I.J. Education and Management Engineering, **2015, 2, 20-29**
Published Online June 2015 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2015.02.03

Available online at http://www.mecs-press.net/ijeme

A Contrast Between Systematic and Automated Sentiment Analysis

R.Nithya[a,*], Dr.D.Maheswari[b]

[a] *Assistant Professor ,RVS College of Arts and Science, Sulur, TamilNadu, India*
[b] *Assistant Professor ,RVS College of Arts and Science, Sulur, TamilNadu, India*

Abstract

Sentiment analysis mainly focuses on subjectivity and polarity detection. Today consumer makes buying decision based on the customer's review that is available in each of the online shopping sites. There are some of the specific websites which discuss about positive and negative facts of those products that comes to market. Hence this type of analysis are socially very needed for sellers to undergo market analysis, branding, product penetration, market segmentation and so on. This paper mainly focuses on difference between systematic and automated methods of determining the positive and negative polarity distribution of Samsung Tablet PC.

Index Terms: Opinion Mining, Feature Extraction, Sentiment Classification.

1. Introduction

Social media are popularly known as 'democracy's pipeline', 'an amplifier of unfiltered emotion', 'an organism with a million tongues and twice as many eyes' and as 'a virtual megaphone with a global reach'. Recent surveys on media by research firm Social Bakers and Semiocast a Paris states that; 75% of web users in India are below the age of 35years, 42% smartphone users in India use device to access news, nearly 72% netizens lives in urban areas, nearly 52% internet users connect to web via a mobile phone and about 1.5Lakh new internet users added every month in India. Furthermore, Forrester estimates that Indians spent around $1.6 billion online on retail e-commerce sites in 2012. By 2016 it can either extend upto $8.8 billion. So that the online shopping sites are engaging with their consumers on the emotional front as well as fulfilling their need for information in order to indicate that they are not limited to satisfy only on their functional needs. Most of the consumers voluntarily make digital footprints while signing up on social media networks, by just signing up on new shopping websites, the likes of Snapdeal or Myntra. That is they do not make buying decision immediately on shopping website by placing an order. Instead they make purchase decisions as they move in and out of TV and print commercials, a friend's recommendation on social media, product information online, product reviews on a trusted blog and the best deals in their local store. They are totally engaged across all the

* Corresponding author. Tel: 98422 64572
E-mail address: nithya.r@rvsgroup.com

places where they are about to access information they need and then move to final step of ordering. Mostly companies share or sell this kind of information with other companies that leverage it to market their products to people around location, age, gender and other such parameters. They do this to decide whether we fit to their target group and such a concept is called data mining. Opinion mining or Sentiment analysis is an important sub discipline of Data mining and Natural Language Processing which deals with building a system that explores the user's opinions made in blog spots, comments, reviews, discussions, news, feedback or tweets, about a product, policy, person or topic.

To be specific, opinion mining can be defined as a sub discipline of computational linguistics that focuses on extracting people's opinion form the web. It analyses from a given piece of text about; which part is opinion expressing; who wrote the opinion; what is being commented. Sentiment analysis, on the other hand is about determining the subjectivity, polarity like positive, negative or neutral and polarity strength. Opinion can be fetched in two different ways. One is of questionnaire where the questions and its answers will be very relevant o product and its feature. So it is easy to make score and finalize the outcome whereas unstructured review that may usually include feedback in the form of text and images from various social monitoring tools and online shopping sites. In market each product may be introduced on the basis of some latest features they hold and they can either uplift or downsize the demand of that product. Researchers have reported lots of approaches towards feature extraction and they are broadly classified as two types like supervised and unsupervised in systematic method. In this paper, through systematic approach the feature words of Samsung Tablet PC are identified from unstructured reviews. And in automated approach some of the common tools available in market for free of cost is used in identifying sentiments.

