Journal of Theoretical and Applied Information Technology
30th November 2016. Vol.93. No.2

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

FEATURE SELECTION IN WEB NER USING GENETIC ALGORITHM APPROACH

[1]MOHAMMED MOATH ABDULGHANI, [2]SABRINA TIUN

Center for Artificial Intelligence Technology (CAIT)

E-mail: [1]albakre2@gmail.com, [2]sabrinatiun@ftsm.ukm.my

ABSTRACT

Named Entity Recognition (NER) is the field of recognizing nouns such as names of people, corporations, places and dates. The process of extracting NEs is mainly relying on supervised machine learning techniques. Hence, utilizing proper features have a significant impact on the performance of recognizing the entities. Several approaches have been proposed for reducing the feature dimensionality of NER. However, these approaches have concentrated on the traditional features or so-called textual features. Recently, extracting information form web pages has caught the researchers' attentions regarding the valuable information that lies on such pages. Extracting NEs from web pages has brought tremendous kinds of features which are inspired from the web nature. Apparently, combining the traditional features with the web features would expand the feature space. Therefore, there is an essential demand for accommodating feature selection for the process of extracting NEs from web pages. This paper proposes a feature selection approach based on Genetic Algorithm for extracting NEs from web pages. The dataset was collected from business web pages. Whilst, the feature set consists of text features such as n-gram and web features such as block position and font type. Finally, a SVM classifier was used to classify the NEs. Results shown that Genetic Algorithm has the ability to identify the most accurate features

Keywords: *Named Entity Recognition, Feature Selection, Genetic Algorithm, Support Vector Machine, Web Pages*

1. INTRODUCTION

With the tremendous increasing of the information over web, there is an essential demand for extracting several types of these information [1]. One of the common information that have been addressed by many researchers are the NEs. Named Entity Recognition (NER) is the task of identifying proper nouns such as person's name, organization's name, location's name and dates [2]. The key challenge behind these entities lies on the frequent occurrences in which these entities are being used widely in various domains. NEs can be classified into main groups; nominal and numeric entities [3]. Nominal entities are the instances that can be expressed via words such as people's name (e.g. John), location's named (e.g. Washington) and organization's name (e.g. Dell). Whereas, numeric entities are the instances that can be expressed via numerals such as dates (e.g. 1937), time (e.g. 5:39) and currencies (e.g. 38 USD).

The traditional approaches that have been proposed for extracting NEs relied on handcrafted rule-based methods in which a set of rules are being constructed to recognize the NEs [4]. These rules were usually developed based on the grammar, context and lexical approaches. One of the common rules is the keyword approach in which the words that frequently occur with NEs such as 'London city', 'Microsoft corporation' and 'Dr. Smith' [5]. However, with the complexity and time-consuming of building the rules, machine learning techniques have caught many researchers regarding to its ability to generate the rules automatically based on a statistical model [6]. The key characteristic behind machine learning techniques lies on utilizing proper and precious features that have the ability to identify the NEs. Features can be defined as the attributes or characteristics that aim to distinguish the occurrences of predefined classes [7].

In fact, there are vast amount of features that could be used for NEs including n-gram feature which aims to address every single word. In addition, term frequency feature has been also used for NER by identifying the most frequent terms within the text. Furthermore, capitalization feature was addressed too since most of the proper nouns are being formed with capital letters. In addition, keyword feature has been considered too for identifying NEs in which the words that are associated with these entities are being stored in lists.

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

Recently, a new type of NEs has been emerged which is the NEs appear in web pages especially Yellow Pages. Such pages contain tremendous amount of NEs such as organization's name, physical location, contacts (e.g. emails, telephone and fax). Hence, researchers tend to address NER from these web pages [8, 9]. In this manner, the feature space has dramatically increased where new features emerged to be used for identifying NEs such as font type, HTML tags, hyperlinks and block position. This expansion of feature space requires accommodating a feature selection task in which the optimized feature set can be identified. This study aims to accommodate such feature selection task using Genetic Algorithm with Support Vector Machine classification.

