Software Usability Evaluation Using Opinion Mining

Alaa M. El-Halees
Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine
alhalees@iugaza.edu.ps

***Abstract*—Usability is critical for any system, but in software it is one of the most important features. In fact, one of the main reasons for software failure is the system lacking to achieve users specified goals and satisfaction. For this reason, usability evaluation is becoming an important part of software development. Software usability evaluation can be costly in terms of time and human. Therefore, automation is promising way to augment existing approaches especially if the evaluation is subjective where the usability concentrated about user's "opinion". This paper proposes to use opinion mining as an automatic technique to evaluate subjective usability. Opinion mining is a research subtopic of data mining aiming to automatically obtain useful opinioned knowledge in subjective texts. We propose a novel model to extract knowledge from opinions to improve subjective software usability. This is the first time opinion mining used in software usability. To evaluate our proposed model, a set of experiments was designed and conducted and we got an average accuracy of 85.41%. Also, we propose to use graphics to visualize user's opinion in software and to compare the usability of two software.**

***Index Terms*—Software usability, usability evaluation, automatic evaluation, usability testing, opinion mining.**

I INTRODUCTION

Usability is critical for any system. It defined by [1] as " to what extent a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Usability is important in general, but in software it is one of the most important features. In fact, one of the main reasons for software failure is the system lacking to achieve users specified goals and satisfaction. For that reason usability evaluation is becoming an important part of software development. The purpose of software usability is to provide a feedback to improve design ; to asses that user and organizational objectives are being achieved; and to monitor long term use of product or system [2]. Software usability evaluation has been used to evaluate usability of many environments such as data mining systems [3], educational software in general [4] and Learning Management System in particular [5] , Mobile Systems [6] , Websites [7] [8] , Web mapping sites [9] and others.
There are many factors effect software usability, the most well-known focus on three aspects:, effectiveness (i.e. can users do what they want to do) , efficiency (i.e. how much effort do users require to do his/her task) and satisfaction (i.e. what do users think about the software ease of use) [10]. To evaluate these factors two broad approaches have been used objective and subjective. In objective evaluation they use metrics to evaluate the factors or dimensions of software quality [11]. The main disadvantage of this evaluation is that it cannot capture the complexity of user expression [11], and using measurement in usability is particularly difficult [12]. On the other hand; subjective evaluation measures participants' opinions or attitudes concerning their perception of software usability [12]. Subjective usability evaluation can be costly in terms of time and human, thus , automation is a promising way to enhance it [13]. Subjective usability focuses on user's "Opinion", therefore, this paper proposes to use opinion mining as an automatic technique to measure subjective usability.
Opinion mining is a research subtopic of data mining aiming to automatically obtain useful opinioned knowledge in subjective texts [14]. This technique has been widely used in real-world applications such as e-commerce, business-intelligence, information monitoring and public-opinion polls [15]. In this paper we propose to use it in software usability. We propose a model to extract knowledge from users' opinions to improve subjective software usability.
The rest of the paper is structured as follows: section two discusses related work, section three contains usability testing, section four is about opinion mining, section five describes the proposed usability model , section six gives the experiments and results and section seven concludes the paper.

II. RELATED WORK

To the best of our knowledge, there is no other research used opinion mining in software usability evaluation. However, the most related work was using data mining to enhance usability of web sites. For example, Marı et. al. in [11] presented an approach that empowers the qualitative usability testing process by extending it through data processing and data mining techniques such as association rules and decisions trees. Their goal was to obtain a general usability diagnosis of given context of use. The final output was a usability report include a list of prioritized usability problems. Jorge in [16] used data mining to enhance the quest for usability. They developed recommendation models that can indicate interesting links to a visitor of the web site

doi:10.4304/jsw.9.2.343-349

according to the preferences of similar users. A clustering methodology and visualization methods have been used . In addition, Tiedtke, et. al. in [17] proposed an automated website usability analyzer. It is a framework for automated website usability evaluation. The proposed framework is based on the combination of information architecture, automated usability evaluation and web mining techniques. Also, Li and Kit in [18] used a web structure mining algorithm which allows the automatic extraction of navigational structures in a website without performing hypertext analysis. They performed several usability experiments to correlate the usability of web sites and the structural design of the web site. Their experimental results showed that the structure mining algorithm gave reasonable prediction about several design issues in web structure.

