Applying Information Extraction and Fuzzy Sets for Opinion Mining

Shaidah Jusoh, and Hejab M. Alfawareh

***Abstract*—** Opinions are always expressed in comments or reviews. An automated opinion mining system has been seen as one of the desirable intelligence business tools. The system can extract public opinion about a certain topic, product or service which is embedded in unstructured texts. Extracting opinions from reviews and comments requires a system to deal with natural language texts. The current focus in opinion mining research is to classify opinion into three categories; positive, neutral, and negative. Classifying opinion which is presented in a phrase still remain a challenge to researchers in this area. This paper has introduced an approach for classifying opinion which is presented in a phrase into two categories; positive and neutral. The approach is obtained by applying information extraction technique and fuzzy sets to the texts which contain opinion.

***Keywords*—** Fuzzy Sets, Information Extraction, Opinion Mining.

I. Introduction

An automated opinion mining system is a sentiment analysis tool. It is a new computing technology which may enhance a decision making process. The tool can be used to extract public opinion which is represented in unstructured texts. Public opinion is a piece of important information for various kinds of areas. These include, marketing, politics, economics, and so on. Opinion mining can be useful in many ways. for example if we are in education, it can help the school or college or university to determine the current mood of students. For example, an academic system can identify which demographics like or dislike with any changes being made at the school or college or university. If we are in marketing, for example, it can help us to judge public whether they like or not with products being introduced to them. Being able to recognize this kind of information in a systematic ways gives us a better picture of public opinion.

Thus the research field of opinion mining and sentiment analysis is an important field and well suited to various types of intelligence applications. Opinion mining is a type of natural language processing for tracking the mood of the public about a particular issue, product, or service . Opinion mining is involved with building a system to collect and examine opinions usually in comments or reviews. Recent research on automated opinion mining has focused on sentiment analysis into positive or negative sentiment. Research work on opinion mining has reported on utilizing natural Language Processing (NLP) and machine learning techniques in extracting sentiment (opinion) is indeed a complex task. There are a number of existing sentiment tools are available at this moment, unfortunately none of them is able to give an accurate reflection of public. In opinion mining, one of the fundamental problem is to recognize whether a text expresses a positive, negative or neutral sentiment. In the work of [1] have shown a technique on finding the degree of positive and negative sentiments by evaluating one sentiment word, such as "*beautiful*", "*excellent*". However, not all the time an opinion is expressed in one word. Sometimes an opinion is expressed in a phrase, such as "*somehow beautiful*", "*very pretty*" or "*not beautiful*". For example, "*not beautiful*" is evaluated as negative opinion, "very beautiful" is evaluated as positive opinion and "somehow beautiful" is evaluated as neutral opinion.

The purpose of this paper is to introduce a new approach for evaluating sentiments which are presented in a phrase either it is positive or neutral. Information extraction (IE) technique and fuzzy approach have been adopted in solving the problem. This paper is organized as follows. Section II presents a brief overview on previous related areas, namely opinion mining, information extraction and fuzzy sets. Section III presents the proposed approach. Section IV presents a summary of the paper.

II. Related Areas

A. Opinion Mining

Mining opinion research work has been conducted on various platforms. These include, extracting opinion from messenger [2], online forums where consumers exchange product opinions [3], social networks [4], Weblogs [5] and many more. The review work of [6] has revealed the great potential of opinion mining methods for the analysis of textual citizens' contributions in public policy debates. Research work of [1], [7] also demonstrated opinions of customers reviews on the Web. Various techniques have been applied for opinion mining. For example, work of [8] had applied sentiment

Shaidah Jusoh is an associate professor of computer science at the College of Computer Science & Information Systems at Najran University, Saudi Arabia.(email: shaidah.jusoh@gmail.com /sbjusoh@nu.edu.sa)

Hejab Alfawareh is an assistant professor at the College of Computer Science & Information Systems at Najran University, Saudi Arabia. (email: alfawareh@gmail.com /hmalfawareh@nu.edu.sa)

analysis to online movies reviews. Their research finding indicated that machine learning techniques, specifically support machines are applicable at detecting sentiment in movie reviews. Techniques used for opinion mining include Information extraction [9], machine learning, graph [10], reinforcement approach [11], lexicon [12], [13] , a linguistic approach [13].