2. RELATED WORK

Since the named entity recognition is mainly relied on supervised techniques thus, the feature plays an essential role in terms of improving the performance of recognition. Several researchers have examined traditional types of textual features such as n-gram, affixes and term occurrence. For example, Pinheiro et al. [10] have addressed Natural Language Processing (NLP) features in terms of identifying named entities from web. The authors have utilized lexical features such as n-gram, as well as, semantic features such as lexicon. In addition, [11, 12] have addressed the affixes features in terms of identifying named entities.

Hence, plenty features are being utilized for the recognition of named entities. Such dimensionality of features requires a selection tasks in order to identify the optimized set of features. This is due to the strength's variations among these features in which some of them could be trivial in some cases and the other could be powerful.

In this manner, Hasanuzzaman et al. [13] have proposed a feature selection approach in order to reduce the dimensionality of features used in NER. Hence, a Genetic Algorithm (GA) is being used to identify the best combination of features. In this manner, multiple features have been used including POS tagging, length of words, affixes and frequency of words. These features have been encoded as chromosomes for identifying the optimized features using GA. After that, the best combination is being used to classify the entities based on a Maximum Entropy (ME) classifier. Results shown that the performance of ME with the best combination of features has outperformed the performance of ME with all features.

Similarly, Ekbal & Saha [14] have used GA as a feature selection approach for determining the optimized feature set. Basically, the authors have a hypothesis emphasis the challenging task for identifying the best features in accordance to the NE classes (i.e. PER, ORG, LOC, etc.). Since there are many classes thus, the process of identifying the appropriate features for such classes is a difficult task. In this manner, GA has been applied on CoNLL-2003 benchmark dataset. The results shown that the application of GA has a significant impact in terms of identifying the best combination of features which significantly improved the accuracy of classification.

Recently, new kind of features have been addressed for the extraction of named entities. This new kind is relying on web features for example, Song et al., [15] have proposed a learning approach that identify the importance of blocks within the web pages. First, they used Vision-based Page Segmentation (VIPS) algorithm in order to segment the web pages into semantic blocks with hierarchy structure. Hence, specific features have been extracted from the block such as font, size, position, number of images and links. These extracted features have been formed as a vector space model where it can be fed to a classifier for learning. For this purpose, Support Vector Machine (SVM) and Neural Network (NN) classifiers have been used in order to train on the vector and then classify each block to its corresponding importance.

Moreover, Zhang et al. [16] have discussed the task of extracting locations from web pages. In particular, the authors have discussed the problem of ambiguity that lies on the location names. Such ambiguity can be represented in the multiple locations that yield the same name, or the locations that yield other location name such as person name. Hence, they have proposed an approach called GeoRank which is similar to the page ranking in terms of ranking the Geo locations based on its relevancy. Their approach has successfully eliminated the ambiguity of Geo location names.

SVM classifier has been employed successfully through classifying the data instances directly into their actual classes. The evaluation explains that the proposed SVM classification is performed better with the optimized feature selected by GA. This kind of achievement is significantly contribute toward improving the information extraction from web especially the NEs.

Journal of Theoretical and Applied Information Technology
30th November 2016. Vol.93. No.2

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

Essentially, in the literature, multiple authors have sorted out the use of feature selection with regards to NER such as Hasanuzzaman et al. (2011) and Ekbal & Saha (2011). However, there approaches of feature selection has been performed on text data such as CoNLL-2003 benchmark dataset. With the surge of research in terms of determining NEs from web pages. Moreover, there is an on-demand application of feature selection regarding to web-NER. This is because of the dimensionality of features that would be utilized for such entities. For that reason, this study aims to propose a feature selection approach by using GA for extracting NEs through web pages.

3. PROPOSED METHOD

The proposed method comprises of five main stages as shown in Fig. 1. First stage is associated with the data that will be used in the experiment where a set of web pages are being utilized for such purpose. Second stage aims to accommodate a transformation task that intended to normalize and segment the web pages in order to turn it into proper form enables processing. Third stage aims to extract features from the transformed form of the web pages in which two types of features are being utilized; web and text features. Fourth stage represents the contribution of this study in which a Genetic Algorithm is being carried out upon the extracted features in order to select the best combination of features. Finally, the fifth stage is associated with the classification process in which a Support Vector Machine (SVM) classifier is being used to classify the named entities.