III. Usability Testing

Software usability testing is a technique used to evaluate a software from users' perspective. It gives direct input on how real users exploit the software system [19]. In usability testing, the system is evaluated during which some data are collected, then an evaluator analyzes and interprets the results. In this research we concentrated on using automatic, remote unmoderated, subjective, testing method. The following sections gives an overall description of usability testing perspectives.

A. Usability Testing Methods

Carvalho in [4] categorized usability testing methods as: *Expert evaluation*, also known as , heuristic evaluation, which is done by experienced people who are asked to describe the potential troubles they expect for less experienced users. *Observational evaluation* involves collecting data that provide information about what users do when using software. *Surveys* help to know users' opinions or about an existing or potential software through the use of interviews or questionnaires. *Experimental evaluation* an evaluator can direct a number of factors associated with the usability and study their effect on user performance.

B. Ways of Usability Testing

Sauro in [20] classified ways of running usability tests to three which are: *Lab-Based*, where users physically come to a lab and are observed by a team of researchers. The main disadvantage of this method is the limited number of users can be used, the increase cost of evaluation and limitation of location. The second method is *remote moderated* where users log into screen sharing software. TimeZone difference is main drawback for international studies and cost is fair because the need of moderation. Third type is *remote unmoderated* where participants walk through tasks and paths are recorded.

C. Usability Testing Data Formats

Usability evaluations gather both subjective and objective data. Objective data are measures of participants' performance [12]. It is associated with the calculation of metrics that assess some software quality factors. On the other hand, subjective evaluation are measures of participants' attitudes concerning their opinion of usability. It measures user comments, interviews, or questionnaire responses. The subjective evaluation are better because it is easier, quicker and less expensive to obtain. It is often generates better insight, and is typically carried out through a process [12].

D. Manual and Automatic Evaluation Testing

The software evaluation can be done manually or automatically. Manually evaluation of usability can be expensive in terms of time and human. Automation is therefore a promising way to enhance existing approaches [21]. The automation of these activities has the following advantages: reduced cost of usability evaluation, reduced need for evaluation expertise among individual evaluators, increased coverage of evaluated features, enable comparisons between alternative designs, and easy incorporation of evaluation within the design phase of user interface development, as opposed to after implementation [21]. Automation of objective evaluation is easier than subjective evaluation because objective has metric data which can be a feature and tools in statistics and data mining can be used to automate the evaluation. Subjective data is more difficult because it contain unstructured text. In this paper we propose to use opinion mining methods to automatically evaluate the subjective data.

E. Usability Quality Factors

Using ISO 9241-11 Standard , there are three basic principles needed for defining usability which are: *Effectiveness* which measures of effectiveness relate the users' goals, accuracy and completeness with which these goals can be achieved [22]. *Efficiency* which measures of efficiency related the level of effectiveness achieved to the costs of resources. And, *Satisfaction* which measures the extent to which users are free from discomfort, and their attitudes towards the use of the product [22]. This research tries to measure these metrics automatically.

IV. Opinion Mining

Opinion mining is a subtask of text mining that automatically extract knowledge from the various user-generated contains such as product reviews, discussion forums and personal blogs. Opinion mining defined by Liu [23] as "Given a set of evaluative text documents D that contain opinions (or sentiments) about an object, opinion mining aims to extract attributes and components of the object that have been commented on in each document $d \in D$ and to determine whether the comments are positive or negative".