B. Information Extraction

IE is an enabling technology which allows an intelligent system for retrieving valuable information and knowledge from free text to be developed. Basically, IE is a process of extracting useful information from the text and storing the information in a structured database. Then a machine learning approach can be applied to the structured data for discovering new knowledge [15]. IE task is defined by its input and its extraction target. The input can be unstructured documents like free text that are written in natural language or the semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists. Programs that perform the task of IE are referred to as extractors or wrappers . The first step in most IE tasks is to detect and classify all the proper names mentioned in a text; a task generally referred to as named entity recognition (NER). Reference defined entity as anything that can be referred to with a proper name. A process of NER refers to the combined task of finding spans of text that constitute proper names and then classifying the entities referred to according to their type. The IE tasks aim at finding specific data in natural language texts. With IE approach, events, facts and entities are extracted before the knowledge mining process is conducted [16]. Consequently IE allows for mining the actual information presented in the texts, rather than the limited set of tags associated to the documents Unlike information retrieval (IR), which concerns how to identify relevant documents from a document collection, IE produces structured data ready for post-processing, which is crucial to many text mining applications, including opinion mining.

C. Fuzzy Sets

A fuzzy set is a set without a crisp, clearly defined boundary. It can contain elements with only a partial degree of membership. The root of fuzzy set theory lies at the idea of linguistic variables. A linguistic variable x is characterized by a term-set which contains a set of names of x. The meaning of each name is given by a fuzzy set, characterized by a membership function. The membership function itself is context dependent. Let us consider the proposition “the old book is cheap”. Here we immediately identify cheap as a member of the term set of the linguistic variable PRICE. For instance,

$$T(PRICE) = \{cheap, moderate, expensive\}$$

where the fuzzy values cheap, moderate, expensive are characterized by fuzzy sets in a certain universe of discourse U. The concept of linguistic variables makes it possible to store and reflect such information into a computer. The fuzzy values of the linguistic variable PRICE can be freely interpreted depending on context. Surely the price of a second hand hardcover book differs from a secondhand paperback cover. Nevertheless, humans have the ability to adjust the meaning of a sentence context dependent. A linguistic variable is a fuzzy variable which can be translated into a fuzzy set as below

$$F_p = \{0.3, 0.6, 0.9\}$$

where 0.3 denotes *cheap*, 0.6 denotes *moderate* and 0.9 denotes *expensive*. By storing linguistic variables into numerical variables, the fuzzy set of *price* can be calculated.

III. PROPOSED APPROACH

In this approach, a review will be an input text. A review may consist of a word, a sentence, or a paragraph, however the evaluation is made at a sentence level. Previous work of [1] has solved the problem for one word using fuzzy sets. This work focuses on a phrase. It is common to find a reviewer to give his/her opinion in a phrase. For example, "*I found that the service is very helpful*", or " *My new IPhone is somehow easy to use*". The proposed approach will evaluate not only a sentiment word such as *easy*, but will also consider the modifying word which comes before the sentiment word. Using the example above, the word *somehow* modify the total semantic of the word *easy*. In the first example, the opinion about the service should be taken as positive while in the second example, the opinion about the IPhone should be taken as neutral.

A. Framework

In this approach there are two main components; opinion extractor and opinion fuzzy sets. Opinion fuzzy sets contains only two types of linguistic variables. The first variable is a word which can represent either a positive sentiment or a negative sentiment. The word which represents sentiment is defined as SenWord [1]. For an example, *beautiful* is normally used to represent a positive sentiment while *ugly* is normally used to describe a negative sentiment. The second variable is a word that is used to modify the total semantic of SenWord, which is defined as *SenWord Modifier*. For example, if *somehow* comes before the word *beautiful*, the total semantic of beautiful has changed. The structure of opinion sets is

$$F(opinion) = \{SenWord, SenWord\ Modifier\}$$

B. Implementation

Step 1: Sentence Tokenization

Sentence tokenization is conducted if a review is in a form of a sentence or a paragraph. Tokenizing a sentence is a process of breaking a sentence into a list of words. In other words, a tokenizer parses a sentence into a list of tokens

(words). An output of sentence tokenization process will be stored in a dynamic list. Technically it is stored as an array. Handcrafted rule is used to determine a word which has two possible parts of speech. If a word can be tagged by more than one, part of speech, handcrafted rules are used to determine the most possible part of speech. For example the word 'place' can be assigned as a verb as well as a noun. The rule 'if the preceding word is an article, then the word is a noun, otherwise it is a verb', is used. In this work, we have adopted a shallow parsing technique.