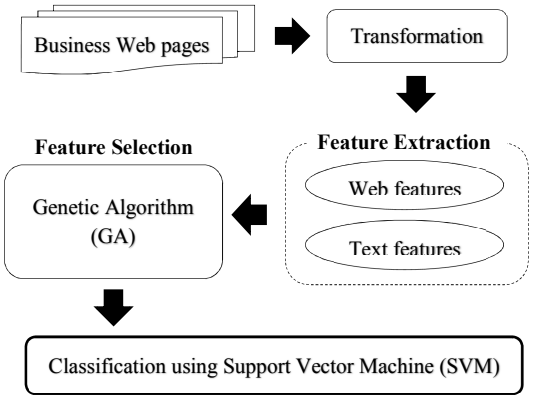

Figure1. Proposed method

3.1. Business Web Pages

The dataset used in this study consists of business web pages that contains various information about firms and corporations. This information represents the named entities such as companies' names, emails, addresses, telephones and fax numbers. A Yellow pages website (http://www.yellowpages.ae) has been used to collect (i.e. download) the business web pages which is associated with UAE companies [17]. Fig. 2. Shows a sample of such Yellow pages.

Figure 2. Sample of Yellow pages

3.2. Transformation

Since the data is consisting of web pages thus, the representation will be based on HTML tags. Hence, in order to exploit the nature of text for feature extraction, it is necessary to accommodate a transformation task in order to get rid of the unwanted data. Fig. 3 depicts such task.

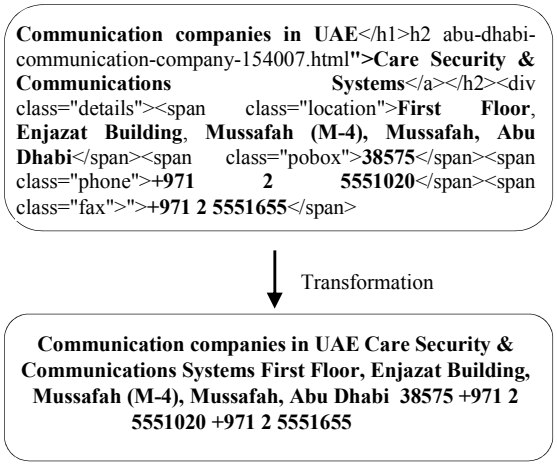

Figure 3. Transforming web pages

3.3. Feature Extraction

This phase aims to extract several kinds of features. Basically, the features play an essential role in terms of identifying NEs [6]. This can be represented by the discrimination and generalization that could be produced by the features which facilitate the process of classifying the NEs. For this manner, two kinds of features are being used by this study which are text features and

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

web features. These kind of features are being illustrated in the following sub-sections.

3.3.1 Text Features

In fact, text features for NER are vary and have been used in the literature frequently. The common textual feature is the *Part-Of-Speech (POS) tagging* feature which has been proposed to identify the syntactic class of the words such as verb, noun, adjective and others. The syntactic analysis of the words could yield significant indication whether the word is named entity or not. This is due to most of the named entities are lying under the 'noun' tag. In this manner, if the POS tag of a particular word is 'noun', then the probability of being named entity will be increased. For this purpose, a Stanford POS tagger [18] has been used to parse the words.

In addition, *keyword* feature is one of the textual features that have been extensively examined in terms of named entity recognition. This feature aims to exploit the words that are frequently occurred with NEs such as the keyword 'city' which usually occurred with locations such as 'Dubai city'. Basically, this feature has been utilized by many authors for identifying NEs [6, 11, 19, 20]. In this manner, there are several kinds of keywords including Organization's keywords, Location's keywords and others. For this purpose, a gazetteer for such keywords has been used, this gazetteer introduced by the study of Nadeau et al. [21].

Furthermore, *capitalization* feature is also one of the textual feature that has been commonly utilized for named entity recognition. This feature aims to investigate whether the word is capitalized or not. This is due to most of the NEs are being capitalized in the web pages. Therefore, this feature could be a useful feature to identify whether the word is NE or not. However, most of the traditional text classification approaches tend to lower all the letters as a pre-processing task. This study has ignored such task and attempted to keep the letters with their original case in order to identify the capitalization.

