

Unsupervised Artificial Neural Nets for Modeling Movie Sentiment

William B. Claster and DINH Quoc Hung

School of Asia Pacific Management
Ritsumeikan Asia Pacific University
Beppu, Japan

wclaster@apu.ac.jp / dqhung215@yahoo.com

Subana Shanmuganathan

Geoinformatics Research Centre
School of Computing and Mathematical Sciences
Auckland University of Technology, New Zealand.
e-mail: subana.shanmuganathan@aut.ac.nz

Abstract-- Sentiment mining aims at extracting features on which users express their opinions in order to determine the user's sentiment towards the query object. Movie sentiment in Twitter provides an excellent base upon which to evaluate sentiment mining methodologies both because of the pervasiveness of discussions devoted to movie topics and because of the brevity of expression induced by twitter's 140 word limitation. In this paper we explore movie sentiment expressed in Twitter microblogs. A multi-knowledge based approach is proposed using Self-Organizing Maps and movie knowledge in order to model opinion across a multi-dimensional sentiment space. We develop a visual model to express this taxonomy of sentiment vocabulary and then apply this model in test data. The results show the effectiveness of the proposed visualization in mining sentiment in the domain of Twitter tweets.

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

I. INTRODUCTION

With the flourishing of the Web, online reviews are becoming an increasingly useful and important information resource. As a result, automatic review mining and summarizing has become a hot research topic. With the explosion of Web2.0 platforms, user participation and user feedback has been sought out and encouraged. Users express opinions through review sites such as Amazon.com, IMDB (<http://www.imdb.com>), and Epinions, as well as through blogs, discussion forums, peer-to-peer networks, user feedback and comments, and various types of social network sites. These reviews are informative both to vendor and readers. Consumers, with unprecedented reach, may express their opinions and experiences. Vendors can ascertain, both from current customers and potential customers, information that previously may have been 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 [1], [2].

In this paper we analyze twitter posts which can be regarded as microblogs. We propose that these microblogs can be used as a source of sentiment expression and we focus on sentiment towards recently released movies as our test arena. The analysis differs from other sentiment mining in that instead of evaluating the sentiment across a unidimensional scale (positive/negative) we analyze the sentiment on an organic multidimensional space derived from Self Organizing Maps (SOM) [3],[4],[5],[6]. We calculate a SOM and use it as a model to visualize sentiment.

II. BACKGROUND

Much work has recently been undertaken in sentiment mining over the last few years. Pang and Lee give an excellent review [7]. Work has been done specifically on sentiment mining movie reviews [8],[9],[10] and even more recently work has been carried out on mining tweets from Twitter. 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 these shorter expressions, although conclusive research has not been conducted¹. Janson et al 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 [11]. A Go et al. build an algorithm to classify sentiment within Tweets as positive or negative. They achieve accuracy as high as 81%². Movie sentiment classification has also drawn the interest of researchers recently. A Kennedy et al. 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 [12]. L. Zhuang et al. mine the IMDB database to derive sentiment scores [8]. They report that the methodology

¹ <http://lifeanalytics.blogspot.com/search/label/twitter>

² A Go, L Huang, R Bhayan, "Twitter Sentiment Analysis", www-nlp.stanford.edu, 2009.

compares favorably to an earlier methodology developed by Minquig Hu for summarizing customer reviews [13]. In our work we employ an artificial neural network to model sentiment expressed via tweets about movies.

III. DATA SET DESCRIPTION

Using twitter's API, we collect and store tweets in a database. We used a data set of 25,570,800 tweets comments or about 3.45 GB data drawn from Dec. 1, 2009 to January 31st, 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 our set of 10 movies (see table 1) were included. The tweets were divided into 2 data sets; a training set and a testing set. Furthermore we selected tweets based on the movie release date. For example, to collect tweets containing the movie "Crazy Heart", only those tweets published after December 6th, 2009 are included since the movie premiered in the US on that date. Table 1 shows the 8 movies for which tweets were collected for the training set and the 2 movies for which tweets were collected to test the model. We further filtered the tweets according to an aggregated collection of words that form a lexicon of "movie sentiment descriptors".