Opinion mining studies attitudes at three different levels: word level, sentence level and document level [24]. In this research we will concentrate at document level, where we consider user review as a document. In this case, systems assign positive or negative sentiment for a whole review [24]. Many approaches have been used in opinion mining the most common ones are lexicon based and machine learning. The lexicon-based

approach represents text as a bag-of-words. Opinion lexicons are resources that associate sentiment orientation and words. It considers a review as a collection of words without considering any of the relations between the individual words. In this approach positive opinion words are used to express desired states while negative opinion words are used to express undesired states [25]. The other approach is machine learning which uses classification methods to classify a document as positive or negative. This paper used an approach proposed by [26] that combined the lexicon based method and machine learning methods. It passes the document from lexicon based method to two classifiers, maximum entropy and k-nearest. The justification of that, using only one approach produces a poor performance. After applying lexicon based method, the classified documents are used as training set for machine learning methods. So, reviews first passed to lexicon-based classifier which classifies as much as possible. Then maximum entropy produces accurate results if they can classify the document, using another classifier, k-nearest, will classify the rest of the documents.

V. Proposed Usability Model

We propose a model in figure 1 to evaluate subjective usability using opinion mining approach. The model contains the following:

1) Group of users write their opinion for certain software in the three usability factors: Effectiveness, Efficiency and Satisfaction. The users do not need to be in lab, and no moderation necessary.

2) The preprocessing step including: stripping out the HTML tags and non-textual contents. Then, separating the documents into review and converted each review into a single file. Then, some of the wrong spelling words are corrected. After that, the sentences are tokenized, stop words removed and light stemmer applied. We also removed some terms with a low frequency of occurrence. Then, we obtained vector representations for the terms from their textual representations by performing TFIDF (Term Frequency–Inverse Document Frequency) which is a well known weight presentation of terms used in text mining [27].

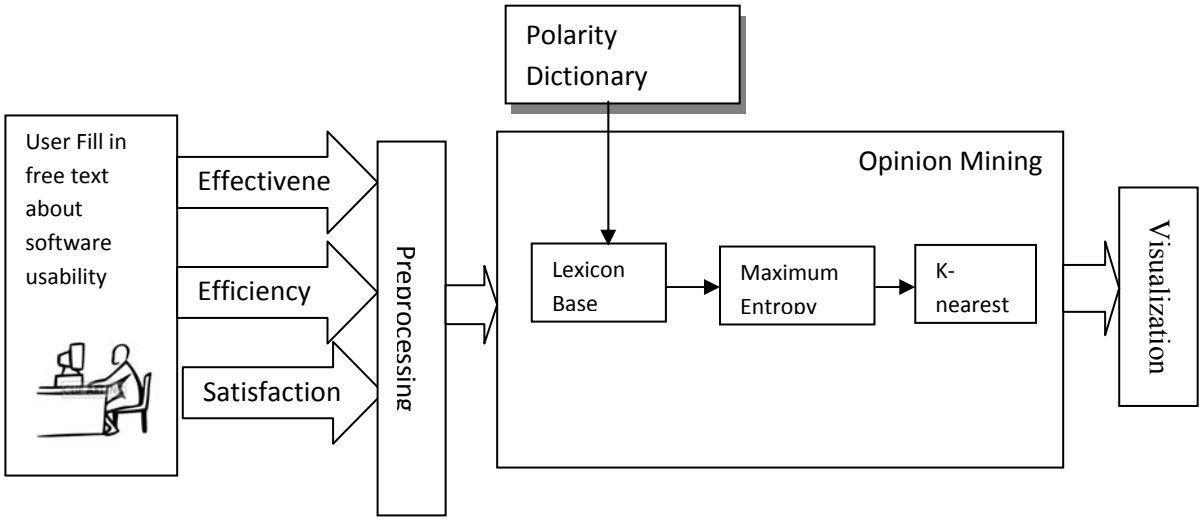

Figure 1: A model to evaluate subjective usability using opinion mining approach

.