3.3.2 Web Features

Since our study is concentrating on recognizing named entities from web pages therefore, identifying web features could be a significant task. Many researchers have used web features to identify NEs from web pages [8, 22, 23]. However, extracting web features is relying on the nature of the data. According to our dataset which is represented using HTML, it is possible to extract multiple features. For example, *font* feature is one of the common characteristic of web which aims to analyze the font type of the words. Basically, in web page, the words are being described with multiple types of font in which the size and color are being distinguished. To identify the font type, the HTML tags of '<h1>', '<h2>', '<h3>' and '<h4>' have been utilized which has the ability to determine the font type. Basically, these tags are being used frequently in HTML to identify the size of fonts. Basically, the tags <h1>, <h2> and <h3> show different font sizes in which the <h1> is the biggest, <h2> is smaller and <h3> is smaller and so on. In this manner, named entities could be formed with specific tag thus, utilizing the font size could be a useful indicator to identify the named entities.

Another web feature is the *URL*. In web pages, assigning the word with a URL is possible in which clicking upon the word by the user would result on opening a new page that may describe such word. In this manner, NEs are being formulated in some web pages as a URL which may indicate an address, contact, email or other entities. The HTML tag used to formulate a URL is '<a href= URL>' in which the user can click on such URL to know more details about such entity. Therefore, it is an opportunity to utilize the URL feature.

Finally, the *block position* is a web feature that aims to determine the position of the word whether the word is being represented in the top, middle or in the bottom. For example, telephone numbers are usually represented in the bottom. Therefore, this feature could be useful to identify some of the entities. For this purpose, the HTML tags of '<header>', '<nav>', '<section>' and '<footer>' are being utilized in which the first tag refers to the top of the page, the second tag refers to the left, the third tag refers to the middle and the fourth tag refers to the bottom.

Basically, all the mentioned features used in this study whether text or web features are being articulated based on *Term Occurrence*. Unlike term frequency, term occurrence aims to identify whether the feature is occurred in the selected word or not [24]. This can be represented as '1' or '0' where '1' refers to the presence of a feature in a particular word, whereas '0' refers to the absence of a feature in a particular word.

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

3.5. Feature Selection

In fact, there are features that have significant impact on the performance, whereas other features have insignificant effect. Hence, there is a vital demand to identify a best combination of the features in order to identify the strong features and weak features. In this study, Genetic Algorithm (GA) feature selection approach has been carried out. GA is one of the local search techniques that mimics the biology of natural selection [25]. Basically, it works by analyzing a population in which the possible solutions are being addressed. Then, a combination process is performed by combining the best population based on a fitness function. Basically, GA begins with generating an initial population in which the possible solutions are being represented. Such population contains multiple genes (i.e. features) which consist of chromosomes. Every chromosome is represented by a binary value (i.e. 0 or 1). After generating the initial population, every gene should be evaluated in terms of feasibility. Such evaluation is conducted based on fitness function which determines the desired effectiveness using a value. After the best genes have been chosen, a re-production process is being conducted which aims to generate the next population. This can be performed by manipulating the chromosomes of the current genes. For this purpose, there are multiple methods can be used to re-produce the genes such as crossover, mutation and elitism. In this study, crossover method has been used to re-produce the next generation. Table 1 depicts the parameters of implementing GA.

Table 1. GA Parameters adjustment

Parameter	Description
Number of features (Chromosomes size)	267
Number of iterations or generations	50
Re-production mechanism	Crossover
Condition for termination	No significant changes in next generations

Basically, many researchers have argued in terms of the GA termination conditions. According to Safe et al. [26] who accommodate a review on the criteria used for stopping GA, there are three conditions can be used for the termination. These conditions can be illustrated as:

- An upper limit on the number of generations is reached,
- An upper limit on the number of evaluations of the fitness function is reached,
- The chance of achieving significant changes in the next generations is excessively low.

The first condition is associated with the cases when a predefined number of generations is being determined. Apparently, such condition is required when the goal is to enhance the efficiency (i.e. minimize the iterations). Second condition is associated with the cases when a predefined target or class label is being determined. In such case, once the fitness function reaches the target GA will stop. Finally, the third condition is simply emphasizing that if there is no significant change has been occurred in the next generation, terminate GA.

Since this study is concerned with the effectiveness (i.e. accuracy of NEs classification) rather than the efficiency thus, the first condition will not be considered. In addition, the second condition will not be an appropriate criterion for this study because the target is not constant but rather it consists of multiple classes. Therefore, the third condition will be considered in which the termination will take a place once there is no significant changes are being achieved.

