Analyzing Web user' Opinion from Phrases and Emoticons

Anil Kumar K.M
Research Scholar
DOS in Computer Science
University of Mysore
Mysore, Karnataka, India

Suresha
Research Supervisor
DOS in Computer Science
University of Mysore
Mysore, Karnataka, India

ABSTRACT

In this paper we present an approach to identify opinion of web users from an opinionated text and to classify web user's opinion into positive or negative. Web users document their opinion at opinionated sites, shopping sites, personal pages etc., to express and share their opinion with other web users. The opinion expressed by web users may be on diverse topics such as politics, sports, products, movies etc. These opinions will be very useful to others such as leaders of political parties, selection committees of various sports, business analysts and other stakeholders of products, directors and producers of movies as well as to the other concerned web users. Today web users express their opinion using opinion elements such as opinion phrases, emoticons, short words etc. These form of opinion expressions are very popular and are used by a large number of web users to document their opinion. In this paper we use semantic based approach to find users opinion from opinionated phrases and emoticons. Our approach detects opinionated phrases and emoticons and uses them to obtain semantic orientation scores. These scores are later used to identify users opinion from opinionated texts. Our approach is effective and provides better results on different data sets.

General Terms

Experimentation, Algorithm

Keywords

Sentiment Analysis, Opinion Mining, Affective Computing

1. INTRODUCTION

The rapid development of World Wide Web and its related technologies have fueled the popularity of the web with all sections of society. The web has been used by many firms such as governments, business houses, industries, educational institutions etc., to make them familiar and accessible globally. An individual web user is provided with an opportunity to seek and share knowledge. The web is the origin of many research activities and one interesting area of research is to mine users opinion from web on diverse topics like politics, movies, educational institutions, products etc. The study of opinions is useful to both producers and consumers of the topic. The producers can be manufacturers of automobiles, movie producers, editor of news article, digital product manufactures etc., who are very much interested to find opinion of a user. The consumers are individual users who express their opinion and want to share it with other web users. Today many web users document their opinion on different platforms like discussions forums, opinionated sites, e-commerce sites, blogs, personal web pages etc., the opinion expressed may be in a single line or multiple lines in an opinionated text. For last few years, web has seen new forms of communications, which are quite popular with vast section of the web users. One such form of communication popular with web users is using opinion elements such as normal phrases, emoticons, short words etc.

Example 1 i cant believe this phone. Its too good to be true. I cant think of any one feature that it doesnt have. I have been using this phone for quite some time now and i am absolutely satisfied. Actually make that more than satisfied. I love listening to songs and clicking pictures because of its outstanding quality

Example 2 Well I will review it later but it is one of the most exciting phones to ever come out. Well I do think it might be behind the times compared to older phones from Nokia like the N95 etc but it is still a nice phone **:-).**

Example 3 I bought this product mainly because of its HDD feature. I liked the fact that the videos and pictures could be transferred and saved in a computer rather than disks and cassettes **:)** But the video and picture quality is extremely poor **:(.**

Example 1 refers to an opinion of a web user using nor-mal opinionated phrases. Similarly, example 2 and 3 shows opinion of web users using normal opinionated phrases and emoticons. The phrases in bold represent normal opinionated words, while those

that are bold and underlined represent emoticons. Afore mentioned opinionated texts were collected from reviewcenter [9] and no attempt was made to correct the grammatical mistakes of web users from these opinionated texts.

In this paper, we find opinion of web users expressed as normal phrases and emoticons from opinionated texts. We aim to find opinion of web users from opinionated texts on products. The remainder of this paper is organized as follows: In Section 2 we give a brief description of related work. Then, in Section 3, we discuss our methodology. In Section 4, the experiments and results are discussed. Conclusion is discussed in Section 5.