3) Opinion mining system which has three parts:

a) Lexicon based opinion classifier which uses opinioned words and phrases to determine the sentiment orientation of the whole review. It tries to find out the words or phrases that indicate the sentiment and determine the orientation of their sentiment (i.e. positive or negative), then classify the sentence. After that it can classify the whole document. To do that, it uses a dictionary of positive and negative words (e.g., love, hate) [28,29]. This step works as follows: It takes unannotated reviews (to be classified), identify all opinion words and phrases (using negations when needed). Then aggregate these words to give a sentiment (positive or negative) to the review. However, some documents do not appoint to any sentiment polarity which is the documents that do not have enough clear opinioned words.

b) The next step is to use maximum entropy classifier. The maximum entropy model estimates probabilities based on the principle of making as few assumption as possible, other than the constrained

imposed. The constraints are consequent from training process which articulate a relationship between the binary features and the outcome [30,31].

The reviews that have been classified from the previous step will be used as training set for the classifier. The goal in this step is to classify as much reviews as possible that remain form the previous step. The unannotated data set is given to the maximum entropy probability systems. Given certain threshold (we used 0.75) if the sentiment greater than this probability it will be classified, if not it will be unclassified review which will pass be to the next step

c) The next step is to use k-nearest neighbor (kNN) which is a simple method to classify document [32]. In this method, given an unannotated review *r*, the system finds the k nearest neighbors among training reviews which are classified in the previous two phases. The similarity score of each nearest neighbor review to the test review is used as the weight of the classes of the neighbor review *r* belongs to the sentiment *s* that has the highest score.

4) Visualization step which visualizes the three factors of software usability in a graph that shows the negative and positive opinions of a certain software usability. Then, given two reviews, the system compares their usability graphically.

VI. Experiments

To evaluate our model, a system was built and a set of experiments was designed and conducted. In this section, we describe the experiments design including the corpus, the classification tools, evaluation metrics and results**.**

A. Corpus

To test our model, we collected usability reviews from different users for four different software. As depicted in table 1, we used total of 565 reviews contain 345 positive reviews and 220 negative reviews.

TABLE 1
DESCRIPTION OF CORPUS USED IN THE EXPERIMENTS

Number of reviews		
Negative	**Positive**	
53	**95**	**Software 1**
82	**110**	**Software 2**
55	**85**	**Software 3**
30	**55**	**Software 4**
220	**345**	**Total**

B. Classification Tools

For preprocessing steps, we used Rapidminer from [33] for tokenization, removing stopwords, light stemming and vector presentation. For opinion mining methods we used Polarity Dictionary from lexicon based opinion classifier project SentiStrength [30]. SentiStrength employs several methods to extract positive and negative sentiment strength from short text. It uses a dictionary of sentiment words with associated strength measures. For maximum entropy method we used maxent software from [34]. For k-nearest we, also, used Rapidminer as data mining tool to classify and evaluate the results.

C. Evaluation Metrics

There are various methods to evaluate the system built based on the model; however, accuracy, precision and recall are the most common in this field. Accuracy measures the percentage of the test set that the classifier has labeled correctly. Furthermore, the precision and recall are calculated. Precision is the percentage of predicted documents class that are correctly classified. Recall is the percentage of the total documents for the given class that are correctly classified. We also computed the F-measure, a combined metric that takes both precision and recall into consideration [35].

D. Results

The input of the implemented system is a set of subjective reviews from users who evaluated the usability factors of four software. First, we measure the accuracy of the system on evaluating positive and negative opinions of the users**.** Evaluation of opinion classification relies on comparison of results on the same corpus annotated by humans [36]. Therefore, to evaluate our system, first we manually assigned a label for each user subjective opinion. Then**,** we evaluated the accuracy of the data sets using proposed classifier. Table 2 gives the accuracy of the four software in each factor.

