ENHANCING THE PERFORMANCE OF SENTIMENT ANALYSIS SUPERVISED LEARNING USING SENTIMENTS KEYWORDS BASED TECHNIQUE

Amira Abdelwahab[1], Fahd Alqasemi[2] and Hatem Abdelkader[3]

[1,2,3]Information Systems Department, Menoufia University, Menoufia, Egypt
[1]amira.ahmed@ci.menofia.edu.eg
[2]fhdahmd16@yahoo.com
[3]hatem6803@yahoo.com

ABSTRACT

Sentiment Analysis (SA) and machine learning techniques are collaborating to understand the attitude of text writer, implied in particular text. Although, SA is an important challenging itself, it is very important challenging in Arabic language. In this paper, we are enhancing sentiment analysis in Arabic language. Our approach had begun with special pre-processing steps. Then, we had adopted sentiment keywords co-occurrence measure (SKCM), as an algorithm extracted sentiment-based feature selection method. This feature selection method had utilized on three sentiment corpora using SVM classifier. We compared our approach with some traditional methods, followed by most SA works. The experimental results were very promising for enhancing SA accuracy.

KEYWORDS

sentiment analysis; opinion mining; supervised learning; feature selection; Arabic language

1. INTRODUCTION

Social networks (SN) have changed the nature of communication between people around the world, due to the big and fast tangible ability of sharing information. It also resulted in different SN applications which gain very important popularity. One of SN precious content is opinions that people share about everything: products, sports, people, politics, science ... etc. SN contents are varying, concerning stored data types. However, the most type of information that had shared is text, specially, in the context of opinion expression, since people are expressing their opinions in written words regardless used language. This made text mining importance, so as natural language processing (NLP), and sentiment analysis (SA), regarding Opinion mining (OM), which is another term for sentiment analysis.

Sentiment analysis, where is text mining, machine learning and NLP techniques are collaborating to discover the direction of text segment. This depends on the implication of an attitude behind written text. The text may be a document, review, article or SN post/comment. The discovered direction is called polarity. Text polarity is determined by agree or disagree of text writer with any subject, which text talks about. Agree and disagree is expressed in SA studies as 'positive' and

David C. Wyld et al. (Eds) : CSITA, ISPR, ARIN, DMAP, CCSIT, AISC, SIPP, PDCTA, SOEN - 2017
pp. 107– 116, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70111

'negative', respectively. Some papers had added the case of 'neutral' to the above two polarities, which make them three directions [1]. In this work, we had applied our proposed technique only on the two first mentioned polarities, i.e. positive and negative.

SA is achieved using two popular approaches, machine learning sentiment analysis (MLSA), and lexicon-based sentiment analysis (LBSA). MLSA is intended for supervised learning approaches that need pre-annotated text data set, to train and test some selected features. On the other hand, LBSA is considered as unsupervised learning approach, which is based on a pre-made sentiment lexicon. That lexicon is used to count sentiment terms appearance on target text [1] [2].

Sentiment mining in Arabic language is faced numerous difficulties. These difficulties related to some factors; firstly, the nature of Arabic as rich morphological language, and the varying of Arabic dialects which are overlapped with the main form of Modern Arabic Standard, i.e. MSA, especially in SN web pages. Secondly, the rare of researches in these new fields in Arabic. That had led to the lack of resources and tools that may help to create more and big techniques. Although, there were good efforts of Arabic researchers in the NLP, text mining, and sentiment analysis, they need more works.

In this paper, we present an SA approach, that have implemented MLSA on Arabic language, used resources had annotated into the two polarities, 'negative' and 'positive', however. We had exploited LBSA in the purpose of improving MLSA. This had been done by using sentiment lexicon as sentiment keywords list, which implemented as a feature selection method, which helped for increasing the performance of supervised learning task on three corpora. We had compared our approach with other traditional and recent approaches, and the results were very promising.

Literature included some methods adopted Arabic text pre-processing. Our approach took an advantage step, thus we begun with the much recommended pre-processing process, which helped in removing corpus noise, and improved the computation complexity by reducing the size of feature set variables, simultaneously. Further explanation of our methodology is in section 3, which based on sentiment keywords co-occurrence measure (SKCM), SKCM is an algorithm we had used as a measure of each corpus term polarity. It was not considered as term polarity, but, sentiment weight which give better results when replaced with term appearance, and outperformed term frequency that had been used in different figures in many literatures [3] [4].

