LEXICAL ANALYSIS TO EFFECTIVELY DETECT USERS' OPINION

Anil Kumar K.M and Suresha

Department of Studies in Computer Science University of Mysore
Manasagangothri Mysore, India
{anilkmsjce@yahoo.co.in, sureshabm@yahoo.co.in}

Abstract:

In this paper we present a lexical approach that will identify opinion of web users popularly expressed using short words or sms words. These words are pretty popular with diverse web users and are used for expressing their opinion on the web. The study of opinion from web arises to know the diverse opinion of 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 stake holders of products, directors and producers of movies as well as to the other concerned web users. We use semantic based approach to find users opinion from short words or sms words apart of regular opinionated phrases. Our approach efficiently detects opinion from opinionated texts using lexical analysis and is found to be better than the other approaches on different data sets.

Key words:

Artificial Intelligence, Sentiment Analysis, Opinion Mining

1. Introduction

Short words also known as SMS language [8] [9] [10] are quite popular with web users. There are many arguments [8] [9] and counter arguments [10] about its use in communication and its impact on linguistic development of future generations. We believe these short words continue to exist and would conceal facts and opinions important to others. For example, word like excellent is written in short words as xllent, xlent etc., the use of short words are found to be very popular as it conveys message at less time. The limitation is that there are no standards for short words, making it very difficult for processing. A few web users use one or more short words in their communication and opinionated text are no exceptions. Following are the examples of opinionated text with regular and short opinionated words collected from opinionated site review centre and retained in same form.

Example 1 *Well it is one of the most* ***exciting*** *phones to ever come out. 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 2 *A* ***gr8*** *TV for the dollar. Samsung has designed and produced a* ***5n*** *TV. The picture quality is* ***xllent*** *with analog and high definition signal. Sound quality is* ***gud*** *to* ***xllent****. It has a wide angle of picture side vision. It is a very compact design for a TV of this size. Instructions are very* ***xllent*** *with simple set-up.*

DOI : 10.5121/ijwest.2011.2408

Example 1 conveys opinion of a web user with regular opinionated words. Similarly, example 2 conveys opinion of a web user using regular and short opinionated words. The words in bold represent regular opinionated words, those that are bold and underlined represent short opinionated words. The afore used short words like gr8, 5n, gud, xllent are commonly used to represent regular words great, fine, good and excellent.

In this paper, we focus on detecting opinions expressed by web users using short and regular opinion words only 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 experimental 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]. We have referred many literatures on opinion mining, due to space constraint only a few are below mentioned.

Hatzivassiloglou and McKeown [15] 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 unlabelled document set.

Turney [13] 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

.

Wang and Araki [16] proposed a variation of the Semantic Orientation-PMI algorithm for Japanese for mining opinion in weblogs. They applied Turney method to Japanese webpage and found results slanting heavily towards positive opinion. They proposed balancing factor and neutral expression detection method and reported a well balanced result.

Kamps et al [11] 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 measure 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.

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.

Anil and Suresha [5] proposed an approach for detecting opinion from short words using short word lexicon. It involves in searching an opinionated text with entries of short word lexicon and translating short words with regular words in an opinionated text for opinion detection.

Our work differs from afore mentioned studies by finding opinion of a user from both regular and short opinionated words in an opinionated text. Our work uses adjectives as well as parts-of-speech like verb, adverb etc., to capture opinionated words for efficient opinion detection. Also the use of lexical analysis with extraction patterns eliminates searching and translation phases for opinion detection from short opinionated words.

3 Methodology

We collected nearly 2000 opinionated texts from sources such as web search engines like Google, Altavista, Exalead etc., opinionated sites like Amazon, CNet, review centre, bigadda, rediff etc., and from researchers [2][6] for experimentation.