TABLE 2
ACCURACY OF THE USABILITY OF FOUR SOFTWARE USING PROPOSED MODEL

Avg	**Satisfaction**		**Efficiency**		**Effectiveness**		
	Neg	Pos	Neg	Pos	Neg	Pos	
85.05	87.56	92.32	81.28	83.62	79.21	86.3	**Soft1**
83.86	85.25	89.25	82.55	85.45	78.62	82.05	**Soft2**
90.92	92.98	95.36	86.63	89.25	88.78	92.56	**Soft3**
81.82	83.27	86.89	81.26	83.28	77.02	79.25	**Soft4**
85.41	89.11		84.17		82.97		**Avg**

From the table we can conclude that using the proposed model we got an average accuracy of about 85%, which is an acceptable result given that the review is subjective. Also, we can notice that *effectiveness* is the most difficult to detect. That is because the language of effectiveness is more general that other two factors. On the other hand, *satisfaction* is the easiest to detect. In

addition, in average positive reviews have better accuracy than negative ones. That is because usually negation gives more complication to the statements.
In addition to accuracy measure, we used measure of recall, precision and f-measure.
Table 3 gives the measures of recall, precision and f-measure for the four software.

TABLE 3
RECALL, PRECISION AND F-MEASURE FOR THE FOUR SOFTWARE.

F-measure	Precision	Recall	
88.76	87.36	90.21	**Software 1**
86.23	83.36	89.32	**Software 2**
94.66	94.11	95.23	**Software 3**
86.20	90.90	81.96	**Software 4**
89.05	**88.93**	**89.18**	**Average**

In general, the results is acceptable in recall and precision, hence f-measure.

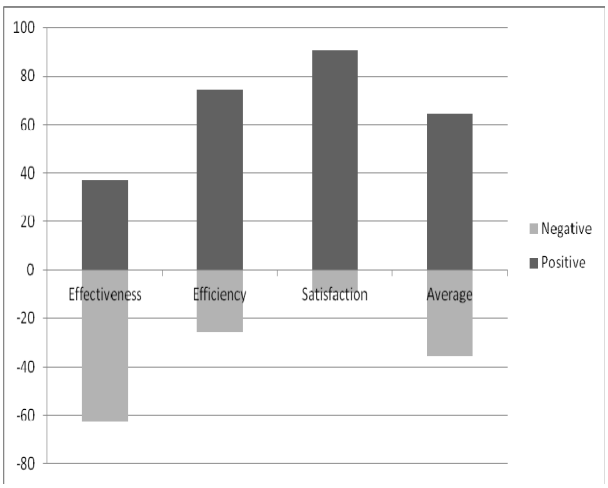

Figure 2: Visualizing effectiveness, efficiency and satisfaction of a software

Also, we propose to visualize the results using graphs. Figure 2 depicted the positive and negative opinions of users about *software1*. As we can see in the figure , it is easy to envisage the positive and negative opinions for each of the three factors. For example, we can figure out that *Effectiveness* has negative attitude while *Satisfaction* has positive attitude from the point of view of the users who review of the software. Also, we can noticed that the average usability of the software is more positive than negative.

In addition, using the visualization, we can compare the users' opinion of two software usability, for example figure 3 compares the user opinion in usability of *software1* and *software3*.

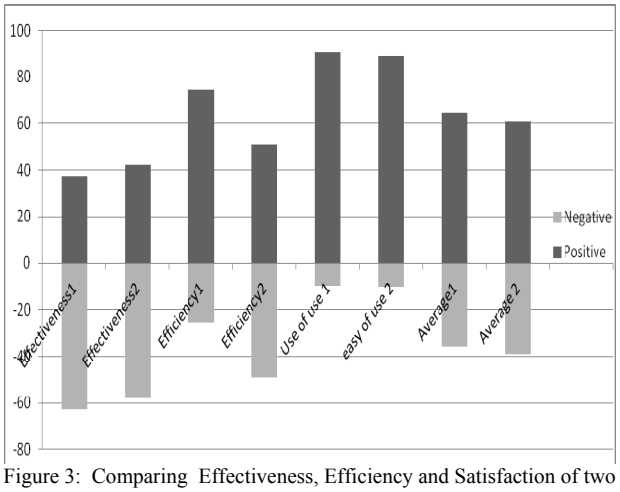

Figure 3: Comparing Effectiveness, Efficiency and Satisfaction of two software.