3.6 Classification using SVM

This phase aims to classify the named entities into their actual classes based on the selected features from GA. In fact, SVM is a non-probabilistic and binary classifier where the classes are divided into two groups (0 or 1). Basically, SVM aims to classify the data by assigning a hyperplane which is a margin that divide the data into two class labels shown in Table 2. This can be performed by identifying a margin that divide the data with the first class (i.e. Organization) and the remaining data (shown in step 2). After that, identifying a margin that divide the data with the second class (i.e. Location) and the remaining data (shown in step 3). Consequentially, identifying a margin that divide the data with the third class (i.e. URL) and the remaining data (shown in step 4). After that, identifying a margin that divide the data with the fourth class (i.e. Phone) and the remaining data (shown in step 5). After that, identifying a margin that divide the data with the fifth class (i.e. Fax) and the remaining data (shown in step 6). After that, identifying a margin that divide the data

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

with the sixth class (i.e. POBOX) and the remaining data (shown in step 7).

Table 2. Pseudo code of the proposed SVM

Algorithm Support Vector Machine
Input Data *D* with classes of Organization, Location, URL, Phone, Fax and POBOX
1. **Divide** the data into 80% training and 20% testing
For each training data
2. Identify the hyperplane between the Organization and not Organization
3. Identify the hyperplane between the Location and not Location
4. Identify the hyperplane between the URL and not URL
5. Identify the hyperplane between the Phone and not Phone
6. Identify the hyperplane between the Fax and not Fax
7. Identify the hyperplane between the POBOX and not POBOX
End For
For each testing data
7. Classify the instances based on the hyperplanes
End For

4. RESULTS

To investigate the impact of feature selection using GA, SVM has been used twice; first with the optimized feature set resulted from GA, and second SVM was conducted with all the features. The evaluation has been performed using the common information retrieval metrics Precision, Recall and F-measure. Precision can be calculated as follow:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

where TP is the number of correctly classified entities, and FP is the number of incorrectly classified entities. On the other hand, recall can be calculated as follow:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

where TP is the number of correctly classified entities, and FN is the number of instances that were not being classified. Finally, f-measure can be calculated as follow:

$$F - measure = \frac{2Pr \times Re}{Pr + Re} \quad (3)$$

Table 3 depicts both results of SVM and SVM with GA using precision, recall and f-measure for all the classes including organization, location, URL, phone, fax and PO-Box.

Table 3. Experimental results

Classes	SVM	SVM-GA
ORG	0.79268	0.96587
LOC	0.80506	0.97324
URL	0.75558	0.95609
Phone	0.76832	0.94670
Fax	0.79206	0.93964
POBOX	0.79074	0.90220
Average	**0.78407**	**0.93952**

As shown in Table 3, the application of SVM with GA has outperformed SVM with all the features for all classes including Organization (i.e. 96% compared to 79%), Location (i.e. 97% compared to 80%), URL (i.e. 95% compared to 75%), Phone (i.e. 94% compared to 76%), Fax (i.e. 93% compared to 79%), POBOX (i.e. 90% compared to 79%). Finally, as an average percentage of f-measure, SVM with GA has outperformed SVM by obtaining 93% compared to 78%.

The evaluation of the proposed approach has been divided into two phases; first SVM will be applied using all the feature set, second SVM will be applied using the optimized feature set. Results shown that the application of SVM with GA as a feature selection approach has outperformed the application of SVM without GA by achieving 93% of f-measure. This emphasis the usability of GA in terms of identifying an optimized feature set for NER. In this manner, the objective of identifying an optimized feature set for extracting NEs from web pages, is being accomplished.

There are three categories of NE features namely; Boolean, Numeric and Nominal. Boolean feature is the one that associated with validation cases in which the representation of such feature is true or false, where true refers to the presence of feature and false refers to the absence of such feature. The common example of this feature is capitalization in which the given word is being checked or validated in terms of 'is-capitalized' condition. Numeric feature is the one that associated with numeric representation in which the given word is being examined in terms of numeric attribute. The common example of this feature is the number of characters where the representation of this feature. Nominal feature is the one that associated with string representation in which the morphology of the given word is being examined. The common

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

example of this feature is the affixes where the word is being analyzed in terms of its prefixes or suffixes. However, there are more analysis can be elaborated on the features of NE such as word-level features, capitalization, special characters, numbers, affixes, syntactic class, and word's length.

