Thailand –Tourism and Conflict. Modeling Sentiment from Twitter Tweets using Naïve Bayes and Unsupervised Artificial Neural Nets

William B. Claster, Malcolm Cooper,
School of Asia Pacific Management
Ritsumeikan Asia Pacific University
Beppu, Japan
wclaster@apu.ac.jp / cooperm@apu.ac.jp

Philip Sallis
School of Asia Pacific Management
Ritsumeikan Asia Pacific University
Beppu, Japan
wclaster@apu.ac.jp / cooperm@apu.ac.jp

Abstract-- **In this paper we mine over 80 million twitter microblogs in order to explore whether data from this social media initiative can be used to identify sentiment about tourism and Thailand amid the unrest in that country during the early part of 2010 and further whether analysis of tweets can be used to discern the effect of that unrest on Phuket's tourism environment. It is proposed that this analysis can provide measurable insights through summarization, keyword analysis and clustering. We measure sentiment using a binary choice keyword algorithm. A multi-knowledge based approach is proposed using, Self-Organizing Maps along with sentiment polarity in order to model sentiment. We develop a visual model to express a sentiment concept vocabulary and then apply this model to maximums and minimums in the time series sentiment data. The results show actionable knowledge can be extracted in real time.**

Keywords; Tourism, Sentiment Mining, SOM, Twitter, Social Networks, Semantic Web, Text Mining.

I. INTRODUCTION

Online reviews are becoming an increasingly useful and important information resource. As a result, automatic review mining and summarizing has become an important research topic. Horrigan [1] noted that 81% of Internet users have done online research on a product at least once. Another study conducted by Comscore[2] in 2007 revealed that user reviews has a significant influence on customers' purchase, and that reviews generated by fellow customers have a greater influence than those generated by professionals. Vendors can ascertain, both from current customers and potential customers, information that previously may have beyond their reach. Subjective information related to objective characteristics such as a customer's subjective view of product design may be available in blogs and review sites. In addition knowledge pertaining to political and policy issues is communicated across the web and may be utilized to formulate policy that is attractive and rooted in the public interest [3], [4]. During the 2006 elections in the United States, 34% campaign internet users use the Internet to gather information and exchanged views about the 2006 elections online [5].

Accessing and measuring the sentiment accumulated in the vast store of blogs, online publications, social network media (such as Facebook) and microblogs such as Twitter can yield tangible and actionable information for Business, Marketing, Social Sciences, and government. Knowledge of consumer opinions, public attitudes, and generally the "wisdom of crowds" can yield highly valuable information. As the World Wide Web has developed, considerable decision making power over the consumption of discretionary products like tourism has been transferred from suppliers to consumers; there is thus a real need to improve market intelligence and market research for private and public tourism organizations to facilitate timely consumer decision making. Here we explore the development of user generated content about the characteristics and value of destinations through analyzing the use of Twitter and seek to answer whether tweets can be mined for industry intelligence.

Twitter posts may be regarded as conversational microblogs. We propose that these microblogs can be used as a source of sentiment expression and for this study we focus on sentiment expressed towardsthe travel destination of Phuket a travel resort region in Thailand.

II. BACKGROUND

Much work has recently been undertaken in sentiment mining over the last few years. Pang and Lee give and excellent review [6]. Work has been done specifically on sentiment mining movie reviews [7] and even more recently work has been carried out on mining tweets from Twitter[8],[9]. It has been suggested that the limited size of the twitter microblogs (140 words) may boost text mining efficiency as subject ambiguity is reduced in the these shorter expressions, although conclusive research has not been conducted[1]. Janson et al[10] analyze 150,000 microblogs from Twitter in terms of frequency, timing, and contents of tweets within a corporate account and focus on the sentiment expressed towards products produced by that company . A Go et al. build an algorithm to classifysentiment within Tweets as positive or negative[2]. They achieve accuracy as high as 81%. A Kennedy et al.