2. Related work

There are so many approaches and methodology followed by researchers in order to undergo sentiment classification. Sentiment analysis for the emotional preference of online comments has gained great achievement since it was raised up by Pang etal.2002 [14] and studied in-depth. A common type of opinion summarization is Aspect-based Opinion Summarization. It contains aspect feature identification, sentiment prediction, and summary generation. Hu and Liu et al [15][16] attempt to find features by using NLP-base techniques, they perform POS tagging and generate n-grams, for sentiment prediction, they choose some seed sentiment words. Popescu and Etzioni et al [17] investigated the same problem. Their algorithm requires that the product class is1597. The algorithm only reckon noun/noun phrase as the candidate features. It determines whether a noun/noun phrase is a feature by computing the Point-wise Mutual Information (PMI) score between the phrase and class discriminators, e.g., "of xx", "xx has", "xx comes with", etc., where xx is a product class. But it calculates the PMI by searching the Web. Querying the Web is time-consuming. Khairrullah khan et al [6] has suggested that Brill Tagger or CST tagger can be used to identify which category of words can be features. Hsiang Hui Let et al [5] recommends the observation that there are relations between the product features or aspects and opinion words. Thelwall, M., Buckley et al [7] has given some confidence that SentiStrength is a robust algorithm for sentiment strength detection on social web data and is recommended for applications in which exploiting only direct affective terms is important. Alekh Agarwal et al [13] tried to focus on adjectival word that increase the polarity score and gained accuracy of about 61.1% compared to non-adjectival word of 55.93%. Hai-bing ma et al [8] suggests a typical approach first to identify k positive words (such as excellent, awesome, fine) and k negative words (such as bad, poor). Later to get the sentiment weight of a word, we should subtract the associated weight with k negative words, These 2k words are often selected by experts. This is a kind of supervised learning algorithm where 2k words have to been taken for further classification. Dipali V.Talele et al [1] used Na ïve bayes classification with tf-idf for summarizing review and its accuracy is 47.8% which is higher than of SVM with 27.0%.

3. Systematic Approach

Feature Extraction and polarity detection is one of the very interesting as well as difficult tasks in opinion mining. Sentiment strength detection is one which predicts the strength of positive or negative sentiment within a text. We tried a very common approach for sentiment analysis by selecting a machine learning algorithm and a method of extracting features from texts and then train the classifier with a human-coded corpus. Corpus is a large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based.

3.1. Feature Extraction

The term stemming refers to the reduction of words to their roots. Porter's stemming algorithm can be used to remove stop words. Brill Tagger, Tree Tagger, CST Tagger are the tool used for annotating text with part-of-speech (POS). POS also called grammatical tagging is the process of marking up a word in a corpus as corresponding to a particular part of speech, based on both its definition, as well as its adjacent and related words in a phrase, sentence or paragraph.

3.2. Steps in Pre-processing

A parser processes input sentences according to the productions of a grammar, and builds one or more constituent structures that conform to the grammar. The assumption here is that positive words will tend to co-occur with other positive words more than with negative words, and vice-versa. Here, in Fig 1.(a), Brill tagger is used to find part of speech of each commented sentence. Most of the adjective words bear sentiment, so they are highlighted in Fig 1.(b). Followed by Fig.1.(c) were feature words are visualized using TagCrowd.

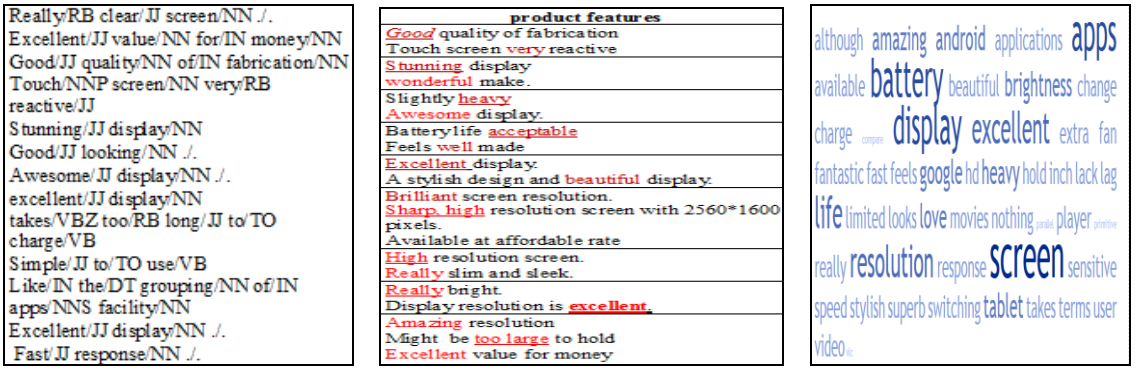

Fig.1. (a) POS-Brill Tagger; (b) sentiment bearing words-highlighted; (c) Most frequent occurring feature words-Tag Crowd