5. DISCUSSION

Through the variety of features that have been utilized for named entities, there is an essential demand to indicate the most accurate feature set. This process is called feature reduction or feature selection and it has a significant effect on the performance [27]. This is because there are plenty of features that could possibly be not important and have an insignificant correlation with the process of determining named entities. For this reason, Genetic Algorithm has been applied to identify the optimized feature set.

Based on the obtained results, the application of SVM with GA has outperformed the application of SVM with all the features. This can demonstrate the capability of GA in terms of identifying the optimized feature set for extracting named entities from web pages. This has been expected from the study of Hasanuzzaman et al. [13] who demonstrated an enhancement in the performance of NER when using GA as a feature selection approach in order to reduce the dimensionality of features used in NER. Similarly, Ekbal & Saha [14] attained the best combination of features for NER using GA. Apart from NER, GA has been enhanced the classification process of SVM in various s [28], power planning [29], optimization [30] and handwritten recognition [31].

This study focused to accommodate a feature selection task using GA with SVM classification. The dataset has been collected from Yellow Pages which contains business web pages. The data has been undergone multiple tasks of preprocessing including normalization where the irrelevant data has been removed, and segmentation which aims to divide the text into series of tokens. After that, the proposed feature set has been extracted including text and web features. Consequently, GA has been performed on the feature set in order to select the optimized features. Then SVM will classify the data based on the optimized feature set. The evaluation has been conducted using precision, f-measure, and recall.

6. CONCLUSION

This paper has intended to identify an optimized feature set for extracting NEs from business web pages using GA. The features dimension contains text features such as POS tagging, keywords and capitalization. As well as, it contains web features such as font, URL and block position. Consequentially, an SVM classifier is being applied in order to classify the NEs. Results of classification shown an improvement in the performance when using the optimized feature set.

This paper has successfully achieved the aims by identifying an optimized feature set for extracting NEs using GA as a feature selection approach. Additionally, SVM classifier has been applied successfully by classifying the data instances into their actual classes. The evaluation reveals that the proposed SVM classification is performed better with the optimized feature selected by GA. Such achievement is significantly contribute toward improving the information extraction from web especially the NEs.

A number of studies have concentrated on determining best combinations of features of NER. In essence, the authors have a hypothesis emphasis the challenging task for identifying the best features in accordance to the NE classes (i.e. PER, ORG, LOC, etc.). Since there are many classes thus, the process of identifying the appropriate features for such classes is a difficult task.

However, there is a surge research these days in the field of NER. Such research is represented by extracting the NEs from web pages. Hence, new features have been introduced for web NER such as font format, URL and block positions. This means that the dimensionality of features would be expanded. Therefore, there is a vital demand to accommodate a feature selection approach in order to identify the best feature set for web NER. Therefore, this study aims to utilize various kinds of features in order to identify the best combination that has the ability to handle large number of entities.

Choosing a few feature selection approaches and the dataset has been collected merely from Yellow Pages website could be a limitation of this study. Therefore, Using different feature selection approaches is a key challenging issue in which Ant Colony, Particle Swarm Optimization and Simulated Annealing are competitive feature selection methods compared to GA. Addressing other domains rather than business web pages

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

would be a significant in future researches where handling new types of NEs such as biomedical NEs could be a challenging task. Using multiple classifiers or even a combination would also improve the effectiveness where each classifier has its own ability.

7. ACKNOWLEDGMENT

This study is supported and funded by the University Kebangsaan Malaysia (UKM).

REFERENCES

[1] J. Cowie and W. Lehnert, "Information extraction," *Communications of the ACM,* vol. 39, pp. 80-91, 1996.

[2] E. Alfonseca and S. Manandhar, "An unsupervised method for general named entity recognition and automated concept discovery," in *Proceedings of the 1st International Conference on General WordNet, Mysore, India*, 2002, pp. 34-43.

[3] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes,* vol. 30, pp. 3-26, 2007.

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

[5] D. Farmakiotou, *et al.*, "Rule-based named entity recognition for Greek financial texts," in *Proc. of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, 2000, pp. 75-78.