---

[1] http://lifeanalytics.blogspot.com/search/label/twitter

[2] A Go, L Huang, R Bhayan, "Twitter Sentiment Analysis", wwwnlp. stanford.edu, 2009.

978-0-7695-4262-1/10 $26.00 © 2010 IEEE
DOI 10.1109/CIMSiM.2010.98

examine the effect of "valence shifters" such as negations, intensifiers, and diminishers in modifying sentiment. Then they extend the study by looking beyond unigram features to bigrams [11]. L. Zhuang et al.[12] mine the IMDB database to derive sentiment scores. They report that the methodology compares favorably to an earlier methodology developed by Minquig Hu for summarizing customer reviews [9].

A binary choice algorithm is employed to produce a one dimensional measure of sentiment (negative to positive) and plot this measure over time. Then focusing on maximum and minimums in the variation over time Weemploy a Kohnen Self Organizing Map (SOM) to visualize multiple characteristics at these peaks and valleys in this time series. The analysis differs from other sentiment mining in that instead of evaluating the sentiment across a one-dimensional scale (positive/negative) we analyze the sentiment on an organic multidimensional space derived from Self Organizing Maps (SOM) [13],[14],[15],[16].

III. DATA SET DESCRIPTION

Using twitter's API, we collect and store tweets in a database. We used a data set of 70,570,800 tweets comments or about 20.42 GB data drawn from Oct. 30, 2009 to May 21, 2010. Although data collected includes time, date, user name, user followers, whom the user is following, location, and the textual comment, we make use of the date and textual comment in this study. The tweet comments were then filtered such that only those containing references within the textual comment to the keywords'Phuket' or 'Bangkok' were included.

IV. Modeling Polarity –Positive/Negative Sentiment.

We first employed a keyword based algorithm to measure sentiment. The algorithm employed binary choice which was then modeled against a gold standard using a naïve bayes algorithm. This has been shown in text mining literature to be an optimal approach to polarity mining [17].This was then tested against the gold standard measurement and achieved r = 0.63 correlation. Using this combination of binary choice and naïve bayes we were able to plot the polarity of the sentiment expressed in the Tweets over the 200 day period towards both Bangkok and Phuket (see figures2 and 3).

From these figures we see that sentiment towards the both areas during this period varied but was always positive. We also see there is a downward trend in the sentiment on Bangkok (figure 2) but not in Phuket (figure 3).

From this initial visualization of the sentiment expressed in twitter we may hypothesize that although Thailand has experienced significant political turmoil from 2008 to present and in particular in a 10 week period from March to May 2010, it may be that Phuket was remote enough so as to be largely insulated from serious damage to its tourist industry.

Additionally examination of the content of tweets may reveal and reflect the nature of concerns during this period towards Bangkok and the events there.

V. Preprocessing Twitter Data For ANN SOM.

In order to visualize the data beyond a one-dimensional positive/negative measure a series of self-organizing mapswere developed. The preprocessing of the data prior to input into the SOM algorithm consisted of two steps. First a series of steps were followed to provide a vector space model. Secondly, these tokens were filtered to produce a lexicon of travel sentiment related words.

A. Vector Space Model

The following are the major steps adopted to produce the vector space model whereby a matrix of meaningful sentiment descriptors whose weights were determined based on their presence in a twitter comment and their overall presence in the corpus of tweets are constructed on the set of keyword filtered tweets.

1. Consolidate all tweets into a single corpora perform the following steps for each corpus.
2. Remove stop words.
3. Reduce verbs to lemmas using a simple nonaggressive stemming algorithm.
4. Discard rare words by giving a lower limit to the frequency of accepted words equal 3. The weighted movie descriptor frequency matrix is calculated using vectorization as developed by Salton [18] in equation (1).

$$w\_i = tf\_i*\log(D/df\_i) \qquad (1)$$

5. Take the transpose of the matrix obtained in step 4 in order to cluster the words instead of the comments.
6. The matrix obtained in step 5 was then trimmed by including only those attributes (rows) which were elements of the sentiment lexicon described above.