3.3. Product Aspects

TextStat is a freely available which can be used for pattern extraction. The aspect words are nothing but the noun or noun phrases like display, fabrication, response, screen, accessories, applications, battery life, speed, weight, size, price, cost, navigation, connectivity with its number of frequency are retrieved. From which most occurring aspect words are taken and they are clustered manually which is shortlisted in Table 1.

Table 1. A Short List of Features and Its Grouping

screen/display/ touch screen	accessories/ applications	batterylife/ speed	weight/size	price/cost	wi-fi
Good quality of fabrication	No apps installed for office and other applications.	takes too long to charge	Slightly heavy.	Excellent value for money	Connected to hoı hassle at all.
Touch screen very reactive	Limited video formats available for viewing movies etc.	battery life acceptable	Solid but not heavy.	Available at affordable rate	
Stunning display	Movies, pages & content looks good!	not superb battery	It's 10 inch! that's all that matters.	very costly.	
Wonderful make.	simple to use applications	Have nothing to compare the battery life with.	Once you get used to the weight, it'll be alright.	A bit more expensive than other 10 inch tablet PCs.	Bluetooth connection to my Nexus phone was quick and easy and so was wifi tethering.
Good looking.		Very long lasting battery	Might be too large to hold		

3.4. Finding the polarity of opinionated sentence

SentiStrength is a lexicon-based classifier that uses additional linguistic information and rules to detect sentiment strength in short informal English text. For each text, the SentiStrength output is of two integers: 1 to 5 for positive sentiment strength and a separate score of 1 to 5 for negative sentiment strength. For instance, 0 indicates no emotion, 1 indicates not positive, 2 indicates slightly positive, 3 indicates normal positive, 4 indicates positive and 5 indicates very positive. These scales are used because even short texts can contain both positivity and negativity.

3.5. Experimenting Dataset

From online shopping site totally 575 reviews and their sent strength binary value are tabulated and it comes totally of columns 116. So only the snapshot of that table is given in Table 2.

Table 2. Snapshot of Features With Sentistrength Value

Display	**Accessories**	**Battery life**	**Weight**	**Cost**
3,-2	1,-1	1,-2	1,-2	3,-1
1,-3	1,-1	1,-1	1,-1	2,-1
2,-1	1,-1	1,-1	2,-1	1,-3
3,-1	1,-1	1,-1	1,-1	1,-2
3,-1	1,-1	1,-1	1,-1	2,-1
4,-1	2,-1	1,-2	1,-3	1,-2
1,-2	3,-1	2,-1	3,-1	2,-1
1,-1	2,-1	3,-1	2,-1	1,-3
1,-1	1,-3	3,-1	1,-3	4,-1

3.6. Classification through Tanagra1.4

Tanagra1.4 is free data mining software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. The main purpose of its project is to give researchers easy-to-use data mining software. The second purpose of TANAGRA is to propose to an architecture allowing them to easily add their own data mining methods, to compare their performances. TANAGRA acts more as an experimental platform. By this way, Tanagra can be considered as a pedagogical tool for learning programming techniques to undergo Na ïve bayes classification.

Step 1. Import dataset specified in above excel sheet which consist of sentiment value for each of 575 unstructured reviews detected using sent strength tool.

Step 2. Define status and set parameters as discrete binary values for all the most basic features that are identified.

Step 3. From supervised learning method select Na ïve bayes classifier and set to each of the defined status thru step 2.

Step 4. Set the classification function to true so that the features gets prior distribution to each of class attributes.