Our selected feature method had tested in support vector machine (SVM) classifier. Thus, we had classified each corpus of our three corpora three times. Firstly, via traditional feature selection method. Then with our SKCM results. After that, we tested them with the combination of these two FS methods. The results in all three corpora had showed the preference of SKCM results versus both two mentioned methods results.

This paper has related works brief in section 2. Then methodology explanation is presented in section 3, whereas section 4 shows experiments details and results illustration. Finally, we concluded this paper at section 5.

2. RELATED WORKS

The work of [5] had built domain-specific sentiment dictionary, automatically. This domain was movie reviews in Korean language. They used huge set of online movie reviews for finding out the joint probability of dictionary words This joint probability is between words and their position wither in negative annotated reviews, or positive annotated reviews. Then they give each word a polarity which was the difference between those two joint probability values. At last, they had

normalized these values and obtained 135,082 words that constitute their outcomes of sentiment dictionary in Korean language. However, the method of [5], was gave each term a polarity value, likewise our approach, but we didn't need annotated data set for finding that polarity, due to our approach do not find that polarity by advanced knowledge of document polarity, where this term appeared in.

The work of [6], had collected and prepared the biggest Arabic language opinion mining data set to-date. It is large Arabic book reviews corpus LABR, which included over 63,000 text segments, distributed between five rating values scaled from 1 and 2 for negative opinionated reviews, 3 as neutral reviews, whereas 4 and 5 as positive sentiments. Then, they study the data set characteristics, and testing the corpus for sentiment analysis. Here, we couldn't take the whole LABR corpus. So, we just take randomly two parts of LABR, one is balanced corpus, and another as unbalanced corpus.

[4] had exhibited one of the popular Arabic sentiment corpora. For the purpose of research, they had called it OCA as opinion corpus for Arabic. OCA is consisted of just 500 movie reviews, gathered from different Arabic blogs and pages in the web. OCA didn't include short text segments. Thus it had articles that led to big number of unique words. They carried on OCA some SA experiments using SVM and other classifiers. The features selection used was traditional feature selection which based on terms frequency (TF), where they concluded that TF had the same results of TFIDF, whether in the form of unigram, bigram, or trigram.

The work of [7], had utilized a mathematical tool based on rough set (RS) theory. They focused on the differences between RS reducts in Arabic SA. They used Term Weighting using TFIDF as a feature selection method, which based on counting terms per text segments, over all corpora. Rosetta toolkit which is designated for RS was used, with cross validation evaluation. They achieved an accuracy of 57% as the best result comparing other RS reducts. They concluded the applicability of using RS for Arabic SA, which need more enhancements, especially, in regard of reduct methods.

[3] had presented Arabic SA for Arabic tweets, they had collected Arabic dataset from twitter. Their results showed that SVM classifier was outperformed NB classifier, where they had used traditional feature selection that based on term frequency in the form of unigram and bigram. They had presented also that eliminating stop-words enhanced the results of sentiment classification accuracy. In our work, we re-implemented feature selection which based on traditional methods, and comparing resulted accuracy with ours.

The work of [8] had exhibited the concept of semantic orientate on (SO) for text segments. This concept was calculated based on counting the number of adjective or adverbs terms appeared in text. Also, SO was resulted using Mutual Information relation between text and polarity advanced known terms, such as "excellent" for positive polarity, or "poor" for negative polarity.

A text is considered as positive or negative based on SO calculated value. They achieved different accuracies in different domain in English language. The concept of semantic orientation had been used in literature in various applications such as in [9] [10] [8]. In our work, we didn't use SO as it is in [8]. We satisfied with the value of terms co-occurrence which is easier to be calculated than mutual information. And replaced the adjective (or adverb) usage, with sentiment keywords list that we had generated in [2] from pre-existed sentiment lexicon [6]. Also, we calculated it for terms not for segments of text. Then distributed resulted terms values over the binary matrix that record the appearance of each term in the corpus. Finally, we compared these results with the results of multiplying them by term frequency values related to each term in each text segment in the corpus.

