

Harvesting Consumer Opinion and Wine Knowledge Off the Social Media Grape Vine Utilizing Artificial Neural Networks

William B. Claster, Maxwell Caughron
School of Asia Pacific Management
Ritsumeikan Asia Pacific University
Beppu, Japan
wclaster@apu.ac.jp, maxwca10@apu.ac.jp

Philip J. Sallis
Geoinformatics Research Centre,
Auckland University of Technology,
Auckland, New Zealand
philip.sallis@aut.ac.nz

Abstract-- In this paper we mine over 80 million twitter microblogs in order to explore whether data from the social media initiative known as Twitter can be used to identify sentiment about red wines. We test to see whether models derived from Twitter data can corroborate industry sales figures and we employ text analysis software developed to assess emotional, cognitive, and structural components of text to analyze the twitter dataset to harvest knowledge about consumer sentiment on different wine varietals. A multi-knowledge based approach is proposed using, Self-Organizing Maps and domain expertise in order to establish view the social network conversation. We show that it is possible to both confirm previously known knowledge and find novel information through the proposed methodology.

Keywords; Wine, Viticulture, 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 towards 8 varietals of red wine: Syrah/Shiraz, Merlot, Cabernet Sauvignon, Malbec, Pinot, Zinfandel, Sangiovese, and Barbera.

II. BACKGROUND

Much work has recently been undertaken in sentiment mining over the last few years. Pang and Lee give an 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 these shorter expressions, although conclusive research has

not been conducted¹. 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 classify sentiment within Tweets as positive or negative². They achieve accuracy as high as 81%. 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 [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].

In our previous work, a binary choice algorithm was employed to produce a one dimensional measure of sentiment (negative to positive) and was plotted over time to understand movie sentiment³.

III. DATA DESCRIPTION AND MODELING

Using twitter's API, we collected and stored 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 only the textual comment in this study. The tweet comments were then filtered such that only those containing references within the textual comment to any of the following keywords; Syrah/Shiraz, Merlot, Cabernet Sauvignon, Malbec, Pinot, Zinfandel, Sangiovese, and Barbera were included.

To extract the sentiment of these tweets automatically, we used LIWC2007 (Linguistic Inquiry and Word Count; Pennebaker, Chung, and Ireland 2007 [13]), a text analysis software developed to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. This software calculates the degree to which a text sample contains words belonging to empirically defined psychological and structural categories. Specifically, it determines the rate at which certain cognitions and emotions (e.g., future orientation, positive or negative emotions) are present in the text. For each psychological dimension the software calculates the relative frequency with which words related to that dimension occur in a given text sample (e.g., the words "maybe", "perhaps", or "guess" are counted as representatives of the construct "tentativeness"). LIWC has been used

widely in psychology and linguistics [14]. For example, Yu, Kaufmann, and Diermeier [15] have used LIWC to measure the sentiment levels in US Senatorial speeches. We focus on 19 dimensions in order to profile wine sentiment: swearing, social, family, friend, affection, posemotion, negemotion, anxiety, anger, sadness, see, hear, feel, health, sexual, work, leisure, home, and money. Following the methodology used by Yu, Kaufmann, and Diermeier [15] we concatenated all tweets published over the relevant timeframe into one text sample for each varietal separately and then also a combination of all the varietals together and used these as inputs into LIWC.

IV. TWITTER AS A REFLECTION OF CONSUMER SENTIMENT ON WINE

The fact that users are discussing wines online does not mean necessarily that we can extract meaningful information from their comments. To explore the question we aggregated the information stream about the 8 varietals and compared the resulting profiles with anecdotal evidence and industry publications. In order to analyze the sentiment of the tweets, we generated multi-dimensional profiles of the wines in our sample using the relative frequencies of LIWC category word counts.

Figure 1 shows these profiles for the 8 varietals taken as a group (AllWines). Overall, positive emotion outweighs negative emotion by nearly a factor of 4 in an LIWC-based analysis. The most prominent dimensions are positive emotion, affective, social, leisure, see, and money. As may be expected, negative emotions such as anger, sadness, and anxiety are substantially missing from tweets relating to wines. It is also interesting to note that money has a significantly smaller presence in the tweet conversation as compared with social and leisure dimensions and this may assist marketers in tuning their message to consumers.