2. RELATED WORK

Opinion mining is a recent sub discipline of information retrieval which is not about the topic of a document, but with the opinion it expresses [1]. In literature, opinion mining is also known as sentiment analysis [7], sentiment classification [8], affective classification [21] and affective rating [16]. It has emerged in the last few years as a research area, largely driven by interests in developing applications such as mining opinions in online corpora, or customer relation-ship management, e.g., customer's review analysis [21].

Hatzivassiloglou and McKeown [19] have attempted to predict semantic orientation of adjectives by analyzing pairs of adjectives (i.e., adjective pair is adjectives conjoined by and, or, but, either-or, neither-nor) extracted from a large unlabeled document set.

Turney [14] has obtained remarkable results on the sentiment classification of terms by considering the algebraic sum of the orientations of terms as representative of the orientation of the document. Turney an Littman [15] have bootstrapped from a seed set, containing seven positive and seven negative words, and determined semantic orientation according to Pointwise Mutual Information-Information Retrieval (PMI-IR) method.

Wang and Araki [20] proposed a variation of the Se-mantic Orientation-PMI algorithm for Japanese for mining opinion in weblogs. They applied Turney method to Japanese webpage and found results slanting heavily to-wards positive opinion. They proposed balancing factor and neutral expression detection method and reported a well balanced result.

Opinion observer [6] is the sentiment analysis system for analyzing and comparing opinions on the web. The product features are extracted from noun or noun phrases by the association miner. They use adjectives as opinion words and assign prior polarity of these by WordNet exploring method. The polarity of an opinion expression which is a sentence containing one or more feature terms and one or more opinion words is assigned a dominant orientation. The extracted features are stored in a database in the form of feature, number of positive expression and number of negative expression. Kamps et al [12] have focused on the use of lexical relations defined in WordNet. They defined a graph on the adjectives contained in the intersection between the Turney's seed set and WordNet, adding a link between two adjectives whenever WordNet indicate the presence of a synonymy relation between them . The authors defined a distance mea-sure d (t1, t2) between terms t1 and t2, which amounts to the length of the shortest path that connects t1 and t2. The orientation of a term is then determined by its relative distance from the seed terms good and bad

Afore mentioned studies attempt to find opinion of a user from an opinionated text that contains only normal opinionated phrases. Our work uses adjectives to capture users opinion and Sentiment Product Lexicon (SPL) [4] to classify users opinion. We find opinion of web users from both normal phrases and emoticons, popularly expressed by users in opinionated texts

3. METHODOLOGY

We collected 100 opinionated texts with emoticons on Nokia 8310 and another 100 opinionated texts with emoticons on O2 mobile network. We would refer it as Data set 1 and Data set 2. Our Data set 3 comprises of nearly 400 opinionated texts with emoticons present in a few (12.5%) opinionated texts. The opinionated texts were collected manually from different web sites like amazon, review centre, bigadda etc., from 12/01/2010 to 20/02/2010. Our data set consist of nearly 50% positive and 50% negative opinionated texts.

In our approach we pass an opinionated text to a sentence splitter program. The sentences obtained from the program were input to a part of speech tagger. The tagger used is Monty Tagger [11]. We use different forms of adjective such as /JJ, /JJS and /JJR as extraction patterns to capture opinionated phrases that expresses users opinion. i.e. we extract a phrase associated with an adjective. The tagged opinionated sentences are parsed to obtain opinionated phrases that are likely to contain user's opinion.

For example consider an opinionated text "**i really like this Sony DSC-T1 digital camera. It's the perfect snap-shot, take everywhere camera**." After passing it through a sentence splitter program and a part of speech tagger pro-gram, we obtain the tagged opinionated sentences such as **i/NN really/RB like/IN this/DT Sony/NNP DSC/NNP -/: T1/NNP digital/JJ camera/NN ./. It/PRP 's/VBZ the/DT perfect/JJ snapshot/NN ,/, take/VB everywhere/RB cam-era/NN ./.**

JJ represent adjective, NN/NNP represent different forms of noun, RB represent adverb, VBZ/VB represent different forms of verbs, DT represent determiner, IN rep-resent preposition and PRP represent personal pronoun. We use adjective as an extraction pattern to find opinion of a user, we obtain phrases like **digital/JJ and perfect/JJ** as potential opinionated phrase