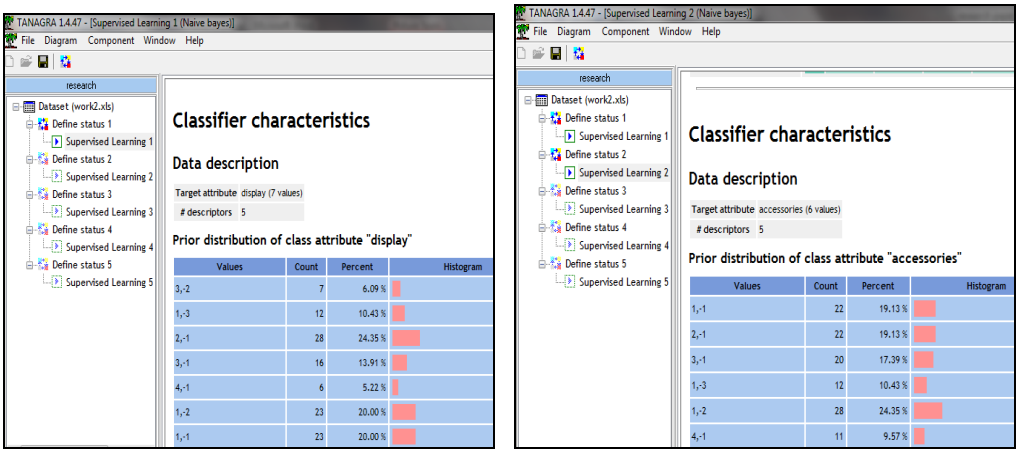

Fig.2. (a) Prior distribution of feature word-display, (b) Prior distribution of feature word-accessories

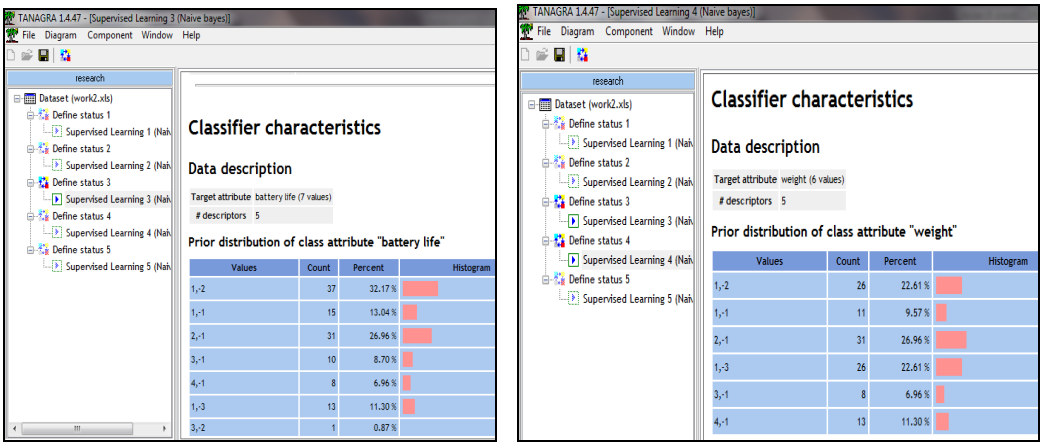

Fig.3. (a) Prior distribution of feature word-battery life, (b) Prior distribution of feature word-weight