We can noticed that the two software have similar *effectiveness* and *satisfaction*, but the *efficiency* of *software1* is better than *software3*.

VII. CONCLUSION

In this paper, we have proposed a model which aims at using opinion mining to automatically evaluate software subjective usability. The model concentrated on three usability quality factors: effectiveness, efficiency and satisfaction

The model used three methods at sequence: First, lexicon based method is used which classifies some reviews. The classified documents used as training set for maximum entropy model which subsequently classified some other reviews. After that, k-nearest method is used to classify the rest of the reviews.

In experiments we applied the model to four software usability reviews contained 345 positive reviews and 220 negative reviews**.** Our system achieved an accuracy of about 85%. The experimental results further showed that recall, precision and f-measure of the evaluated reviews are acceptable. The model also proposed to visualize a subjective usability of any software and compare the usability of two software visually.

In the future work we may use different and more general software usability factors. Also, it may be able to automatically classify the type of factors from subjective user opinion.

REFERENCES

[1] ISO 9241-11. *Ergonomic requirements*. Part 11: Guidance on usability, 1998
[2] ISO 13407. *Human centered design processes for interactive systems*, 1999.
[3] Galiano FB, Cubero JC, Marín N, Serrano JM and Blanco IJ. *Usability Issues in Data Mining Systems.* Proceedings of the 5th International Conference on Enterprise Information Systems, Angers, France, April 22-26, 2003.
[4] Carvalho, A. (2002) 'Usability testing of educational software methods, techniques and evaluators', 3 Simposio Internacional de Informática Educativa, Portugal.
[5] Blecken A, Bruggemann D and Marx W. *Usability Evaluation of a Learning Management System.* In the

Proceeding of the 2010 43rd Hawaii International Conference on System Sciences. IEEE Computer Society Washington, DC, USA 2010.

[6] Beck E, Christiansen M , Kjeldskov J, Kolbe, N. and Stage J. *Experimental Evaluation of Techniques for Usability Testing of Mobile Systems in a Laboratory Setting.* In the Proceedings of OzCHI, Brisbane, Australia. CHISIG. 2003.

[7] Tobar, L.M., Latorre Andrés, P.M. & Lafuente Lapena, E., 2008. *WebA: a tool for the assistance in design and evaluation of websites.* Journal of Universal Computer Science, pp.1496-512.

[8] Fernandez A, Insfran E and Abrahão S. *Usability evaluation methods for the web: A systematic mapping study.* Information and Software Technology Volume 53, Issue 8, Pages 789–817. August 2011.

[9] Nivala AM, Brewster S and Sarjakoski T. *Usability Evaluation of Web Mapping Sit* . The Cartographic Journal, Volume 45, Number 2, pp. 129-138. May 2008.

[10] Bevan N. *What is usability?* UsabilityNet 2006. http://www.usabilitynet.org.

[11] González MP, Lorés, J and Granollers A. *Enhancing usability testing through datamining techniques: A novel approach to detecting usability problem patterns for a context of use.* Information and Software Technology, Volume 50, issue 6, , p. 547-568. May, 2008.

[12] Lewis JR. *IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use*. International Journal of Human-Computer Interaction, 7:1, 57-7, 1995.

[13] Kostaras N, Stavrinoudis D , Sokoli S and Xenos M . *Combining experimental and inquiry methods in software usability evaluation*: *The paradigm of LvS educational software*, Journal of Systems and Information Technology, Vol. 12 Iss: 2, pp.120 – 139. 2010.

[14] Harb A and Dray MP. *Web opinion mining: how to extract opinions from blogs?*. In the Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology. ACM New York, NY, USA. 2008

[15] Ding , Liu B and Zhang L. *Entity Discovery and Assignment for Opinion Mining Applications.* In the Proceeding KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. Pages 1125-1134 .ACM New York, NY, USA. 2009.