Similarly, consider the opinionated texts such as "**this phone has a weaker battery**" and "**this phone is weakest in its battery**". We obtain **this/DT phone/NN has/VBZ a/DT weaker/JJR battery/NN** and **this/DT phone/NN is/VBZ weakest/JJS in/IN its/PRP$ battery/NN** as tagged opinionated texts. Here, JJR and JJS represent comparative and superlative form of adjective. We use porter stemmer program described in [10] and apply it to only adjectives in tagged opinionated texts. The objective is to convert the comparative and superlative form of an adjective to its base form.

3.1 Polarity Detector

The phrases obtained after stemming are subjected to Sentiment Product Lexicon for capturing only subjective or opinionated phrases. This is necessary as some phrases obtained after application of extraction patterns may be non subjective or non opinionated such as digital in afore example.

Sentiment Product Lexicon is collection of General lexicon and Domain lexicon. General lexicon will maintain a list of positive and negative words by collecting opinion words that are positive or negative from sources like General Inquirer [17], Subjective clues [18] and list of adjectives [13]. Domain lexicon will maintain a list of positive or negative words from the domain context. We found words like cool, revolutionary etc., appeared in negative list of General lexicon. These words were used to express positive opinion by web users. Hence we created a domain lexicon to have opinion words from the domain perspective. The details of construction of General lexicon and Domain lexicon are made available in [4].

In this paper we use these lexicons to identify neutral phrases. Sentiment product lexicon can be expressed as

SPL = {GLP, GLN , DLP, DLN } (1)

Where
GLP : Positive words in General lexicon
GLN : Negative words in General lexicon
DLP : Positive words in Domain lexicon
DLN : Negative words in Domain lexicon

Consider for example the opinionated texts "**This phone is bad".** After application of part of speech tagging we obtain **This/DT is/VBZ bad/JJ**. When the extraction patterns are applied, we obtain bad/JJ as the opinionated phrase This phrase is passed to SPL to find the polarity. If polarity of the phrase is found in positive list of Domain lexicon, then the phrase is considered as positive and it is assigned a score +1.
If polarity of the phrase is found in negative list of Domain lexicon, then the phrase is considered as negative and it is assigned a score -1. If the polarity of the phrase is found neither in positive nor in negative list of Domain lexicon, then the positive and negative list of General lexicon is consulted to find the polarity of the phrase.
After determining the polarity of the phrase, the opinionated sentence is parsed to find a negator, such as not, at a window size of 10 from the occurrence of the opinionated phrase. Our intuition is the presence of a negator within a window size of 10 would reverse the polarity of opinionated phrase.

For example, consider an opinionated sentence "**This phone is not bad".** The opinionated phrase obtained is bad, after application of extraction patterns. The polarity of the phrase, after subjecting it to SPL, is found to be negative with a score of -1. The opinionated sentence is parsed to find a negator, the presence of not will reverse the polarity of the phrase and the score of opinionated sentence will be-come +1. If there is no negator in an opinionated sentence, the score of opinionated sentence will be similar to the score of opinionated phrase. We compute the score of phrases in an opinionated sentence using Equation 2.

$$\text{Score sentence}_i\text{ (phrase)} = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{sentence (phrase }_j\text{)} \quad (2)$$

3.2 Emoticon Lexicon

There are two types of emoticons namely textual emoticons and graphical icons. In this paper we use only textual emoticons for detection opinion of web users. We collected variety of emoticons from many sources like wikipedia, bigadda, rediff, amazon, review centre, yahoo etc. The total number of emoticons collected is 240. One major issue with emoticons is that, there are no standards for emoticons as a result we had two tasks at hand

1. To identify emoticons that express opinion
2. To classify emoticons as positive opinionated emoticons and negative opinionated emoticons.