IV. SENTIMENT LEXICON

In order to improve the efficiency of the modeling estimation procedure we incorporated domain expertise through the aggregation of a movie language expression list. A lexicon of subjective words which may relate to movie sentiment was constructed. The words were gathered from various websites including About.com and websites devoted to English language education. A list of 138 words which express

TABLE 1. MOVIE TITLE AND NUMBER OF ASSOCIATED TWEETS. 8 MOVIES SELECTED AND TWEETS WERE COLLECTED CORRESPONDING TO A 1 MONTH POST RELEASE PERIOD TO YIELD THE TRAINING SET.

	Movie Name	Number of Comments
Training set	A Serious Man	3444
	District 9	1637
	New Moon	11031
	Precious	8141
	Inglorious Bastards	442
	The Hurt Locker	1257
	Up In The Air	11320
	Zombie Land	1718
Test Set	The Blind Side	1243
	Crazy Heart	758

sentiment and which are commonly used to describe movies was prepared –words such as, absurd, adventurous, agile, amusing, animated, annoying, awesome, beautiful, big-budget, bland, blockbuster, bloody, ... and so on. These words would be used to subset the dictionary constructed through vector space model described below.

V. PRE-PROCESSING

The following are the major steps adopted to produce 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 collected on a particular movie.

- 1) Consolidate all tweets into two separate corpora as shown in table 1 and perform the following steps for each corpus.
- 2) Remove stop words.
- 3) Reduce verbs to lemmas using a simple non-aggressive 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 [14] in equation (1).³

$$w_i = tf_i * \log(D/df_i) \quad (1)$$

- 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.

VI. SELF-ORGANIZING MAP MODEL TO DEVELOP MULTI-DIMENSIONAL MEASURE OF MOVIE SENTIMENT.

Using the matrix resulting from step 6 in the preprocessing mode, we analyzed the data using 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. The resultant map is shown in figure 1. The SOM map shows the filtered data set clustered according to their overall similarity. This display

³ Also referred to as a bag of words model.

shows that the comment lexicon was separated into 20 clusters with nominations such as 'Charming, 'Blockbuster, 'Masterpiece' etc. as the subjects of the clusters. These nominations use a simple naming process whereby the most frequently expressed sentiment in a cluster is deemed the name of the cluster.

We can also view the corresponding attribute maps visualized as layovers of the original SOM. These are shown in figure 2. From this view we can see, for example, that the attribute 'grip' (stemmed from 'gripping') falls primarily in the cluster nominated as Masterpiece although it does spread somewhat to the neighboring clusters with nominations of Blockbuster, Fantasy, Sensitive, and Towering. This illustrates a fundamental property of Self-Organizing Map algorithm, - it is a topologically preserving mapping from a high-dimensional space to a lower-dimensional viewable space [15].

Figure 7 shows the clusters expressed as a bar graph. It shows how the test movies compare for each particular sentiment. For example we can see that along the Masterpiece dimension "Up In the Air" and "Precious" score high, whereas "Inglorious Bastards" scores lowest.

VII. APPLYING THE MODEL -THE BLIND SIDE AND CRAZY HEART

Our first test was on the movie "The Blind Side". We consider whether there are any tokens in the set of (training) tweets that are not contained in the (test) dictionary and vice versa. Of course, some tokens from the (training) dictionary may occur in these test tweets and some may not. For example, the token

'violent' does not occur in the test set, whereas the token 'great' does. This is illustrated in figures 3 and 4. An additional problem may occur when applying the model to a new movie. It may be that tokens within the corpus of tweets about the new movie may not be a subset of the tokens in the dictionary of tokens associated with the model. However in both "The Blind Side" and "Crazy Heart" no new words are necessary for the model.

Next we consider to which clusters "The Blind Side" belongs. We recognize that tweets pertaining to this movie will contain many tokens, and thus it may have a presence in many clusters."The Blind Side" will have a vector representation in the overall token space. This representation is a function of the frequency of each token in the corpus of The Blindside Tweets. We ask which are the primary clusters to which this movie is distributed? This can be explored visually in a SOM as shown in figure 5. From this map we can see that all comments about "The Blind Side" express sentiment across the dimensions of Thoughtful, Good, Comely, Charming, Beautiful, Enjoyable, Blockbuster, and Funny, but primarily these comments have presence in the Thoughtful, Good, and Comely dimensions. With this analysis we can infer these characteristics about this movie without either watching it or reading all the relevant twitter comments.