B. Sentiment Lexicon

In order to improve the efficiency of the modeling estimation procedure we sought to refine the dictionary of tokens obtained above. In earlier work we used domain expertise to develop a lexicon of words related to movie sentiment. Here we sought a combination of automated pruning based on word frequencies and manual deletion. Manual deletion consisted of removal of most time related words, all non-English words, words relating to size, numbers, and certain twitter specific tokens like #, @, and rt. A lexicon of 213 subjective words which may relate to travel sentiment was constructed.

VI. ANN NEURAL NETWORK SELF-ORGANIZING MAP FOR MULTI-DIMENSIONAL SENTIMENT VISUALIZATION.

Using the matrix resulting from step 6 in the preprocessing mode, we analyzed the data using a SOM. We employed a recognized software package called

Viscovery SOMiner. Viscovery SOMiner is a desktop application for explorative data mining, visual cluster analysis, statistical profiling, segmentation and classification based on self-organizing maps (SOMs) and classical statistics in an intuitive workflow environment.

Our strategy was to examine the entire set of tweets and then just the tweets that include the keyword 'Phuket'. We next utilized the one dimensional sentiment measure to partition these two datasets based on extreme values of the sentiment measure. We broke down these two sets of data so as to identify the tweets in periods of peak sentiment (as derived from the sentiment analysis in section V) and valley or minimum sentiment. In particular we identified a relative peak in the sentiment of the Thailand data between February 15th to March 4th and a relative minimum in sentiment between March 5th and March 25th. For the Phuket data we found a peak from November 29th to December 20th and then a long valley from January 17th to March 5th.We analyzed each of these periods using self-organizing maps. These maps are displayed in figures1 and 4 – 8.They cluster the lexicon derived in section IV.

VII. HYBRID VIEW: APPLYING A MULTI-DIMENSIONAL ANALYSIS ON THE UNIDIMENSIONAL MODEL.

The self-organizing maps may be viewed as providing multi-dimensional view of the sentiment beyond that of the one dimensional numeric polarity variable described above.The SOM map shows the filtered data set clustered according to their overall similarity. These present the lexicon separated into clusters.The peaks and valleys in the sentiment time series (figures 1 and 2) may suggest shifts in the balance of components across the multi-dimensional sentiment view. We thus trained the SOM on time frames where these peaks and valleys occur. Figures 4,5,7 and 8 show these results.

A. SOM of all Thailand tweets over entire period of data collection: Oct. '09 to May '10.

Figure 1 shows the distribution of sentiment for tweets taken over the entire 7 months for tweets containing either of the query words 'Bangkok' or 'Phuket'. We see that there are clusters which seem to refer to the political tension and others that do not. 'Redshirts' are prominent in the Kohonen map and there are clusters relating to 'curfew', 'military', and 'wounded', but there are also a significant number of tweets that pertain to positive aspects of tourism.

B. SOM of all Thailand tweets between Feb15 Mar4 - Sentiment Peak modeled using 15 Clusters.

Figure 4 focuses on the tweets where sentiment peaked. In this Kohonen map the conversation is considerably freer of references to the political unrest with the exception of one cluster that references security concerns. When looking deeper into the 'Resort' cluster we see it subsumes concepts like nightlife, golf, beaches, pool, and Samui. (see table 1).

Table 1. Word decomposition of the resort, work, cheap, deal, and play clusters with relative weights.

Resort		Mall	Trip
Phuket	June	Australia	discount
golf	king	bad	fly
money	nightlife	full	offers
restaurant	villas	mall	trip
beaches	krabi	vacation	**Grand**
bank	photos	weekend	boutique
thb	march	**Tourists**	friends
safe	pool	visit	grand
samui	good	tourists	spa
rates	airways	love	temple

C. SOM of all Thailand tweets betweenMarch 5 and March 25 -Sentiment valley modeled using 9 Clusters.

During this period (figure 5) the sentiment seemed to fall precipitously and the clusters point to violence and war and moreover the Kohonen map shows Bangkok surrounded by the redshirts. Inspecting the clusters more deeply confirms that the conversation here was focused on the unrest. Looking deeper into the clusters (table 2) we see the Redshirts cluster subsumes followers, troops, and anti-government. The Wars cluster includes the concepts of military, money, security. The BBC cluster includes the concepts of wounded, tv, explosions, fire, and mafia.