Our data sets comprised of predominantly of normal opinionated words with a few short words used for expressing opinions. We passed these data sets to group of 10 engineering students of diverse disciplines to rephrase regular opinion words with their popular short opinion words, while retaining a copy of original data sets for further processing. We refer to original opinionated texts as Data Set 1 and rephrased opinionated texts as Data set 2. All opinionated texts, both original and rephrased, are subjected to a parts of speech tagger. The tagger used is Monty Tagger [7]. The tagged opinionated texts are then subjected to extraction patterns to obtain opinionated phrases that are likely to contain user's opinion. Table 1 and 2 shows a few extraction patterns used to find opinionated phrases, where JJ represent adjective, CD represent cardinal and NN/NNS, VB/VBD/VBN/VBG, RB/RBR/RBS represent different forms of noun, verb and adverb.

Table 1. Extraction pattern-1

Slno.	First Word	Second Word	Third Word
1	JJ	NN or NNS	anything
2	RB,RBR or RBS	JJ	not NN nor NNS
3	JJ	JJ	not NN nor NNS
4	NN or NNS	JJ	not NN or NNS
5	RB,RBR or RBS	VB,VBD,VBN or VBG	anything
6	NN, NNS or NNP	NN or NNS or NNP	anything
7	RB, RBR or RBZ	VB or VBD or VBG or VBN	anything
8	RB, RBR or RBZ	CD	anything
9	CD	NN, NNS or NNP	anything

Table 2. Extraction pattern-2

Slno.	Word
1	JJ or JJS or JJR
2	RB or RBR or RBS or RBZ
3	NN or NNS or NNP
4	VB or VBD or VBG or VBN
5	CD

We use Sentiment Product Lexicon (SPL) to capture only subjective or opinionated phrases and perform polarity shifting of a few phrases to detect opinion of web users from opinionated texts. The detail of SPL is provided in [3]. We also employ the short word lexicon as discussed in [5] to aid detecting short words from opinionated texts.

We compute the average semantic orientation of the opinionated text by considering all scores of opinionated phrases as shown in Equation 1. We classify an opinionated text as positive, if the average semantic orientation of opinionated text is greater than a threshold and negative when the average semantic orientation is less than a threshold. The threshold used here is 0.

$$\text{SO (Opinionated Text)} = 1/n \,.\, \sum_{i=1}^{n} (\text{Opinionated Phrase}_{i}) \qquad (1)$$