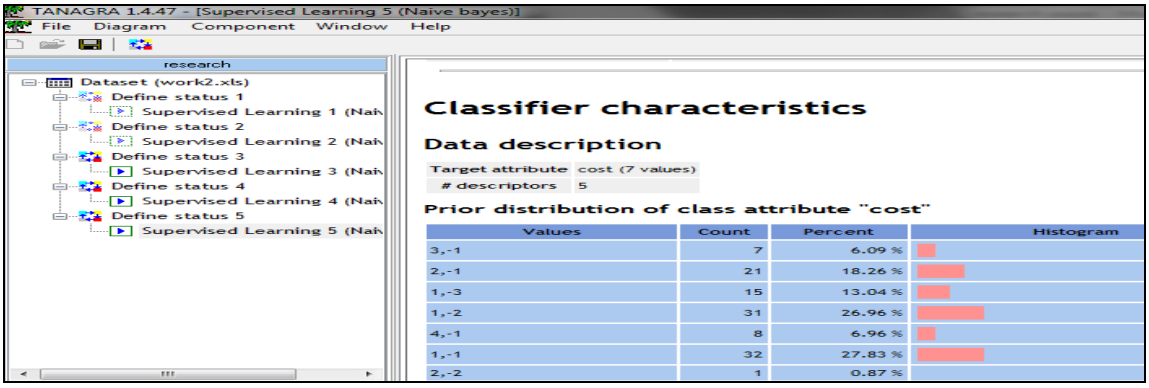

Fig.4. Prior distribution of feature word-cost

3.7. Equations

Inorder to calculate the positive, negative and neutral polarity percentage following is the formula to be used.

positive % = {total number of positive value count} / {total number of comment} X 100 (1)

negative % = {total number of negative value count} / {total number of comment} X 100 (2)

neutral % = {total number of neutral value count} / {total number of comment} X 100 (3)

On using these three formulas the overall polarity distribution of each feature for Samsung Tablet PC is calculated and it is also tabulated in Table 3 and visualized using bar chart in Fig. 5.

Table 3. List of feature and its polarity Distribution

Feature	**% of positive Distribution**	**% of Negative Distribution**	**% of Neutral Distribution**
Display	43.5	36.5	20
Accessories	42.6	44.4	13
Battery life	46	35	19
Weight	45.2	45.2	9.6
Cost	32	40	28

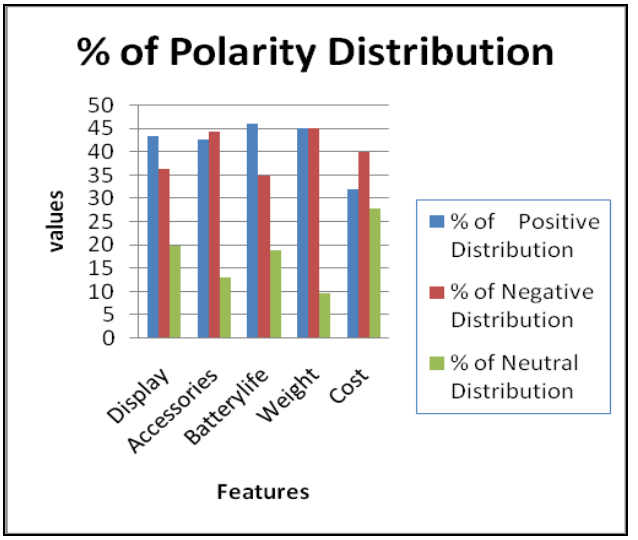

Fig.5. Bar chart indicating polarity Distribution-Samsung Tablet PC

On observing the chart, it is clear that the feature 'batterylife' bears highest positive value and the feature 'cost' indicates very low positive distribution. It is also able to identify that the feature 'weight' scores equal positive and negative distribution. So, with this result the seller can concentrate on promoting the product by revising the cost of tablet PC inorder to promote it in global market and also shall create demand for the product by taking some more steps in making tablet lightweight or handy.

4. Automated Approach

There are huge numbers of automated sentiment analysis tool available in online market. For our proposal, three famous tool were considered and they are also available at free of cost for use.

4.1. Social Mention—track and measure what people are saying about you, your company, a new product, policy or any topic across the Web's social media landscape (100+ social media platforms)

Fig.6. Social mention feedback on Samsung tablet pc

4.2. Tracker—Online reputation and social media monitoring tool to track trends, understand influence, receive alerts and tag sentiment.