[16] Alípio J. *Web site usability and data mining.* Sol-EU-Net project. Data Mining and Decision Support. 2004.

[17] Tiedtke T, Märtin C , Gerth N . *AWUSA – A Tool for Automated Website Usability Analysis*. In the Proceedings of the 9th International Workshop on Interactive Systems. Design, Specification, and Verification. Rostock, Germany June 12-14, 2002.

[18] Li CH, Kit CC. *Web structure mining for usability analysis*. In the Proceedings of International Conference on IEEE/WIC/ACM 17 October 2005.

[19] Lodhi, A. *Usability Heuristics as an Assessment Parameter: for performing Usability Testing.* In the proceedings of 2nd International Conference on Software Technology and Engineering (ICSTE) Islamabad, Pakistan. 3-5 Oct. 2010.

[20] Sauro J. Measuring Usability. *Comparison of Usability Testing* Methodshttp://www.measuringusability.com/blog/method-comparison.php

[21] Ivory MY and Hearst M . *The State of the Art in Automating Usability Evaluation of User Interfaces.* Journal of ACM Computing Surveys (CSUR). Volume 33 Issue 4, Pages 470-516. December 2001.

[22] ISO 9126-11: *Software engineering – Product quality* – Part 4: Quality in use metrics 2004.

[23] Liu B, Hu M. and Cheng J. *Opinion Observer: Analyzing and Comparing Opinions on the Web.* In the Proceedings of International World Wide Web Conference .2005.

[24] Pang B, Lee L, and Vaithyanathan S. *Thumbs up? Sentiment Classification using Machine Learning Techniques*. In the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79--86, 2002

[25] Ohana, B and Tierney B. *Sentiment classification of reviews using SentiWordNet*. In the 9th. IT&T Conference, Dublin Institute of Technology, Dublin, Ireland, 22nd.-23rd. October, 2009.

[26] El-Halees AM. "Arabic Opinion Mining Using Combined Classification Approach". In the proceeding of: 2011 International Arab Conference on Information Technology ACIT'2011 Riyadh, 2011.

[27] Salton G and , Buckley C. *Term-weighting approaches in automatic text retrieval* . Information Processing & Management 24 (5): 513–523. 1988.

[28] Liu B. Opinion Mining Encyclopedia of Database Systems, 2008.

[29] Hu M, and Liu B. *Mining and Summarizing Customer Reviews*. In the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA, 2004

[30] Thelwall M, Buckley K, Paltoglou GA and Cai D. *Sentiment strength Detection in short informal text* .Journal of the American Society for Information Science and Technology vol. 61. issue 12. 2010.

[31] Nigam K, Lafferty J and A McCallum A. *Using Maximum Entropy for Text Classification*. In the proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67. 1999. [32] Dasarathy B. *Nearest neighbor (NN) norms: NN pattern classification techniques* . IEEE Computer Society Press, 1991.

[32] Mierswa I, Wurst, M, Klinkenbe, R, Scholz M and Euler T. *YALE: Rapid Prototyping for Complex Data Mining Tasks.* In the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006

[33] Phillips S J, Dudik M and Schapire RE. *A maximum entropy approach to species distribution modeling*. In the Proceedings of the 21st International Conference on Machine Learning. ACM Press, New York 2004.

[34] Makhoul J, Kubala F, Schwartz R and Weischedel R. *Performance measures for information extraction*. In the Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999.

[35] Osman D and Yearwood J. *Opinion search in web logs*. In the Proceedings of the Eighteenth Conference on Australasian Database, 63. Ballarat, Victoria, Australia.

Alaa El-Halees is an associate professor in computer Science and Deputy Dean of Research Affairs and graduate studies of the faculty of Information Technology Department at Islamic University of Gaza, Palestine. He holds PhD degree in data mining in 2004, MSc degree in software development in 1998 from Leeds Metropolitan University, UK. He received his BSc degree in computer engineering in 1989 from University of Arizona, USA. His research activities are in the area of data mining, in particular text mining and opinion mining, machine learning and e-learning.