In order to achieve the two tasks, we asked undergraduate students from diverse fields of engineering to evaluate emoticons that could be used to express opinion. We used a voting approach to include or exclude emoticons into Emoticon Lexicon.
We included an emoticon into Emoticon Lexicon and labeled it as positive or negative when there is 60% agreement among evaluators. The total number of emoticons made available, after evaluation, in Emoticon Lexicon is 125, with 62 positive opinionated emoticons and 63 negative opinionated emoticons.

Table 1: List of Emoticons

Slno.	Positive Emoticons	Negative Emoticons
1.	:-)	D=
2.	:-O	D:
3.	:D	b(
4.	:-D	:@
5.	:))	:-!
6.	:)	:-(
7.	:,')	8-O
8.	:-"	:c
9.	:-	X[
10.	:-I	:(

Our intuition is "picture provides a better sentiment expression of a web user." Consider the following examples

Example 4 I am feeling happy.
Example 5 I am feeling happy :)

Example 1 conveys opinion of a user using a phrase "happy". Similarly, example 2 conveys opinion of a user using a phrase "happy" as well using an emoticon :). The use of emoticon provides an accelerated positive opinion of a user compared to opinion expressed in example 1. We believe emoticons convey opinion of a user and accelerate or decelerate the opinion of a web user when used with opinionated phrases.

Table 1 lists only a few positive and negative opinionated emoticons due space constraints. Opinionated sentences from opinionated texts are parsed to identify emoticons using Emoticon Lexicon. A score of +2 or -2 is assigned to positive opinionated emoticons or negative opinionated emoticons. We compute the score of emoticons using Equation 3.

$$\text{Score sentence}_i\text{(emoticon)} = \sum_{i=1}\sum_{j=1} \text{sentence}_i\text{(emoticon}_j\text{)} \quad (3)$$

4. EXPERIMENT

Our objective is to find opinion of web users from opinionated texts consisting of opinionated phrases and emoticons. Our approach of finding opinion from phrases with adjectives using SPL is found to be better than the base line approach of finding opinion with adjectives using [15].

Our data sets consist of opinionated text consisting of both opinionated phrases and emoticons. Opinionated text from our data set is passed to sentence splitter program to obtain opinionated sentences.

The sentences are then subject to part of speech tagger and SPL to determine the score of each opinionated phrase.

Table 2: Result of Opinion Elements

Slno.	Opinion element	Data set	Accuracy
1.	Opinion Phrase	Data set 1	68%
2.	Emoticon	Data set 1	78%
3.	Opinion Phrase and Emoticon	Data set 1	84%
4.	Opinion Phrase	Data set 2	52%
5.	Emoticon	Data set 2	64%
6.	Opinion Phrase and Emoticon	Data set 2	73 %
7.	Opinion Phrase	Data set 3	64.24%
8.	Emoticon	Data set 3	10.5%
9.	Opinion Phrase and Emoticon	Data set 3	68.4 %

The opinionated sentence is once again parsed to find the presence of any positive or negative opinionated emoticons. A score of +2 or -2 or 0 is added to score of the sentence, if positive opinionated emoticons are present or negative opinionated emoticons are present or no emoticons are present. Opinion of the user from an opinionated text is determined using the Equation 4.

$$\text{Average SO (Text)} = \frac{1}{n}\sum_{i=1}^{n} \text{Score sentence (phrase + emoticon)} \quad (4)$$

SO is semantic orientation. An opinionated text is classified as positive, if average semantic orientation is greater than the threshold. It is classified as negative if average se-mantic orientation is less than the threshold. We use zero as the threshold to classify opinionated texts.

Table 2 shows the accuracy of our approach with different opinion elements like opinion phrases, emoticons and combination of opinion phrases and emoticons on data set 1, data set 2 and data set 3. We have implemented well known approaches discussed in Liu [6] and Turney [14] on all the data sets. Table 3 provides results of our approach against well known approaches.