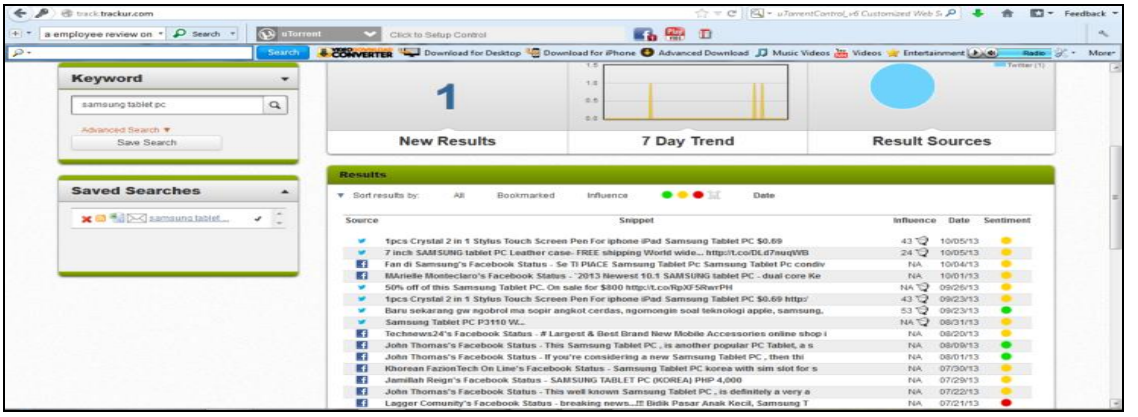

Fig.7. Trackur feedback on Samsung tablet pc

4.3. Opinion Crawl—Simply analyse terms based on a pre-defined glossary, and give highly simplified and unreliable results.

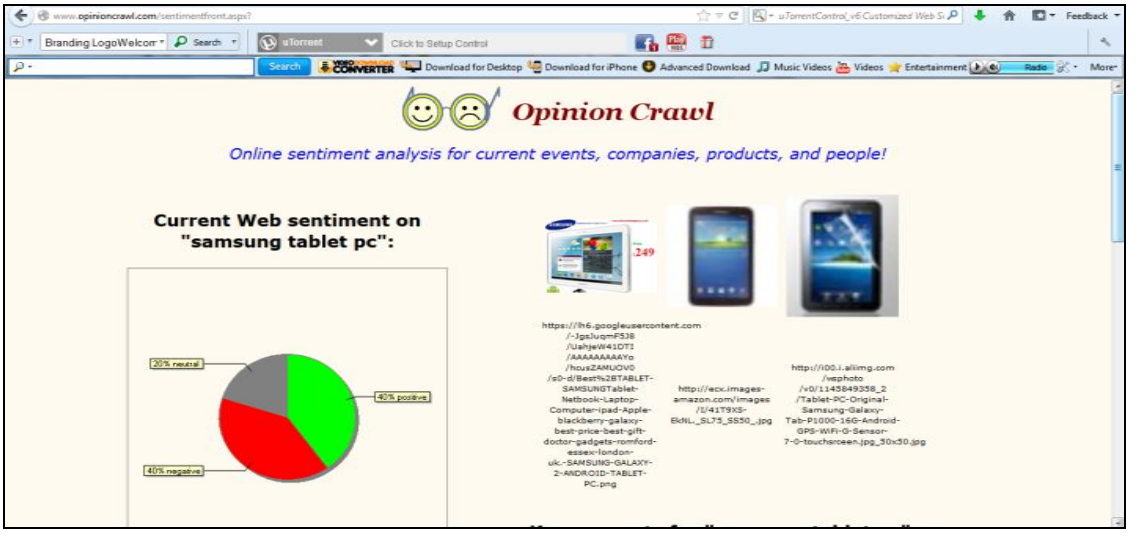

Fig.8. Opinion crawl feedback on Samsung tablet pc

5. Conclusion

The automated tool also captures and visualizes the positive, negative and neutral distribution of a product but, in systematic approach it is able to clearly focus on each and every feature of the product to take necessary steps in promoting branding. Thus the proposed paper differentiates two different approaches of doing sentiment analysis.

References

[1] Dipali V.Talele, Sonal Patil, Extracting and Analyzing Sentiments of the Crowd Using Naive Bayes Classification, Asian Journal of Computer Science and Information Technology, ISSN 2249 – 5126,2013.

[2] V.K. Singh, R. Piryani, A. Uddin, P. Waila, Marisha, Sentiment Analysis of Textual Reviews,Evaluating Machine Learning, Unsupervised and SentiWordNet Approaches, 5th International Conference on Knowledge and Smart Technology (KST), 2013.