We can apply a similar analysis to the second test movie -"Crazy Heart". Figure 6 shows the application of the SOM model to "Crazy Heart". In this case the sentiment appears to be spread primarily across the dimensions Good and Enjoyable, however there are elements of Fantasy, Beautiful, and Comely.

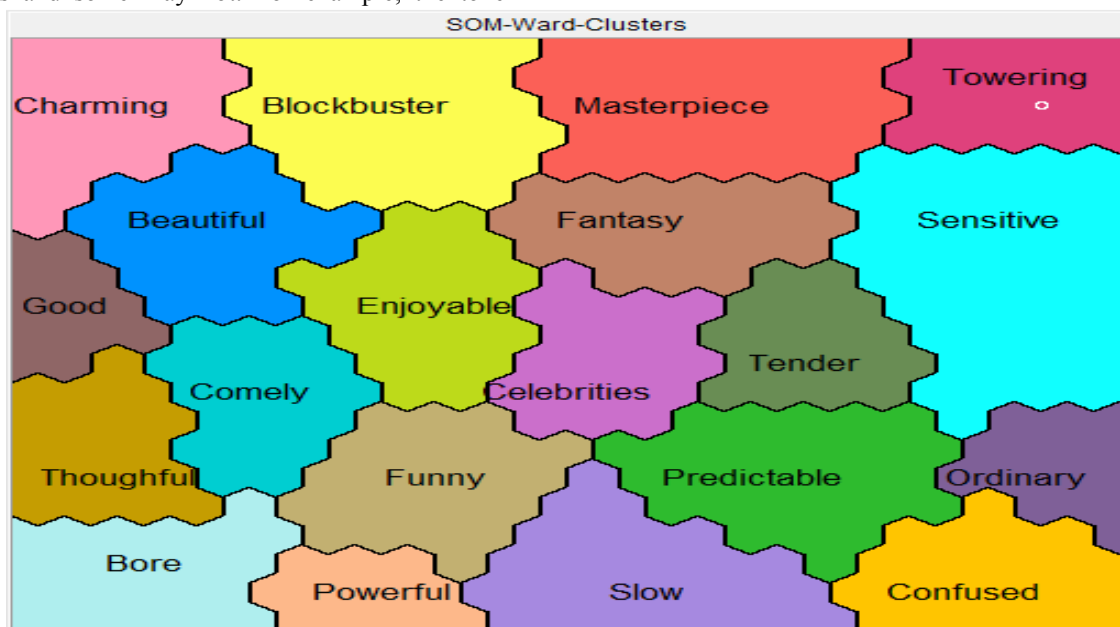


Figure 1. SOM derived from tweets on training set of 8 movies. Each cluster is considered a dimension upon which movie sentiment is measured.