In [11], we had been implemented number of classifiers with number of feature selection methods, comparing these accuracies resulted via number of corpus states. Different corpus states had generated from the same original corpus, after implementing NLP varied operations. We had performed the top accuracy results with the fifth corpus state and the feature selection method that based on LBSA approach, where they give each text segment a weight. This weight is based on the resulted weights from LBSA calculating, which count the number of sentiment lexicon terms appeared in each text segment, and produce the difference between positive words count and negative words count.

The feature selection in this work is different from the one used in [11]. It is actually improved from it, where in [11] we just satisfied with LBSA weights. Nevertheless, we here had take care of the occurrence of all corpus terms, utilizing the list of sentiment terms for give polarity of each term instead of text segments. Then, for each text segment, we had employed these polarities as FS vector.

In [12], seven feature selection methods had been compared, using SVM classifier via each one of these seven methods, those seven methods are: (Principal Components Analysis, Chi-squared, Relief-F, Gini Index, Support Vector Machines (SVMs), Uncertainty, and Information Gain). OCA corpus had been tested. In all SA classification experiments, they found out that SA performance is based on the method and the number of feature selection applied on it. Also, they found out that SVM is better than other method, whether as feature selection method or in classification accuracy.

They had deal with Arabic twitter for opinion mining in [13], by presenting a framework to analyze Arabic tweets into three sentimental classes: negative, positive, and/or neutral. Handling Arabizi, emoticons, and Arabic dialects was one of those sentiment framework aspects, which had done after collecting a dataset from twitter. They had used SVM with other classifiers, since they compare the results of stop-word eliminating and stemming of the corpus, using traditional methods of feature selection that based on counting terms appearing in each tweet on the corpus.
In [14], they had tried to discover useful knowledge about the opinions of Tunisian users from their posts on Facebook, and constructing an emoticons lexicon to be used in sentiment analysis tasks. SVM and NB had been utilized in SA tasks, via feature selection method based on traditional terms counting, where they had used popular terms appearance matrix, with the aid of Part-Of Speech POS, which is defined the type of each corpus term, based on grammar context.

3. METHODOLOGY

We have a hypothesis supposed that, in a number of the same domain text segments, each term has its own polarity, this polarity may be discovered by utilizing a list of sentiment keywords, and find the bigram relation between text terms and this list of keywords. We don't consider it as term polarity. We just used this value for feature selection method, which had led to enhance SA task.
In this work, we had increased the SA accuracy, using feature selection based on above hypothesis, which achieved results outperformed traditional methods in two opinion mining domains. The first domain is movie reviews, using OCA corpus collected by [4], since we used the same corpus and achieved more accuracy measure than traditional methods that applied in [4], we re-implemented it here for comparison purpose. The second domain is book reader reviews, using part of the huge corpus, i.e. LABR [6], since we re-implemented traditional method used in [3], and compared it with our approach which achieved better results.

Different literature had used different text pre-processing schemes; however, we determined the best pre-processing scheme, as a rule of thumb. This scheme is applied on each used corpus before any sentiment analysis had done, as illustrated in fig. 1. And we had noticed the efficient

benefit of this scheme, whether in feature set size reduction, which decrease computation time, or in the results of SA enhancement.

Our approach, generally, consisted of number of phases, illustrated in fig. 1, these phases are included pre-processing steps which had further details in the following subsection.