4 . Experiments and Results

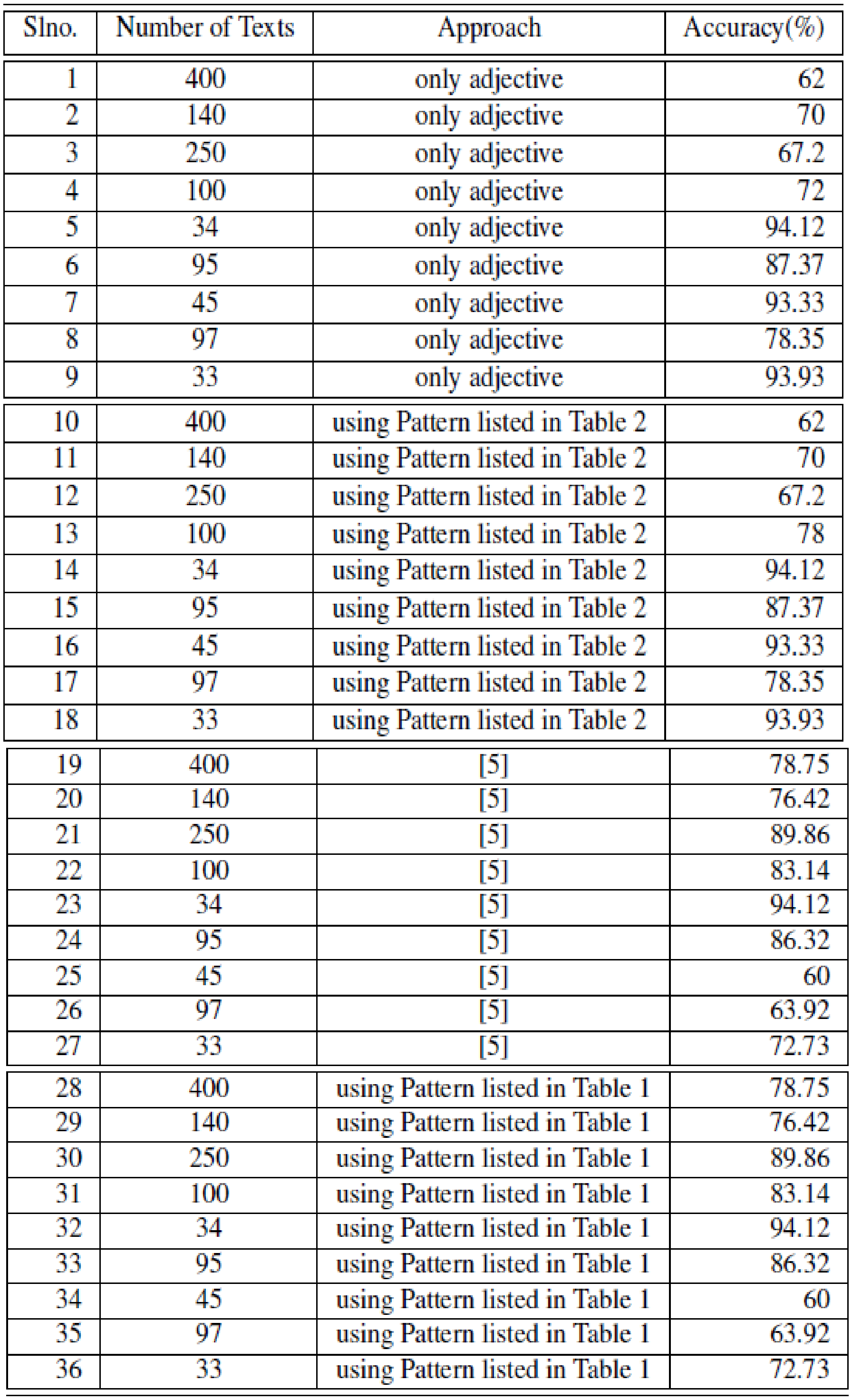
The extraction patterns listed in Table 1 and 2 are used with afore mentioned approach to obtain opinion from opinionated texts. Consider the following opinionated texts

Example 1 "i luv this product it is gr8 to have such a nice phn". Short words **luv, gr8** and **phn** is used by author of the review to mean love, great and phone as per Short word Lexicon. Application of tagger will result in the following tagged **texts i/NN luv/VBG this/DT product/NN it/PRP is/VBG gr8/CD to/TO have/VB such/JJ a /DT nice/JJ phn/NN**. Patterns in Table 1 will extract opinionated phrase **nice/JJ phn/NN** without considering other opinionated phrases like **luv** and **gr8**. Patterns in Table 2 will extract all opinionated phrases such as **luv, gr8** and **nice** from the text.

Example 2 "**apple iphone simply bst i really liked it**". Short word bst is used by author of review to mean best. Its tagged text is **apple/NN iphone/NN simply/RB bst/VB i/NN really/RB liked/ VBD it/PRP**. Patterns in Table 1 will extract phrases such as **simply/RB bst/VB and really/RB liked/ VBD**. Extraction patterns from Table 2 will yield opinionated phrases such as bst and liked.