[3] Lizhen Liu, Zhixin Lv, Hanshi Wang, Opinion Mining Based on Feature-Level, 5th International Congress on Image and Signal Processing (CISP 2012), 2012.

[4] P. Szabo and K. Machova, Various Approaches to the Opinion Classification Problems Solving, 10th IEEE Jubilee International Symposium on Applied Machine Intelligence and Informatics • January 26-28, 2012.

[5] Hsiang Hui Lek and Danny C.C.Poo, Sentix: An Aspect and Domain Sensitive Sentiment Lexicon, 24th Internation Conference on Tools with Aritificial Intelligence, 2012.

[6] Khairullah khan and Baharum B.Baharudin, Analysis of Syntactic Patterns for Identification of Features from Unstructured Reviews, 4th International Conference on Intelligent and Advanced Systems, 2012.

[7] Thelwall, M., Buckley, K., & Paltoglou, G, Sentiment strength detection for the social Web, preprint of an article published in the Journal of the American Society for Information Science and Technology, 63(1), 163-173, © copyright 2011 John Wiley & Sons, Inc.

[8] Hai-Bing Ma, Yi-Bing Geng, Jun-Rui Qiu, Analysis Of Three Methods For Web-Based Opinion Mining, Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, 10-13 July, 2011.

[9] Jingbo Zhu, Member, IEEE, Huizhen Wang, Muhua Zhu, Benjamin K. Tsou, Member, IEEE, and Matthew Ma, Senior Member, IEEE, Aspect-Based Opinion Pollingfrom Customer Reviews, IEEE Transactions On Affective Computing, Vol.2, No. 1, January –March 2011.

[10] Evgeny A. Stepanov, Giuseppe Riccardi, Detecting General Opinions from Customer Surveys, 11th IEEE International Conference on Data Mining Workshops, 2011.

[11] Shen Jie, Fan Xin, Shen Wen, Ding Quan-Xun, BBS Sentiment Classification Based on Word Polarity, International Conference on Computer Engineering and Technology, 2009.

[12] Khairullah Khan, Baharum B.Baharudin, Aurangzeb Khan, Fazal-e-Malik, Mining Opinion from Text Documents: A Survey, 23rd IEEE International Conference on Digital Ecosystems and Technology, 2009.

[13] Alekh Agarwal and Pushpak Bhattacharyya, Augmenting WordNet with Polarity Information on Adjectives, Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 3–8. c Masaryk University, 2005.

[14] Pang,L.Lee and S.Vaithyanathan,Thumbs up?: sentiment classification using machine Learning techniques. In EMNLP' 02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Association for Computational Linguistics, Morristown, NJ, USA, 79-86,2002.

[15] M. Hu and B.Liu,Mining and summarizing customer reviews. In KDD' 04: Proceedings of the Tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA,168-177, 2004.

[16] M. Hu and B. Liu,Mining opinion features in customer reviews. In AAAI' 04: Proceedings of the 19th national conference on Artificial Intelligence. AAAI Press,2004.

[17] Popescu, Ana-Maria and Oren, Etzioni, "Extracting product features and opinions from reviews," In Proceedings of EMNLP, 2005.

Author(s) Profile

Nithya Ramachandran: She is currently working as Assistant Professor in School of Computer studies(UG) Department at R.V.S College of Arts and Science, Sulur, Coimbatore, Tamil Nadu, India and pursuing Ph.D in part time in the area of Data mining.Her research work focusses on sentiment analysis and pattern mining techniques.

Maheswari: She is currently working as Assistant Professor in School of Computer Studies (PG) at R.V.S College of Arts and Science, Sulur, Coimbatore, Tamil Nadu, India. She has also completed her Ph.d in Computer Science by exploring Image Processing Techniques. Currently she is guiding many Ph.d and M.Phil Scholars and an active member of various editorial boards.

How to cite this paper: R.Nithya, D.Maheswari,"A Contrast Between Systematic and Automated Sentiment Analysis", IJEME, vol.5, no.2, pp.20-29, 2015.DOI: 10.5815/ijeme.2015.02.03