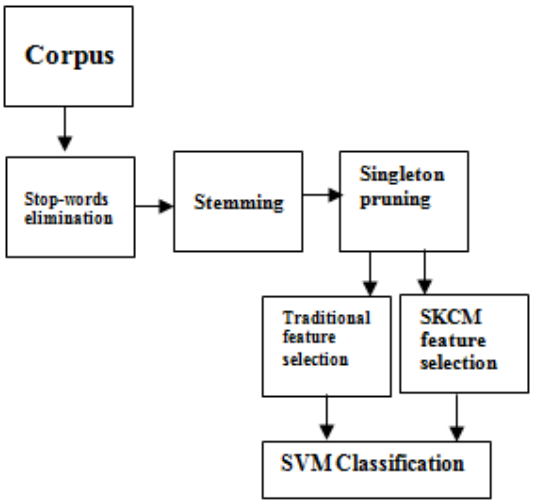

Figure 1. proposed approach general phases

3.1. Stop-words elimination:

Stop-words elimination is done without eliminating negation terms. Where stop-words are terms that known to do not have semantic affect on the written texts [3], specially, in the direction of opinion expressions. However, we preserved some words, with special exception; those are called negation terms, which are considered as sort of stop-words too. But they have an obvious impact on the polarity of terms that come after any one of them. Negation terms examples are terms like: 'no', 'but', 'more', 'less', etc, in English or opposite terms in any language, some negation words inverse the polarity of terms come after, whereas some of them affect the degree of the polarity.

3.2. Stemming:

We have done root stemming, where all unique terms, after stop-words elimination, are processed in premade tool. That converts each term into its origin: stem or root, stemming operation helps to reduce the number of terms. More precisely, those terms which have the same root. Stemming also helped to enhance feature selection, as in [4], because terms of the same root would have the same statistics, which improve semantic content on the feature selected, whereas most feature selection techniques on text mining depended on terms occur one the text.

We had utilized stemming tool that adopted by [15], the tool they had name: Alkhalil morpheme sys., which took the list of corpus unique terms as an input. Then, generated a list of terms, they were the roots of each input term. Then, we had made some refining to these root-stems before input them to SKCM algorithm.

3.3. Singleton terms pruning

Third pre-processing step was pruning corpus content. This was done by removing less impact terms in whole corpus. We choose to prune terms that occurred only once inside all corpus portions, pruning phase reduces the least significant terms in the corpus.

3.4. Traditional feature selection

The most popular method for text mining is using terms frequency TF as feature selection method, although, there were various ways to use TF, whether as one term frequency, i.e. unigram, or two terms as bigrams...etc. Whereas better TF results instance is TF-IDF, which had replaced TF mostly.

Both TF and TF-IDF are tested in [4], with unigram, bigram, and trigram variations. The results of them has supposed that no big differences between them via SA operations. This was compatible with our empirical experiments. Although, our objective herein is to proof the advantage of SKCM weights, comparing with TF weights traditional feature selection method.

3.5. SKCM algorithm

As illustrated in fig. 1, SKCM algorithm is implemented after the third pre-processing phase has finished, in parallel with the implementation of traditional FS method. SKCM algorithm is listed in fig. 2.

The outcome of SKCM algorithm had given each term a weight. This weight is different from the polarity values that had found out in semantic orientation (SO) as adopted in [8] or [9], since OS is used for constructing sentiment lexicons form seeds predefined polarity terms [9], or for determining text polarity as in [8], whereas we used this weight as feature selection method for SA. Moreover, we had used in SKCM distinct way to find terms weight, than mathematical way in [8]. Our approach is based on corpus terms discrimination using sentiment keywords co-occurrence measure (SKCM). This measure is used to find selected feature that depends on the number of corpus terms co-occurrence with sentiment keywords used for this objective.

```
SKCM algorithm
Input:
        crps; //target corpus reviews.
        utrms; // corpus unique terms.
        plst; // positive keywords list.
        nlst; // negative keywords list.
        tfm;  //terms frequency matrix.
        tdm;  // terms appearance matrix.
Output:
        skcm1;  // skcm with tdm matrix.
        skcm2;  // skcm with tfm matrix.
        tplrty; //terms polarity vector.
Begin
 nt = count corpus unique terms.
 nr = count corpus reviews.
 for i is 1 to nt
     neg=0;
     //negative co-occurrence with ith term.
     pos=0;
     //positive co-occurrence with ith term.
     for j is 1 to nr
       if uterms(i) is existed in crps(j)
         n = count nlst terms existed in crps(j).
```

```
        neg = neg + n.
        p = count plst terms existed in crps(j).
        pos = pos + p.
       end if;
     end for;
     tplrty(i) = pos – neg;
end for;
skcm1 = {replace each term's 1 over tdm, with tplrty value of the
          same term}.
skcm2 = {multiply each terms frequency over tfm, by tplrty value of
          the same term}.
END.
```