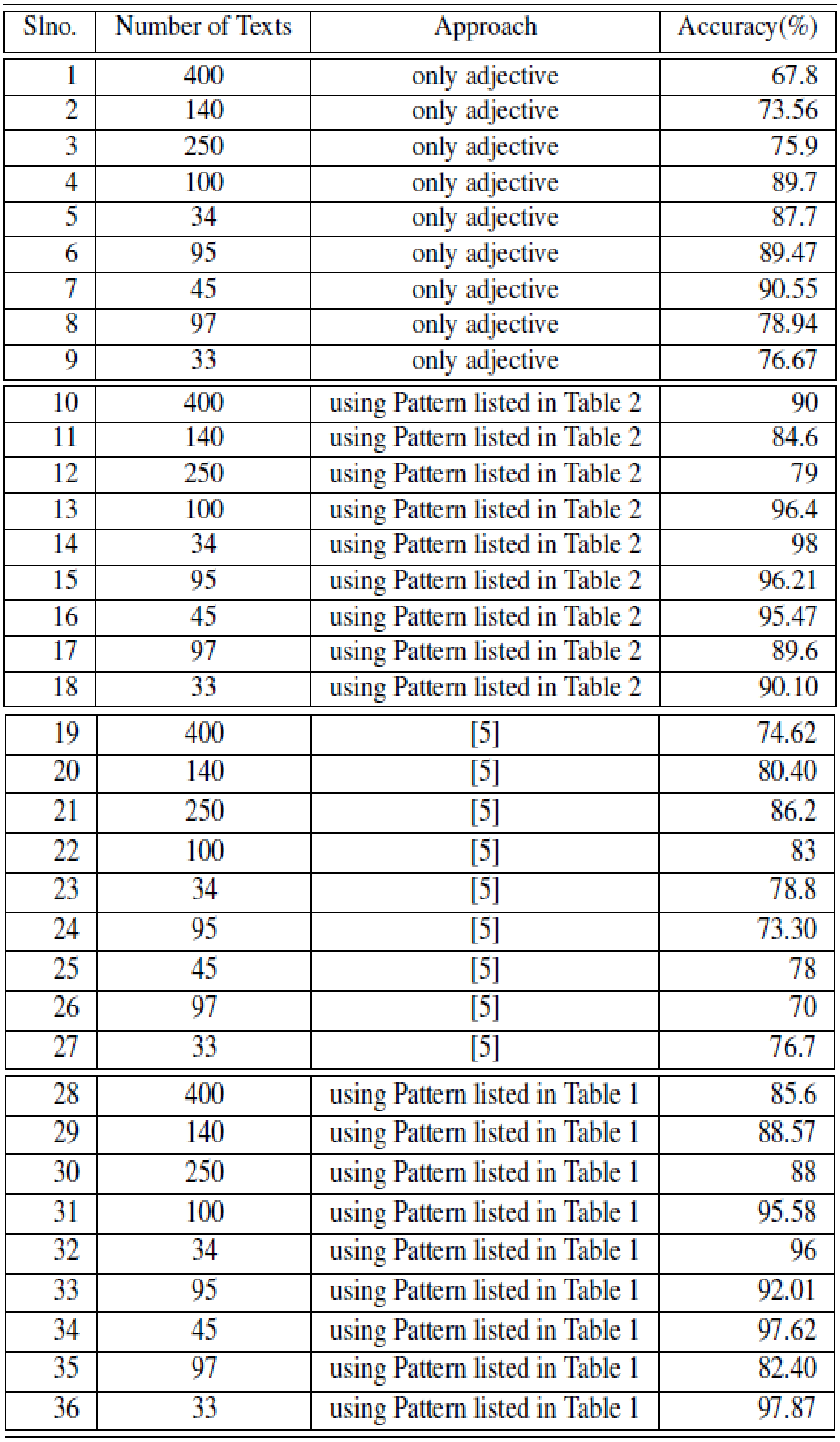
Example 3 "**keep doing gud and come with new camera soon nikon**". Short word gud is used by author to mean good. Its tagged text is **keep/VB doing/ VBG gud/NN and/CC come/VB with/IN new/JJ camera/NN soon/RB Nikon/VB**. Patterns in Table 1 will extract a non opinionated phrase **new/JJ camera/NN** which will be eliminated using SPL. Table 2 patterns will capture an opinionated phrase gud from the text.