Step 2: Identify and Extract SenWord

In identifying a SenWord, in a review/comment, a lexicon of SenWord of positive sentiment and a lexicon of negative sentiment are developed. Each word in the list of tokens is compared to the word exists in both lexicon. If they are match, a SenWord has been recognized and the word in the dedicated word in the list of tokens is labeled as SenWord.

Each SenWord in the lexicon is attached with the value of positive degree or negative values. Table 1 below illustrates the positive SenWord with its values. The values indicate the degree level of positive or negative SenWord. Human common sense and linguistic knowledge have been used in deciding the degree values.

TABLE I
EXAMPLE OF SENWORD LEXICON WITH ITS POSITIVE VALUE

SenWord	Positive Value
Excellent	0.9
Great	0.9
Wonderful	0.8
Good	0.7
Beautiful	0.8

***S*tep 3: Identify and Extract SenWord Modifier**

In this work, a lexicon of SenWord Modifier is also developed. Adverbs such as *somehow*, *very* , *quite*, are considered SenWord Modifier. Once a SenWord is recognized, the Opinion Extractor will identify the SenWord Modifier. A dynamic list of SenWord and its modifier is developed which denote a fuzzy set of an opinion in the form of linguistic variables. Let us consider a sentence such as "*the offered program is somehow excellent*". The word *excellent* is taken as a SenWord, and the word *somehow* is taken as SenWord Modifier

$$F_{opinon} = \{excellent, somehow\} \quad (1)$$

Step 4:Assign Fuzzy Values to Fuzzy Sets Opinion

Once Opinion Fuzzy Sets is constructed, a fuzzy linguistic variable is converted into fuzzy values. The value of SenWord is depending on the value which is stored in the SenWord lexicon. The value of SenWord Modifier is depending on the value which is stored in the SenWord Modifier lexicon. Linguistic variables in opinion fuzzy sets in Equation 1 is replaced with numerical values as shown in Equation (2).

$$F_{opinon} = \{0.9, 0.5\} \quad (2)$$

Step 5: Fuzzy Sets Operation

Fuzzy set operation is conducted on the opinion fuzzy sets, to determine types of a sentiment either it is positive or neutral. In this approach, fuzzy *min* operator is used to decide the category of sentiments. Using the min fuzzy operator, the fuzzy set in Equation 2 is calculated as 0.5

$$F_{opinon} = \min\{0.9, 0.5\} \quad (3)$$

Step 6: Finalize Type of Sentiments

In this work, we have determined that if the value after the fuzzy sets operation is in the range from 0.4 to 0.6, then opinion is considered as *neutral*, and the value in the range of 0.6 to 0.9 represents a *positive* sentiment.

IV. SUMMARY

This paper has presented a novel approach for mining opinions from reviews which are presented in phrases. The approach is conducted at a sentence level. Information extraction technique has been applied to extract words that are representing a sentiment in a sentence. Fuzzy sets approach has been applied to determine the type of sentiment either it is positive or neutral. Negative sentiment has not been considered in this work. Thus the future work will evaluate the negative sentiment in the review.

ACKNOWLEDGMENT

This research is supported by Najran University, Saudi Arabia under the research grant entitled *Opinion Analysis from Reviews and Comments,* with code No. (NU ESCI/13/23).

REFERENCES

[1] Shaidah Jusoh and Hejab M.Alfawareh,in *Proceeding of the International Conference in Computer Applications Technology (ICCAT)*, 2013, pp 1-5

[2] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 168–177.

[3] J. S. Ryu, W. Y. Kim, K. I. Kim, and U. M. Kim, "Mining opinions from messenger," in *Proceedings of the 2nd International Conference on Interaction Sciences*: Information Technology, Culture and Human, ICIS '09. New York, NY, USA: ACM, 2009, pp. 287–290.

[4] C. Kaiser and F. Bodendorf, "Opinion and relationship mining in online forums,"in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* - Volume 01, ser. WI-IAT '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 128–131.

[5] K. S. Cho, J. Y. Yoon, I. J. Kim, J. Y. Lim, S. K. Kim, and U.-M. Kim, "Mining information of anonymous user on a social network service," in *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, ser. ASONAM '11*. Washington, DC, USA: IEEE Computer Society, 2011, pp. 450–453.