Figure 2. SKCM algorithm

Sentiment keywords are brought from premade sentiment lexicon, which we had improved in [2]. This lexicon is stemmed and used not to count lexicon words as in LBSA [9]. These keywords had extracted from a sentiment lexicon and stemmed to look for each keyword in each corpus portion. And then it gives each term in that corpus a counter. This counter is increased in both two polarities, i.e. positive and negative. Then for each corpus term, it have its own weight, which resulted by subtracting the summation of all target term co-occurrences with negative keywords, from the summation of all target term co-occurrences with positive keywords. Finally, as illustrated in fig. 2, SKCM algorithm output two selected features. The first resulted in pure terms weights, the second is the production of multiplying this results by term frequency that had been calculated in the phase of traditional FS method, as illustrated in our approach in fig. 1.

4. EXPERIMENTS AND RESULTS

We had used support vector machine (SVM) classifier, which had achieved good results in SA experiments. Also, it is a binary classifier that used to split each group of data into two categories, which is very suitable for SA tasks.

Three data sets are used. The first one is opinion corpus in Arabic (OCA) that collected by [4]. The second, two corpora are generated from the very big corpus large Arabic book reviews (LABR) [6], LABR consisted of 64000 reviews. Therefore, we, randomly, extracted two small corpora from it, a balanced and unbalanced one. UNBSLABR is unbalanced corpus that consisted of 2500 reviews, 1500 positive and 1000 negative. Whereas BLNSLABR is a balanced one which consisted of 1000 positive reviews and 1000 negative reviews, corpora statistics are available in table I.+

SA keywords list is based on SA lexicon that includes of positive bearing terms, and negative bearing terms, it was improved by [2]. This list terms are stemmed and used directly with each one of the three corpora.

Our experiments are begun with pre-processing steps, which applied to the three corpora. Then, we implemented SVM classifier using traditional selected features, which based on terms frequency (TF).

Table 1. Corpora statistics

Corpus	Reviews	Positive	Negative	Unique terms	Domain
unbSlabr	2500	1500	1000	14955	Book reviews
blnSlabr	2000	1000	1000	17801	Book reviews
OCA	500	250	250	42586	Movie reviews

Then, we applied SKCM algorithm, to assign each terms with a weight value that output from SKCM. After that, we had distributed each resulted value over text segment, based on the appearance of a term inside text segment, i.e. reviews text. These values will illustrate the relation between each term and text segment that belongs to. Then we carried on classification process using this selected feature.

Finally, for all corpora, we multiplied the selected feature of SKCM by TF values that used previously. This step was to measure and compare the effect of SKCM. If their results are better alone, or with traditional methods, feature selection based on terms frequency.

In all classification operations, we used SVM with cross validation k fold. We chose to use k=5 folds instead of 10 folds, due to computation complexity considerations.

Table II illustrates the results of our experiments. Where SKCM results are better than traditional method, i.e. term frequency depending selected features, even if we combined between SKCM results and TF, as illustrated in third column of table II, this achieved good results too. However, SKCM method results still better, when used solely.

Table 2. SVM accuracy results comparison.

	TF	SKCM	SKCM * TF
unbSlabr	74.20 %	**79.30 %**	78.00 %
blnSlabr	70.30 %	**73.25 %**	73.00 %
OCA	89.00 %	**93.00 %**	90.80 %

About the accuracy percentages, it is noticeable that movie reviews corpus made high number than book reviews two corpora. This may interpreted by that OCA has less reviews as illustrated in table I. And in the same time it is the bigger in the number of unique terms. Both two reasons helped in supervised learning, according to our practical experience. Also, another factor related to the nature of book readers reviews; this is noticed from [11] and [6], since LABR didn't achieved high accuracy results, which happened due to the language that used in that corpus, it was more complicated than another domain. And this is relevant to the richness of Arabic language morphology, which becomes as hard as movie reviews domain, perhaps due to used language, far away from dialectic or colloquial language of Arabic.

5. CONCLUSION

We had adopted new approach for enhancing sentiment analysis in Arabic language. Our approach based firstly on pre-processing steps, which reduces the size of dataset. Then we had implemented SKCM algorithm to apply our new feature selection method. After that we had used this new FS method on three corpora via SVM classifier.