Table 3. Results on Data Set 1



Slno.	Number of Texts	Approach	Accuracy(%)
1	400	only adjective	62
2	140	only adjective	70
3	250	only adjective	67.2
4	100	only adjective	72
5	34	only adjective	94.12
6	95	only adjective	87.37
7	45	only adjective	93.33
8	97	only adjective	78.35
9	33	only adjective	93.93
10	400	using Pattern listed in Table 2	62
11	140	using Pattern listed in Table 2	70
12	250	using Pattern listed in Table 2	67.2
13	100	using Pattern listed in Table 2	78
14	34	using Pattern listed in Table 2	94.12
15	95	using Pattern listed in Table 2	87.37
16	45	using Pattern listed in Table 2	93.33
17	97	using Pattern listed in Table 2	78.35
18	33	using Pattern listed in Table 2	93.93
19	400	[5]	78.75
20	140	[5]	76.42
21	250	[5]	89.86
22	100	[5]	83.14
23	34	[5]	94.12
24	95	[5]	86.32
25	45	[5]	60
26	97	[5]	63.92
27	33	[5]	72.73
28	400	using Pattern listed in Table 1	78.75
29	140	using Pattern listed in Table 1	76.42
30	250	using Pattern listed in Table 1	89.86
31	100	using Pattern listed in Table 1	83.14
32	34	using Pattern listed in Table 1	94.12
33	95	using Pattern listed in Table 1	86.32
34	45	using Pattern listed in Table 1	60
35	97	using Pattern listed in Table 1	63.92
36	33	using Pattern listed in Table 1	72.73

Table 4. Results on Data Set 2



Slno.	Number of Texts	Approach	Accuracy(%)
1	400	only adjective	67.8
2	140	only adjective	73.56
3	250	only adjective	75.9
4	100	only adjective	89.7
5	34	only adjective	87.7
6	95	only adjective	89.47
7	45	only adjective	90.55
8	97	only adjective	78.94
9	33	only adjective	76.67
10	400	using Pattern listed in Table 2	90
11	140	using Pattern listed in Table 2	84.6
12	250	using Pattern listed in Table 2	79
13	100	using Pattern listed in Table 2	96.4
14	34	using Pattern listed in Table 2	98
15	95	using Pattern listed in Table 2	96.21
16	45	using Pattern listed in Table 2	95.47
17	97	using Pattern listed in Table 2	89.6
18	33	using Pattern listed in Table 2	90.10
19	400	[5]	74.62
20	140	[5]	80.40
21	250	[5]	86.2
22	100	[5]	83
23	34	[5]	78.8
24	95	[5]	73.30
25	45	[5]	78
26	97	[5]	70
27	33	[5]	76.7
28	400	using Pattern listed in Table 1	85.6
29	140	using Pattern listed in Table 1	88.57
30	250	using Pattern listed in Table 1	88
31	100	using Pattern listed in Table 1	95.58
32	34	using Pattern listed in Table 1	96
33	95	using Pattern listed in Table 1	92.01
34	45	using Pattern listed in Table 1	97.62
35	97	using Pattern listed in Table 1	82.40
36	33	using Pattern listed in Table 1	97.87