D. SOM of Phuket specific over all 7 months, short peak and longer valley.

Figures 6, 7, and 8 show tweets relating to Phuket. Figure 6 shows a 21 cluster Kohonen map for the entire period. Figure 7 shows a map when there is a relative maximum in the sentiment time series and Figure 8 shows a map when there is a longer valley in the sentiment time series. As was noted earlier, the time series for Phuket sentiment suggests that although there is variation in the sentiment over time, there is no downward trend. Examination of theKohonen maps indicates that focus among tweets pertaining to Phuket do not seem to be concerned with the same security and political matters that were prominent in the tweets that included Bangkok. This suggests that although there was extreme violence in the capital the tourism economy outside of Bangkok may have been insulated from the detrimental effects of the situation. This hypothesis is supported by a reportfrom the Tourism Authority of Thailand which states that “About 30 per cent of the trips to Phuket planned by Chinese travellers have been cancelled due to the red shirts' demonstration. However, tourists from other markets have made no changes to their travel plans[3]". As the tweet analysis in our study relates to tweets in English, this suggests that the hypothesis may be reasonable.

[3] www.nationmultimedia.com/home/2010/03/03/business/Chinese-cancelling-Phuket-Trips-Amid-redshirt-prot-30123799.html

VIII. CONCLUSIONS AND FURTHER WORK.

Data mining applied to social media can explain and reflect both numerically and visually societal trends. Inparticular we test, in real time, the effect of political instabilityin Bangkok on nearby tourism. In this paper we demonstrate a hybrid methodology for tracking sentiment towards a well-known resort area, Phuket, in Thailand. A Naïve Bayes algorithm and binary choice keyword are brought together to create a single dimensional measure of sentiment embedded in tweets from Twitter. Then a multidimensional visualization is trained on the resulting time series data to provide a richer understanding of the data. Using this approach we find practical information that can assist businesses in understanding concerns of tourists and travelers. The analysis implies that further work in this area may provide researchers and business with a rich source of sentiment knowledge should the methodology be focused on other resorts and regions.

Table 2. Time series of sentiment tracked through tweets from a data set of 70,570,800 tweets comments or about 20.42 GB data drawn from Oct. 30, 2009 to May 21, 2010

Redshirts	Wars
followers	govt
troops	guys
anti-government	military
rally	money
Phuket	security
redshirts	wars
hotels	work
protests	

Violence	Holidays	BBC
army	fly	wounded
violence	holidays	bbc
district	live	bad
china	luxury	explosions
super	road	TV
bookings	special	bank
forces	thb	British
march	villas	fire
armored	waiting	mafia
krabi		pictures

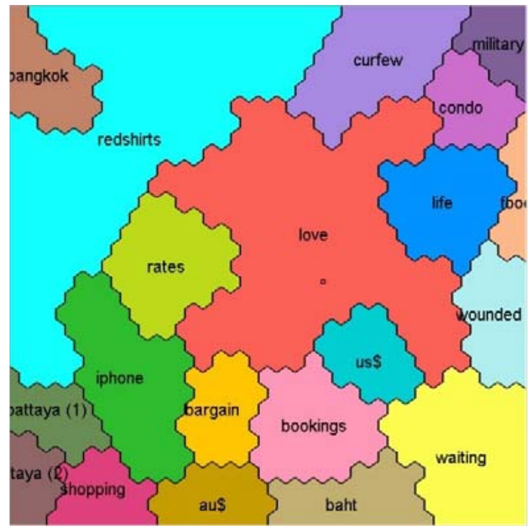

Figure 1.SOM of all Thailand tweets over entire period of data collection: Oct. '09 to May '10.

