL	83.43%	79.44%	81.35%

Table. 4 presents the accuracy of the F-measure, recall, and precision for each class of NE (person name, location, organization, crime type, date, and time) using the tested system. The results shows that the measurements by applying the Decision Tree approach of the classes is almost all within the same range and the overall results for 30 crimes reveal an F-measure of 81.35%, recall of 79.44%, and precision of 83.43%. That is somewhat considered a good performance, but the approach suffers when multiple features are applied.

Table. 5 shows the overall measurements with different datasets with data sizes 50 news, and 70 news, and the findings present the values of precision, recall, and f-measure for all classes.

Experimental results shows DT applied with varying datasets to check the performance with larger amount of data. From Table. 5, it can be seen that increasing the data set produced lesser values for the measurements used compared with 30 crimes. We therefore propose based on our results that DT is a good classifier for small data set. As the dataset increases, there is a corresponding increase in features (scalability), which decreases the performance of DT for proper nouns recognition in the Arabic crime documents [19]. The results shows that the F-measure achieves better result when the dataset is small (81.35% with dataset of size 30).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table. 5 The overall measurements using different sizes of dataset in crime field

Dataset size	Words	Recall %	Precision %	F-measure %
30 crimes	1200	83.43%	79.44%	81.35%
50 crimes	2100	75.30%	68.80%	71.90%
70 crimes	2940	71.80%	66%	68.77%

This is due to the fact that Arabic language has more features when the dataset size increases. When the size increases, the features also increase. Thus, this may affect the classifier results when larger dataset is used.

4. CONCLUSION

The recognition of proper nouns in Arabic NER provides important information for crime analysis and could be used to improve other NLP applications. The NER approach standout to enhance crime analysis and contributes towards the implementation of the NER model that extracts and recognizes proper nouns from Arabic crime documents. This paper supported that proper nouns can be extracted from a given Arabic texts (whether crime texts or others) within a confined dictionary set. Future studies can be planned to improve the extraction of proper nouns from crime domains and to enlarge the dataset to include other domains, provide solutions to issues of proper noun extraction and extend the system for the extraction of entities including weapon names, and crime type. In addition, the proposed Arabic NER approach can be used with other classifiers. Thus, this study contributed to the body of knowledge and provided appropriate platform to enhance crime analysis in Arabic.

REFRENCES:

[1] R. Alfred, L. C. Leong, C. K. On, P. Anthony, T. S. Fun, M. N. B. Razali*, et al.*, "A Rule-Based Named-Entity Recognition for Malay Articles," in *Advanced Data Mining and Applications*, ed: Springer, 2013, pp. 288-299.

[2] K. Shaalan and M. Oudah, "A hybrid approach to Arabic named entity recognition," *Journal of Information Science,* vol. 40, pp. 67-87, 2014.

[3] H. Traboulsi, "Arabic named entity extraction: A local grammar-based approach," in *IMCSIT*, 2009, pp. 139-143.

[4] N. F. Mohammed and N. Omar, "Arabic named entity recognition using artificial neural network," *Journal of Computer Science,* vol. 8, p. 1285, 2012.

[5] M. S. Habib and J. Kalita, "Language and domain-independent named entity recognition: experiment using SVM and high-dimensional features," in *Proc. of the 4th Biotechnology and Bioinformatics Symposium (BIOT-2007), Colorado Springs, CO*, 2007.

[6] R. Al-Shalabi, G. Kanaan, B. Al-Sarayreh, K. Khanfar, A. Al-Ghonmein, H. Talhouni*, et al.*, "Proper noun extracting algorithm for arabic language," in *International conference on IT, Thailand*, 2009.

[7] M. Asharef, N. Omar, M. Albared, Z. MINHUI, W. WEIMING, Z. JINGJING*, et al.*, "Arabic named entity recognition in crime documents," *Journal of Theoretical and Applied Information Technology,* vol. 44, pp. 1-6, 2012.

[8] M. Oudah and K. F. Shaalan, "A Pipeline Arabic Named Entity Recognition using a Hybrid Approach," in *COLING*, 2012, pp. 2159-2176.

[9] S. V. Nath, "Crime pattern detection using data mining," in *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*, 2006, pp. 41-44.

[10] V. Pinheiro, V. Furtado, T. Pequeno, and D. Nogueira, "Natural language processing based on semantic inferentialism for extracting crime information from text," in *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on*, 2010, pp. 19-24.

[11] S. Mesfar, "Named entity recognition for arabic using syntactic grammars," in *Natural Language Processing and Information Systems*, ed: Springer, 2007, pp. 305-316.

[12] Y. Benajiba, P. Rosso, and J. M. Benedíruiz, "Anersys: An arabic named entity recognition system based on maximum

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

entropy," in *Computational Linguistics and Intelligent Text Processing*, ed: Springer, 2007, pp. 143-153.

[13] C. Shihadeh and G. ünter Neumann, "ARNE-A tool for Namend Entity Recognition from Arabic Text," in *Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4), located at the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012, pp. 24-31.

[14] I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig, and N. A. Mahyoub, "Automatic Arabic text categorization: A comprehensive comparative study," *Journal of Information Science,* p. 0165551514558172, 2015.

[15] M. Oudah and K. Shaalan, "Person name recognition using the hybrid approach," in *Natural Language Processing and Information Systems*, ed: Springer, 2013, pp. 237-248.

[16] A. K. Mandal and R. Sen, "Supervised Learning Methods for Bangla Web Document Categorization," *International Journal of Artificial Intelligence & Applications,* vol. 5, pp. 93-105, 2014.

[17] J. Cai, "Decission Tree Pruning Using Expert Knowledge Dissertation," p. 231, 2006.

[18] G. Paliouras, V. Karkaletsis, G. Petasis, and C. D. Spyropoulos, "Learning decision trees for named- entity recognition and classification," in *ECAI Workshop on Machine Learning for Information Extraction*, 2000.

[19] H. Chauhan and A. Chauhan, "Implementation of decision tree algorithm c4. 5," *International Journal of Scientific and Research Publications,* vol. 3, 2013.