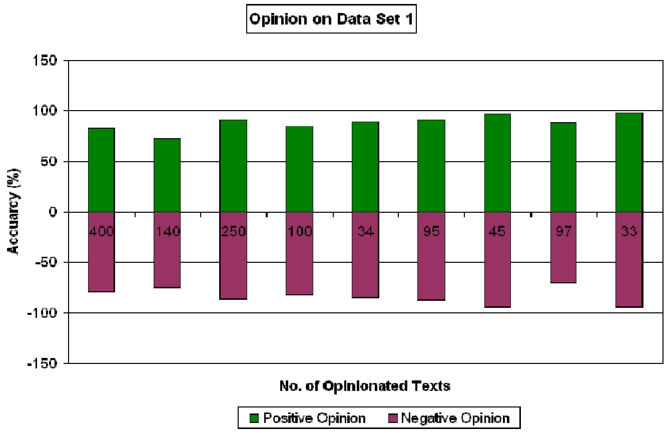


Fig. 1. Summary of Users Opinion on Data Set 1

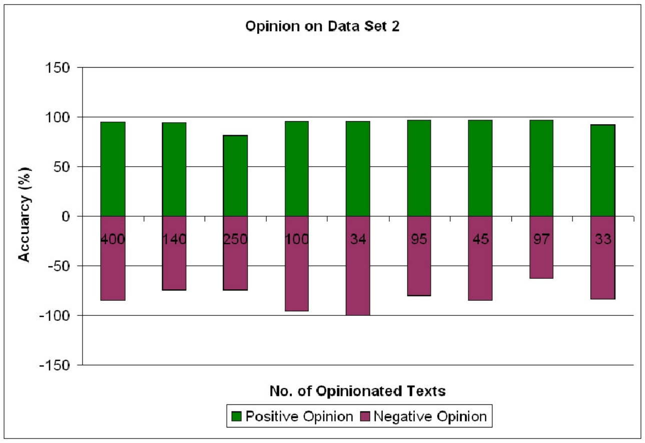


Fig. 2. Summary of Users Opinion on Data Set 2

We compute the accuracy of our approach by considering true positives and true negatives divided by total number of opinionated texts. True positives represent number of opinionated texts classified correctly as positive. Similarly, true negatives represent number of opinionated texts classified correctly as negatives.

Table 3 shows the result of various approaches using extraction patterns listed in Table 1 and 2 on Data set 1. We found the results of approaches using adjective and using patterns listed in Table 2 to be very similar. This is because majority of opinionated texts in Data set 1 uses normal opinion phrase to expresses users opinion. Results of other approaches using patterns listed in [5] and patterns listed in Table 1 were very similar on Data set 1. Similarly, Table 4 shows the result of various approaches using extraction patterns listed in Table 1 and 2 on Data set 2. We found results obtained using patterns listed in Table 1 to be better than the other patterns employed to detect users opinion from opinionated texts. Results obtained from the combined list of patterns from Table 1 and 2 was found to be very similar to result obtained using patterns listed in Table 2. This combined list of patterns can be used to effectively detect web users opinion from opinionated texts. Figure 1 and 2 shows positive accuracy and negative accuracy of users opinion on Data Set 1 and Data Set 2.

5 . Conclusion

We have discussed the use of different extraction patterns to detect opinion from opinionated texts consisting of normal and short opinionated words. We have achieved a better accuracy with patterns listed in Table 1 and 2 on both Data set 1 and Data set 2. An increased accuracy of 2.1 %, 0.66% and 2.11% is obtained with patterns listed in Table 2 as compared to patterns listed in Table 1, using only adjective and using pattern mentioned in [5] on Data set 1. We obtain a better accuracy for patterns listed in Table 1 as against other approaches on Data set 2. An increased accuracy of 0.47%, 10.37% and 13.62% is obtained with patterns listed in Table 1 as compared to patterns listed in Table 2, using only adjective and using pattern mentioned in [5] on Data set 2. A combined list of patterns from Table 1 and 2 obtained accuracy similar to patterns listed in Table 1 on Data set 2. But, they can be used to capture effectively user opinion spanning different parts of speech as well as opinionated phrases consisting of normal and short phrases.

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. 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)

3. Anil, kumar, K.M., Suresha: Identifying Subjective Phrases From Opinionated Texts Using Sentiment Product Lexicon. International Journal of Advanced Engineering & Applications. 2, 63-271 (2010)

4. 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, 43-58 (2010)

5. Anil, Kumar, K.M., Suresha: Detection of Web Users' Opinion from Normal and Short Opinionated Words. In Proc: International Conference on Data Engineering and Management, Bishop Heber College(Autonomous), July 29-31 (2010)

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

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

8. http://www.dailymail.co.uk/news/article-483511/ -h8-txt-msgs-How-texting-wrecking-language.html.

9. http://news.bbc.co.uk/2/hi/uk news /education/2197173.stm

10. http://entertainment.timesonline.co.uk/tol /arts and entertainment/books/non-fiction/article4356458.ece

11. Jaap, Kamps, Maarten, Marx, Robert, J., Mokken, Maarten, De, Rijke: Using wordnet to measure semantic orientation of adjectives. In: Proceedings of 4th International Conference on Language Resources and Evaluation, pp. 1115-1118. Lisbon, Portugal (2004)

12. Livia, Polanyi, Annie, Zaenen: Contextual Valence Shifters. Computing Attitude and Affect in Text: Theory and Applications, pp. 1-10. (2006)

13. Peter, D., Turney: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pp. 417-424. Philadelphia, US (2002)

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

15. 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)

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