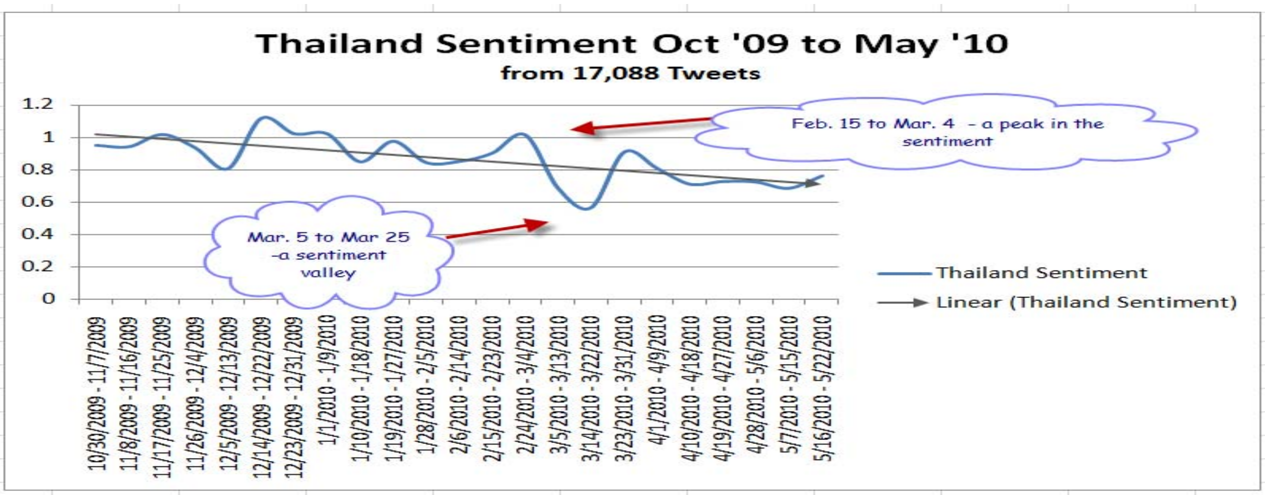

Figure 2. Time series of average sentiment tracked through tweets from a data set of 70,570,800 tweets comments or about 20.42 GB data drawn from Oct. 30, 2009 to May 21, 2010. Tweets are selected if they contain either the query word 'Bangkok' or 'Phuket'

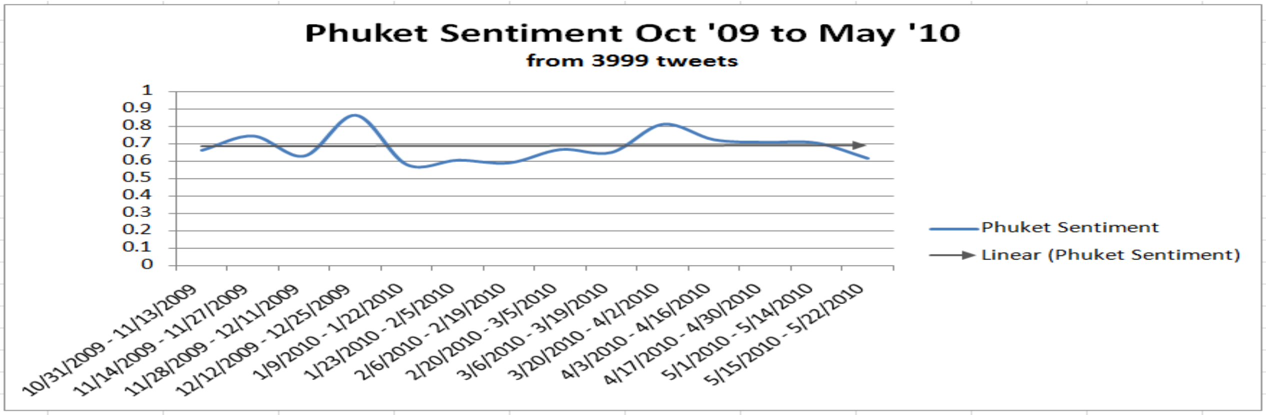

Figure 3. Time series of average sentiment tracked through tweets from a data set of 70,570,800 tweets comments or about 20.42 GB data drawn from Oct. 30, 2009 to May 21, 2010. Tweets are selected if they contain the query word ‘Phuket’