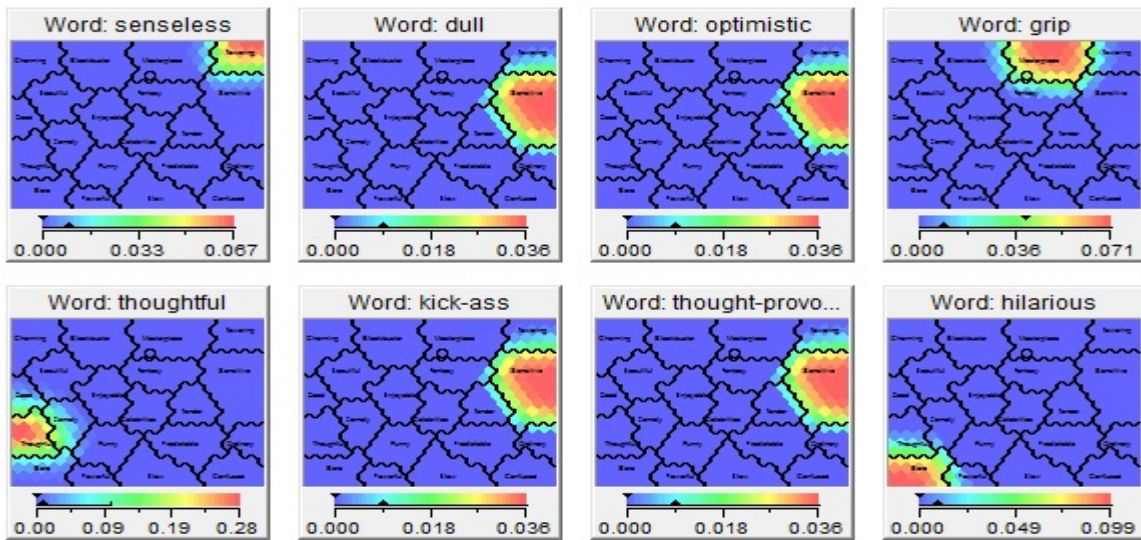


Figure 2. Component maps showing a subset of the attributes ('senseless', 'dull', etc.) and their location within the clusters.

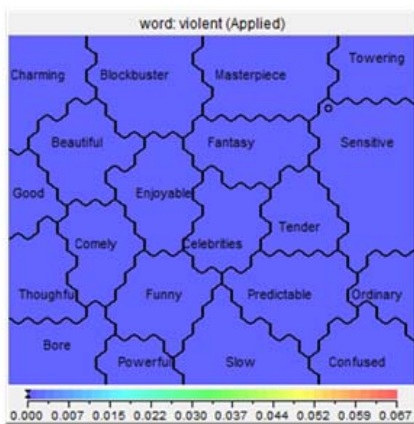


Figure 3. The word violent does not appear in any of the comments about The Blind Side.

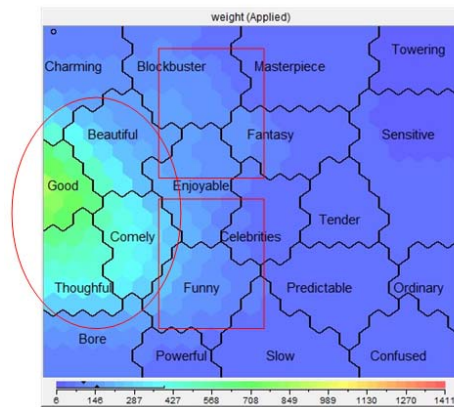


Figure 5. SOM model applied to The Blind Side.

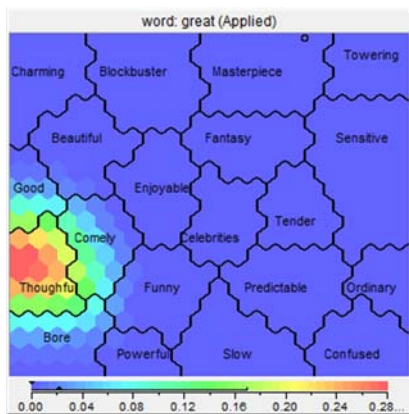


Figure 4. The word 'great' appears in "The Blind Side" comments and it resides mainly in the cluster named "Thoughtful", however it also extends to the "Bore", "Good" and "Comely" Cluster.

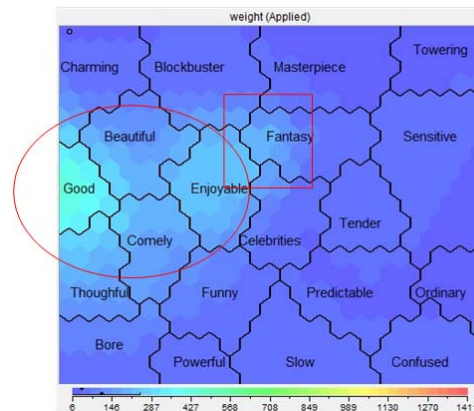


Figure 6. SOM model applied to Crazy Heart.

VIII. CONCLUSION

In this paper we suggest a methodology that proposes a multi-dimensional model on which to measure and identify sentiment. We use as a test case expressions of sentiment present in the microblog domain Twitter. We present an example of a 20 dimensional space on which to locate movies based on the derived SOM map. We propose that this provides a methodology for evaluating sentiment in relation to movies without reading the entire corpus of comments in the twitter database. We test the methodology on 2 movies which are not a part of the training data by visualizing their location on the SOM map. Possible improvements in the methodology

include exploring different nomination procedures other than “most frequent sentiment”, improvements in prior filtering of comments utilizing a subjectivity sensitive filter, extending and refining the aggregated movie sentiment lexicon, and improvement of the stemming algorithm.