Our approach results outperformed traditional methods, followed by most related works, which proof the significance impact of using the three pre-processing steps, which we had adopted in this work, and then illustrated the advantage of using sentiment keywords co-occurrence measure SKCM for discovering term weights that helped in enhancing the SA accuracy.

REFERENCES

[1] B. Liu, Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Morgan & Claypool, 2012.

[2] F. Alqasemi, A. Abdelwahab, and H. Abdelkader, "Adapting Domain-specific Sentiment Lexicon using New NLP-based Method in Arabic Language," International Journal of Computer Systems (IJCS), 2016, pp. 188-193.

[3] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," Collaboration Technologies and Systems (CTS), 2012 International Conference on. IEEE, 2012.

[4] M. Rushdi-Saleh, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and José M. Perea-Ortega, "OCA: Opinion Corpus for Arabic," Journal of the American Society for Infomration Science and Technology, 2011, pp. 2045–2054.

[5] H. Cho and S. Choi, "Automatic construction of movie domain Korean sentiment dictionary using online movie reviews," International Journal of Software Engineering and Its Applications 9.2, 2015, pp. 251-260.

[6] M. Nabil, M. Aly, and A. F. Atiya, "LABR: a large scale arabic book reviews dataset," CoRR, abs, 2014 .

[7] Q. A. Al-Radaideh and L. M. Twaiq,Rough, "Set theory for Arabic sentiment classification," Future Internet of Things and Cloud (FiCloud), 2014 International Conference on. IEEE, 2014.

[8] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[9] A Bai, H. Hammer, A. Yazidi, and P. Engelstad, "Constructing sentiment lexicons in Norwegian from a large text corpus," IEEE 17th International Conference on Computational Science and Engineering, 2014.

[10] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational linguistics, 2011, pp. 37(2), 267-307.

[11] F. Alqasemi, A. Abdelwahab, and H. Abdelkader, "An enhanced feature extraction technique for Improving se ntiment analysis in Arabic language," IEEE conference CSIST2016 Morocco, in press.

[12] N. Omar, M. Albared, T. Al-Moslmi, and A. Al-Shabi, "A Comparative Study of Feature Selection and Machine Learning Algorithms for Arabic Sentiment Classification," Asia Information Retrieval Symposium. Springer International Publishing, 2014.

[13] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment analysis in arabic tweets," Information and communication systems (icics), 2014 5th international conference on. IEEE, 2014.

[14] J. Akaichi, Z. Dhouioui, and M. J. Pérez, "Text mining facebook status updates for sentiment classification," System Theory, Control and Computing (ICSTCC), 2013 17th International Conference. IEEE, 2013.

[15] A. L. Boudlal, and et al, "Alkhalil Morpho Sys1: A Morphosyntactic analysis system for Arabic texts," In International Arab conference on information technology, 2010.

AUTHORS

Amira Abdelwahab, received BSc degree in computer science and information systems from Faculty of Computers and Information, Helwan University, Egypt in 2000. and Ph.D. in information systems from Chiba University, Japan in 2012. In 2013, she was a postdoctoral fellow in Chiba University, Japan. Since 2012, she has been an assistant professor in information systems department, Faculty of Computers and Information, Menofia University, Egypt. Her research interests include Software Engineering, Decision Support System, database Systems, Data Mining, Machine Learning, Recommendation Systems, Intelligent Web applications, e-commerce, and knowledge discovery in Big Data.

Fahd A. Alqasemi, received his Bsc. in Mathematical and Computer from Ibb university,Yemen, then received his Master of Computer Information Systems from Arabic Academy in Sana'a, Yemen. He worked as an instructor in UST, Sana'a, Yemen. Currently, he is pursuing PhD In Menofia University, Menofia, Egypt. His major fields of interest is Image Retrieval, Information Retrieval, Data and Text Mining.

Prof. Hatem Abdelkader, obtained his BS and M.SC., both in electrical engineering from the Alexandria University, faculty of Engineering, 1990 and 1995, respectively. He obtained his Ph.D in electrical engineering also from faulty of engineering, Alexandria University, Egypt 2001. His area of interest are data security, web applications, artificial intelligence, and he is specialized neural networks. He is currently a professor in the information system department, faculty of computers and information, Menofia University, Egypt.