We compute the classification accuracy by dividing the sum of true positives and true negatives with total number of items to be classified. True positive represent number of opinionated texts classified correctly as positive, similarly true negative represent number of opinionated texts classified correctly as negatives. Our result obtained is good considering that we use only adjectives to find opinionated phrases.

We obtain an accuracy of 68%, 52% and 64.24% for Data set 1, Data set 2 and Data set 3 considering only opinionated phrases.

An accuracy of 84%, 73% and 68.4% were obtained for Data set 1, Data set 2 and Data set 3 considering both opinionated phrases and emoticons. While, an accuracy of 78%, 64% and 10.5% were obtained considering only emoticons on Data set 1, Data set 2 and Data set 3. Figures 1, 2 and 3 show positive accuracy and negative accuracy of opinionated texts on Data set 1, Data set 2 and Data set 3 with different opinion elements. Similarly, Figures 4, 5 and 6 show positive accuracy and negative accuracy using different approaches on Data set 1, Data set 2 and Data set 3.

Table 3: Result of Our Approach

Slno.	Approach	Data set	Accuracy
1.	Liu [6]	Data set 1	63%
2.	Turney [14]	Data set 1	67.5%
3.	Our Approach	Data set 1	84%
4.	Liu [6]	Data set 2	64%
5.	Turney [14]	Data set 2	60.2%
6.	Our Approach	Data set 2	73%
7.	Liu [6]	Data set 3	61.78%
8.	Turney [14]	Data set 3	63.5%
9.	Our Approach	Data set 3	68.4%