[6] S. Abdallah, *et al.*, "Integrating rule-based system with classification for Arabic named entity recognition," in *Computational Linguistics and Intelligent Text Processing*, ed: Springer, 2012, pp. 311-322.

[7] Y. Benajiba, *et al.*, "Arabic Named Entity Recognition: A Feature-Driven Study," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 17, pp. 926-934, 2009.

[8] O. Etzioni, *et al.*, "Unsupervised named-entity extraction from the web: An experimental study," *Artificial Intelligence,* vol. 165, pp. 91-134, 2005.

[9] S. Sumathipala, *et al.*, "Protein Named Entity Identification Based on Probabilistic Features Derived from GENIA Corpus and Medical Text on the Web," *International Journal of Fuzzy Logic and Intelligent Systems,* vol. 15, pp. 111-120, 2015.

[10] V. Pinheiro, *et al.*, "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] H. Shabat, *et al.*, "Named Entity Recognition in Crime Using Machine Learning Approach," in *Information Retrieval Technology*, ed: Springer, 2014, pp. 280-288.

[12] S. A. H. Al-Shoukry and N. Omar, "Arabic named entity recognition for crime documents using classifiers combination," *International Review on Computers and Software,* vol. 10, pp. 628-634, 2015.

[13] M. Hasanuzzaman, *et al.*, "Feature subset selection using genetic algorithm for named entity recognition," in *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, PACLIC*, 2011, pp. 153-162.

[14] A. Ekbal and S. Saha, "Weighted vote-based classifier ensemble for named entity recognition: a genetic algorithm-based approach," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 10, p. 9, 2011.

[15] R. Song, *et al.*, "Learning block importance models for web pages," in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 203-211.

[16] Q. Zhang, *et al.*, "Extracting focused locations for web pages," in *Web-Age Information Management*, ed: Springer, 2012, pp. 76-89.

[17] U. Y. Pages. (2016). *UAE Online Business Directory*. Available: http://www.yellowpages.ae/

[18] Stanford, "Part-of-Speech Tagger," ed, 2014.

[19] M. Marrero, *et al.*, "Named entity recognition: fallacies, challenges and opportunities," *Computer Standards & Interfaces,* vol. 35, pp. 482-489, 2013.

[20] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *Journal of biomedical informatics,* vol. 46, pp. 1088-1098, 2013.

[21] D. Nadeau, *et al.*, "Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity," 2006.

[22] M. Pasca, "Acquisition of categorized named entities for web search," in *Proceedings of the thirteenth ACM international conference on*

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

Information and knowledge management, 2004, pp. 137-145.

[23] M. Paşca, "Weakly-supervised discovery of named entities using web search queries," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 683-690.

[24] M. Yamamoto and K. W. Church, "Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus," *Computational Linguistics,* vol. 27, pp. 1-30, 2001.

[25] L. Ferreira*, et al.*, "Using a Genetic Algorithm Approach to Study the Impact of Imbalanced Corpora in Sentiment Analysis," in *The Twenty-Eighth International Flairs Conference*, 2015.

[26] M. Safe*, et al.*, "On stopping criteria for genetic algorithms," in *Brazilian Symposium on Artificial Intelligence*, 2004, pp. 405-413.

[27] F.-b. Wang and X.-s. Xu, "A new feature selection method in text categorization," *Journal of Shandong University (Engineering Science),* vol. 4, p. 001, 2010.

[28] A. H. Kassem*, et al.*, "Improving Satellite Orbit Estimation Using Commercial Cameras," *International Review of Aerospace Engineering (IREASE),* vol. 8, pp. 174-178, 2015.

[29] N. F. Ab Aziz*, et al.*, "Reactive Power Planning for Maximum Load Margin Improvement Using Fast Artificial Immune Support Vector Machine (FAISVM)," *International Review of Automatic Control (IREACO),* vol. 7, pp. 436-447, 2014.

[30] L. Demidova*, et al.*, "Use of Fuzzy Clustering Algorithms' Ensemble for SVM classifier Development," *International Review on Modelling and Simulations,* vol. 8, pp. 446-457, 2015.

[31] T. R. V. Lakshmi*, et al.*, "Hybrid Approach for Telugu Handwritten Character Recognition Using k-NN and SVM Classifiers," *2015,* vol. 10, 2015.