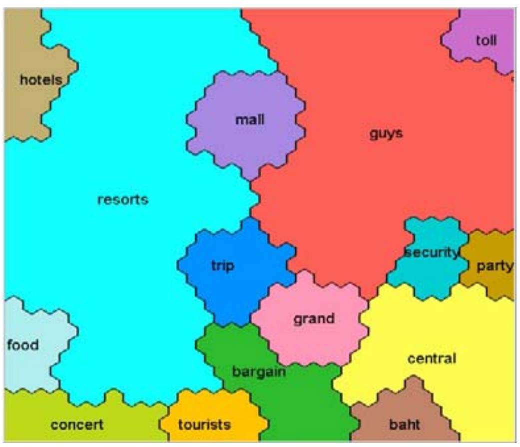

Figure 4.SOM ThailandFeb15 - Mar4; sentiment peak.

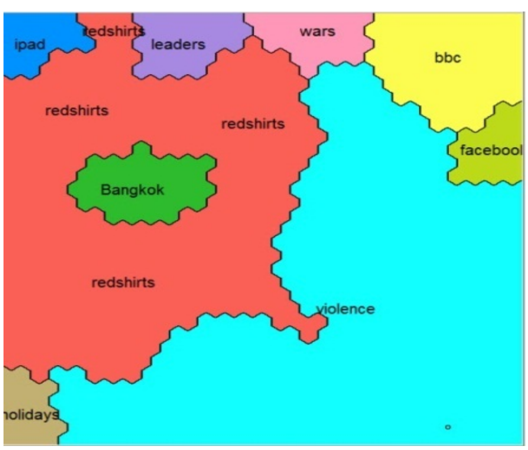

Figure 5.SOM March 5 - March 25;sentiment valley.

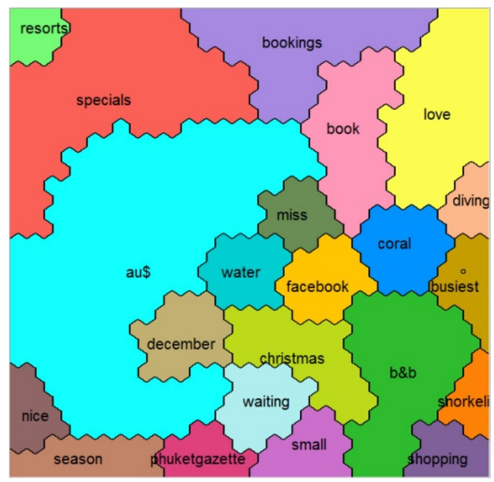

Figure 6. SOM Phuket tweets Oct. ’09 – May ‘10

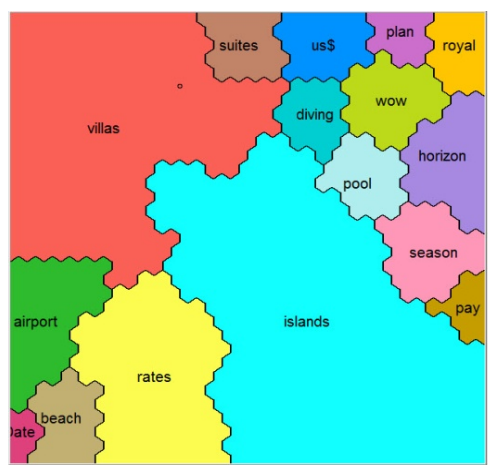

Figure 7. SOM Phuket tweets of relative maximum of sentiment for period 3/13 – 3/31.