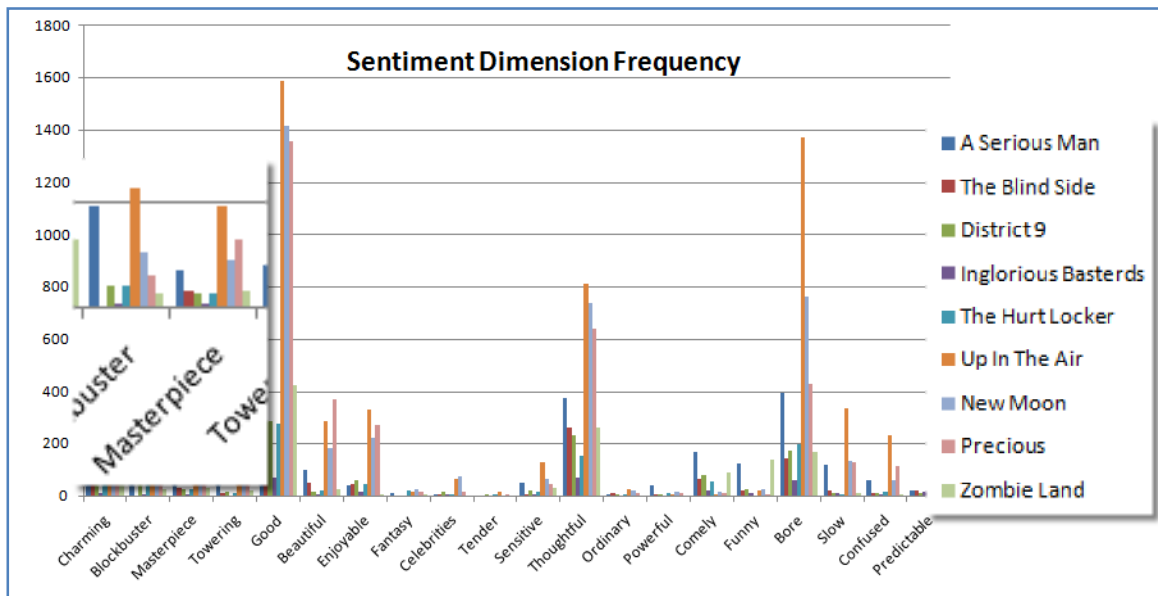


Figure 7. Comparison of 10 movies with respect to each particular sentiment. The sentiment Masterpiece is magnified to view it more easily.

REFERENCES

- [1] 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.
- [2] N. Kwon, S. Shulman, and E. Hovy, “Multidimensional text analysis for eRulemaking, in Proceedings of Digital Government Research (dg.o), 2006.
- [3] 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.
- [4] Timo Honkeda (1997) Using Self-Organizing Maps in Natural Language Processing, Helsinki University of Technology, Neural Network research center, Finland.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] N.Jakob, S.H. Weber, M.C. Muller, I. Gurevych, "Beyond the Stars: Exploiting Free-Text User Reviews to Improve the Accuracy of Movie Recommendations," Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, 2009.

- [10] B Schuller, J Schenk, G Rigoll, T Knaup, I Pingsta, "“The Godfather” vs. “Chaos”: 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 Recognition, 2009.
- [11] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. JASIST, 2009.
- [12] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, vol. 22, pp. 110–125, 2006.
- [13] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of ACM-KDD 2004, pp.168-177.
- [14] Salton Gerard (1983) in Introduction to Modern Information Retrieval. McGraw-Hill. 448 pages.
- [15] Hearst, M. A., Hurst, M., and Dumais, S. T. 2008. What should blog search look like?. In Proceeding of the 2008 ACM Workshop on Search in Social Media (Napa Valley, California, USA, October 30 - 30, 2008). SSM '08. ACM, New York, NY, 95-98. DOI= <http://doi.acm.org/10.1145/1458583.1458599>.