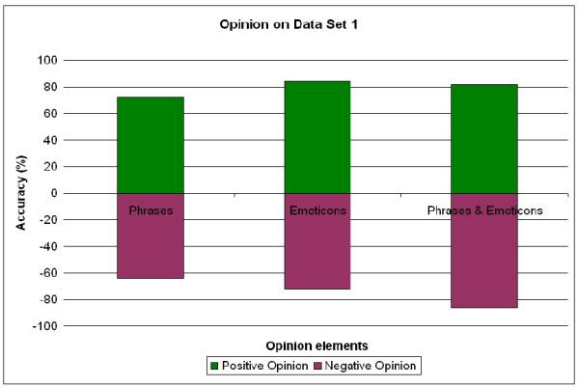

Figure 1: Accuracy of Opinion elements on Data Set 1

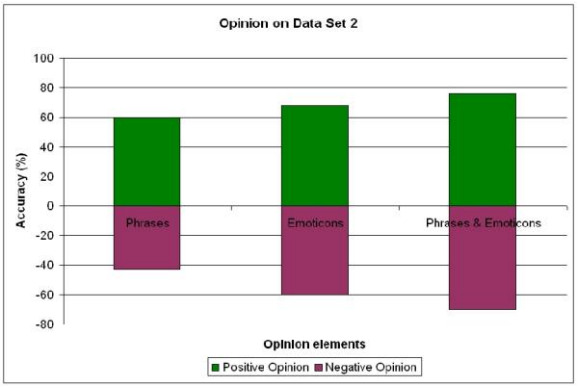

Figure 2: Accuracy of Opinion elements on Data Set 2

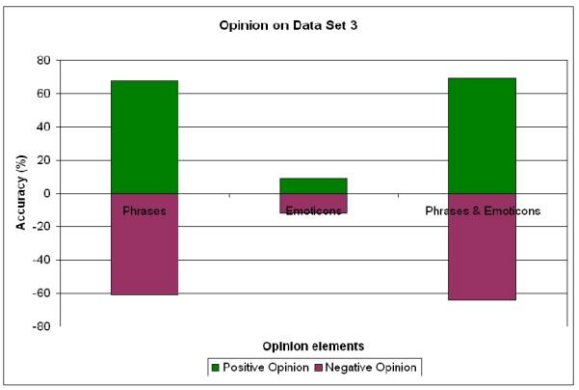

Figure 3: Accuracy of Opinion elements on Data Set 3

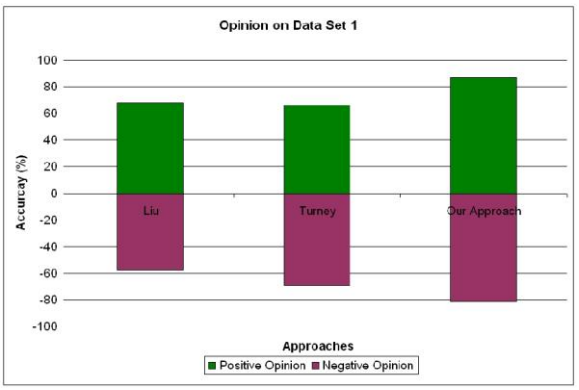

Figure 4: Accuracy of Approaches on Data Set 1

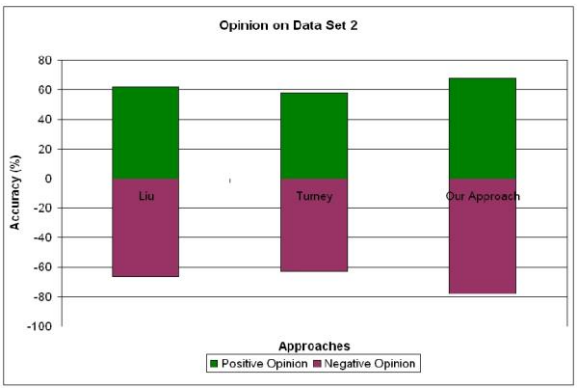

Figure 5: Accuracy of Approaches on Data Set 2

We found that, many users expressed their opinion considering adjectives and also other part-of-speech like verb, adverb, noun etc, hence obtaining above mentioned accuracy for Data set 1, Data set 2 and Data set 3.

We also observed a few opinionated texts contained single emoticon that summarizes overall opinion of users aiding in correct classification of users opinion. Also, a few other opinionated texts contained more than one emoticon expressing opinion of users about the features of the product and an overall opinion about the product. This use of multiple emoticons to express opinion on features contradicts the overall opinion of product, leading to drop in accuracy of opinion detection with only emoticons compared to usage of both opinion phrase and emoticons.

Conclusion

We have discussed an approach that detects opinion of web users from opinionated texts. Opinions are expressed by the web users using normal phrases or emoticons or both normal phrases and emoticons or both. Our approach finds opinion of a web user from normal phrases, emoticons or both normal phrases and emoticons, classifies the opinion as positive or negative. We found 16% rise in accuracy considering both opinion phrases and emoticons on Data set1, 21% rise considering both opinion phrases and emoticons on Data set 2 and 4% rise considering both opinion phrases and emoticons on Data set 3 as against finding opinion only from phrases. Our approach is found to be better than the popular approaches on different data sets used in the experiment.

References

[1] Andrea, Esuli, Fabrizio, Sebastiani: Determining term subjectivity and term orientation for opinion mining. In: Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy(2006)

[2] Andrea, Esuli, Fabrizio, Sebastiani: Determining the semantic orientation of terms through gloss classification. In: Proceedings of 14th ACM International Conference on Information and Knowledge Management, pp. 617-624. Bremen, Germany(2005)

[3] Alistair, Kennedy, Diana, Inkpen: Sentiment Classification of Movie and Product Reviews Using Contextual Valence Shifters. In: Proceedings of FINEXIN 2005, Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations. canada(2005)

[4] Anil, kumar, K.M, Suresha: Identifying Subjective Phrases From Opinionated Texts Using Sentiment Product Lexicon, International Journal of Advanced Engineering & Applications. 2(263-271), (2010)

[5] Anil, Kumar, K.M, Suresha: Detection of Neutral Phrases and Polarity Shifting of Few Phrases for Effective Classification of Opinionated Texts, International Journal of Computational Intelligence Research. 6(1) 43-58, (2010)