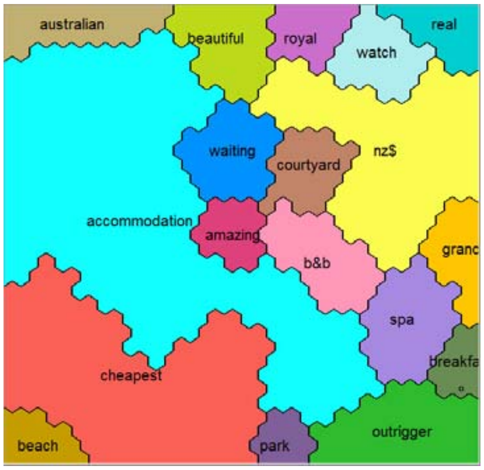

Figure 8.Tweets from long valley period –Jan. to Mar.

REFERENCES.

[1]Horrigan J (2008), Online Shopping, Pew Internet & American Life Project Report

[2]comScore (2007), Online consumer-generated reviews have significant impact on offline purchase behavior." Press Release, Retrieved from http://www.comscore.com/press/release.asp?press=1928

[3]C. Cardie, C. Farina, T. Bruce, and E. Wagner, "Using natural language processing to improve eRulemaking," in Proceedings of Digital Government Research (dg.o), 2006.

[4]N. Kwon, S. Shulman, and E. Hovy, "Multidimensional text analysis for eRulemaking, in Proceedings of Digital Government Research (dg.o), 2006.

[5]Rainie L. and Horrigan J. (2007), Election 2006 online. Pew Internet & American Life Project Report.

[6]B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, nos. 1–2, pp. 1–135, 2008.

[7] B Schuller, J Schenk, G Rigoll, T Knaup, I Pingsta, "“The Godfather” vs. “Chos”: Comparing Linguistic Analysis based on On-line Knowledge Sources and Bags-of-N-Grams for Movie Review Valence Estimation", 10th International Conference on Document Analysis and Recgnition, 2009.

[8]Phelan, O., McCarthy, K., and Smyth, B. 2009. Using twitter to recommend real-time topical news.In Proceedings of the Third ACM Conference on Recommender Systems (New York, New York, USA, October 23 - 25, 2009).RecSys '09. ACM, New York, NY, 385-388. DOI= http://doi.acm.org/10.1145/1639714.1639794[9]Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of ACM-KDD 2004, pp.168-177.

[10]B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury.Twitterpower: Tweets as electronic word of mouth. JASIST, 2009.

[11] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, vol. 22, pp. 110–125, 2006.

[12]L. Zhuang, F. Jing, X.-Y. Zhu, and L. Zhang, "Movie review mining and summarization," in Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM), 2006.

[13]Krista Lagus, Samuel Kaski, and Teuvo Kohonen (1997) Mining massive document collections by the WEBSOM method, Helsinki University of Technology, Neural Networks research center, Finland.

[14]TimoHonkeda (1997) Using Self-Organizing Maps in NaturalLangue Processing, Helsinki University of Technology, NeuralNetwork research Center, Finland.

[15]B.H.ChandraShekar, Dr.G.Shoba (2009) Classification of Documents Using Kohonen's Self-Organizing Map, International Journal of Computer Theory and Engineering, Vol. 1, No. 5, December, 2009.

[16]Shanmuganathan, S., P. Sallis and A. Narayanan (2009) Unsupervised Artificial Neural Nets for Modeling the Effects of Climate Change on New Zealand Grape Wines, 18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009.

[17]Jindal, N. and Liu, B. 2006. Identifying comparative sentences in text documents.In Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, USA, August 06 - 11, 2006).SIGIR '06. ACM, New York, NY, 244-251. DOI= http://doi.acm.org/10.1145/1148170.1148215[18] Salton, G. and M. J. McGill (1983). Introduction to modern information retrieval.McGraw-Hill.ISBN 0070544840.