[6] J. G. Conrad and F. Schilder, "Opinion mining in legal blogs," in *Proceedings of the 11th international conference on Artificial intelligence and law, ser. ICAIL '07*. New York, NY, USA: ACM, 2007, pp. 231–236.

[7] M. Maragoudakis, E. Loukis, and Y. Charalabidis, “A review of opinion mining methods for analyzing citizens’ contributions in public policy debate,” in *Proceedings of the Third IFIP WG 8.5 international conference on Electronic participation*, ser. ePart’11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 298–313.

[8] X. Ding, B. Liu, and P. S. Yu, “A holistic lexicon-based approach to opinion mining,” in *Proceedings of the international conference on Web search and web data mining*, ser. WSDM ’08. New York, NY, USA: ACM, 2008, pp. 231–240.

[9] D. R. Swanson, N. R. Smalheiser, and A. Bookstein, “Information discovery from complementary literatures: categorizing viruses as potential weapons. journal of the american society for information science,” *Journal of the American Society for Information Science*, vol. 52, no. 10, pp. 797–812, 2001.
http://dx.doi.org/10.1002/asi.1135.abs

[10] F. L. Cruz, J. A. Troyano, F. Enr´ıquez, F. J. Ortega, and C. G. Vallejo, “A knowledge-rich approach to feature-based opinion extraction from product reviews,” in *Proceedings of the 2nd international workshop on Search and mining user-generated content*s, ser. SMUC ’10. New York, NY, USA: ACM, 2010, pp. 13–20.

[11] G. Berend and R. Farkas, “Opinion mining in hungarian based on textual and graphical clues,” in *Proceedings of the 8th conference on Simulation, modelling and optimization*. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2008, pp. 408–412.

[12] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su, “Hidden sentiment association in chinese web opinion mining,” in *Proceedings of the 17th international conference on World Wide Web*, WWW ’08. New York, NY, USA: ACM, 2008, pp. 959–968.

[13] O. Vechtomova, “Facet-based opinion retrieval from blogs,” *Inf. Process. Manage*., vol. 46, pp. 71–88, January 2010.
http://dx.doi.org/10.1016/j.ipm.2009.06.005

[14] T. T. Thet, J.-C. Na, and C. S. Khoo, “Aspect-based sentiment analysis of movie reviews on discussion boards,” *J. Inf. Sci.*, vol. 36, pp. 823–848, December 2010.
http://dx.doi.org/10.1177/0165551510388123

[15] Shaidah Jusoh and Hejab M. Alfawareh, Techniques, Applications and Challenging Issue in Text Mining, IJCSI, vol. 9, Issue 6, no 2, pp. 431-436

[16] R. Hale, “Text mining: Getting more value from literature resources,” *Drug Discovery Today*, vol. 10, no. 6, pp. 377–379, 2005.
http://dx.doi.org/10.1016/S1359-6446(05)03409-4

Dr. Shaidah Jusoh is an associate professor of Computer Science at College of Computer Science & Information System, Najran University,Saudi Arabia. She completed her PhD in Engineering (Engineering System and Computing, December 2005) from University of Guelph, Canada. Her PhD research is in the area of intelligent systems. She also received Master of Science in Computer Science with specialization in Distributed Information System from the University of Guelph, Canada, and Bachelor of Information Technology (with Honors) from Universiti Utara Malaysia, Malaysia. Previously she worked at Zarqa University, Jordan, Taibah University, Saudi Arabia, Universiti Utara Malaysia, Malaysia and University of Guelph, Canada. Dr. Shaidah has been an active member of editorial board committees, reviewer committees and technical program committees of international journals and proceedings. She has published numerous number of publications in international refereed journals and high quality international conference proceedings. She has awarded a various number of research grants and has completed 3 PhD students.

Hejab M. Alfawareh is currently an assistant professor at the the of College of Computer Science & Information Technology, Najran University, Saudi Arabia. Dr. Hejab Al Fawareh obtained his PhD in Information Technology with specialization in Artificial Intelligence from Universiti Utara Malaysia in 2009/2010. Dr. Hejab has taught 11 different courses in various university namely, Zarqa University, Al-Albayt University (both in Jordan), Taibah University and Saudi Arabia He is also actively involved with various committees at Zarqa University. His research interests include information extraction, ambiguity resolution, social networks, text mining and fuzzy applications. He has presented his research work in France, Malaysia, Philippines, Ukraine, and Jordan. He has published papers in journal, book chapters, and international proceeding.