[6] Bing, Liu, Minqing, Hu, Junsheng, Cheng: Opinion Observer: Analyzing and Comparing Opinions on the Web. Chiba, Japan(2005)

[7] Bo, Pang, Lillian, Lee: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of 42nd Meeting of the Association for Computational Linguistics, pp. 271-278. Barcelona, Spain(2004)

[8] Bo, Pang, Lillian, Lee, Shivakumar, Vaithyanathan: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of 7th Conference on Empirical Methods in Natural Language Processing, pp. 79-86. Philadelphia, US(2002)

[9] http://www.reviewcentre.com/

[10] http://tartarus.org/ martin/PorterStemmer

[11] Hugo:MontyLingua: An end-to-end natural language processor with common sense. (2003)

[12] Jaap, Kamps, Maarten, Marx, Robert, J., Mokken, Maarten, De, Rijke:Using word-net to measure seman-tic orientation of adjectives. In: Proceedings of 4th Inter-national Conference on Language Resources and Evaluation, pp. 1115-1118. Lisbon, Portugal(2004)

[13] Maite, Taboada, Jack, Grieve:Analyzing appraisal automatically. In: Proceedings of the AAAI Symposium on Exploring Attitude and Affect in Text: Theories and Applications. (2004)

[14] Peter, D, Turney: Thumbs up or thumbs down? Se-mantic orientation applied to unsupervised classification of reviews. In: Proceedings of 40th Annual Meeting of the association for Computational Linguistics, pp. 417-424. Philadelphia (2002)

[15] Peter, D, Turney, Michael, L, Littman: Measuring praise and criticism: Inference of semantic orientation from association, ACM Transactions on Information Systems, pp. 315-346. (2003)

[16] Sara, Owsley, Sanjay, Sood, Kristian, J, Hammond: Domain specific affective classification of document. In: Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs. California, US (2006)

[17] Stone, P.J: Thematic text analysis: New agendas for analyzing text content. In C. Roberts (Ed.), Text Analysis for the Social Sciences, Mahwah. Erlbaum (1997)

[18] Theresa, Wilson, Janyce, Wiebe, Paul, Hoffmann: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: Proceedings of HLT/EMNLP, Vancouver, Canada, (2005)

[19] Vasileios, Hatzivassiloglou, Kathleen, R., McKeown: Predicting the semantic orientation of adjectives. In: Proceedings of 35th Annual Meeting of the Association for Computational Linguistics, pp. 174-181, Madrid, Spain(1997)

[20] Wang,Araki: Modifying SO-PMI for JapaneseWeblog Opinion Mining by Using a Balancing Factor and Detecting Neutral Expressions. In: Proceedings of NAACL HLT 2007, pp. 189-192, Rochester, New York, US (2007)

[21] Youngho, Kim, Sung-Hyon, Myaeng: Opinion Analysis based on Lexical Clues and their Expansion In: Proceedings of NTCIR-6 Workshop Meeting, pp 308-315, Tokyo, Japan (2007)

ABOUT THE AUTHORS.

Dr. Suresha is presently working as Associate Professor, Department of Studies in Computer Science, University of Mysore. He received his B.Sc. degree in 1987 from University of Mysore, M.Sc degree in 1989 from University of Mysore and M. Phil degree in 1991 from DAVV, Indore. He received M.Tech degree from Indian Institute of Technology, Kharagpur in 1996 and Ph. D. from prestigious Indian Institute of Science, Bangalore in 2007. He was awarded second prize in IRISS-2002 competition, which is an all India level research student competition called" Inter Research Institute Student Seminar". He has a teaching experience of 20 years at post graduate level, has several publications to his credit and currently supervising five research scholars towards their doctoral work.

Anil Kumar K.M is a Research Scholar, Department of Studies in Computer Science, University of Mysore. He received his B.E. Degree in 1999 from University of Mysore and M.Tech Degree in 2006 from Visvesvaraya Technological University, Karnataka. He is working towards doctoral degree under the supervision of Dr. Suresha