Figure 2 shows some variation among varietals in the dimensions of affective, positive emotion, see, and leisure. Affective is a dimension that Relate to moods, feelings, and attitudes. 'See' has to do with visual perception including color and brightness. In terms of positive emotion Merlot, the most popular wine in the US market is the highest, with Pinot and Malbec tied for second place. Merlot is also highest with regard to the affective dimension. Cabernet Sauvignon is lowest on both of these dimensions. On the social dimension Merlot again stands out. On the visual dimension, 'see', Zinfandel scores highest with Sangiovese coming in a close second. The other

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

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

³ <http://cicsyn2010.info/>

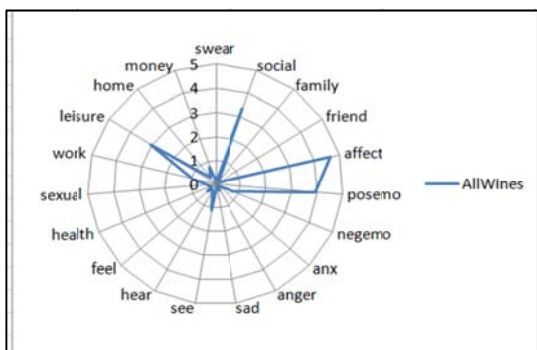


Figure 1. All 8 wines together evaluated on 19 dimensions using LIWC.

varietals are significantly lower on this dimension. In fact this dimension exhibits the greatest variation across the varietals. Pinot and Malbec are lowest here. In terms of leisure Zinfandel and Sangiovese are virtually tied for first place with Cabernet Sauvignon coming in second. In terms of the money dimension Cabernet Sauvignon is highest. This may reflect the fact that this varietal is generally the most costly among those considered in the study. In terms of negative emotion, Malbec seems to have the minimum and Merlot the maximum.

Figure 3 shows sales figures for 2008 in the US. It is interesting to note that none of the dimensions we considered can accurately predict sales rank. However, as table 1 shows, ranking according to positive emotion is a good predictor of sales rank with the exception of the rank of Cabernet Sauvignon.

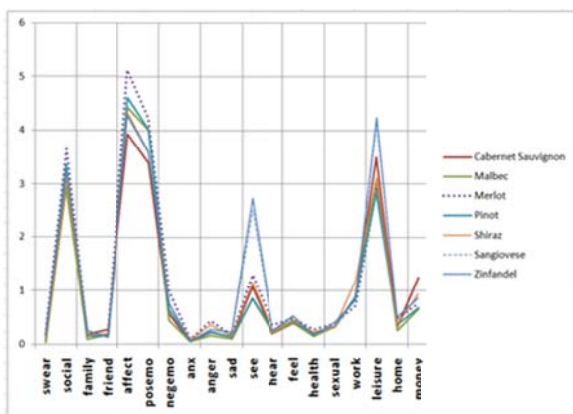


Figure 2. Individual varietals measured across 19 dimensions

Table 1. Varietals ranked by positive emotion

Dimensions	posemo
Merlot	4.23
Pinot	4.02
Malbec	4.01
Sangiovese	3.63
Shiraz	3.62
Zinfandel	3.59
Cabernet Sauvignon	3.39

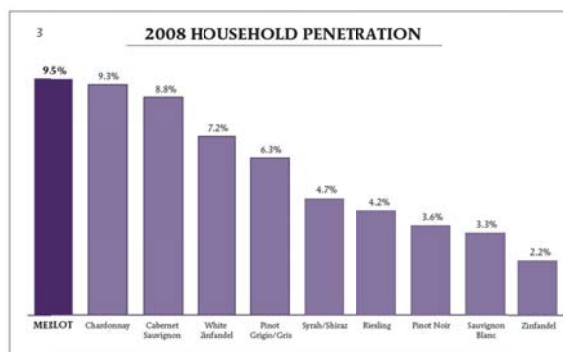


Figure 3. Industry report ranking varietals by sales.

V. PREPROCESSING TWITTER DATA FOR ANN SOM.

In order to visualize the data beyond the LIWC scale a series of self-organizing maps were 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. The document-term matrix was generated according to the standard methodology developed by Salton [16]

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.

The document-term matrix was analyzed with a self-organizing map algorithm. 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 from the entire dataset and then filter the dataset by varietal. Figure 4 shows the SOM for the entire set. We also present SOMs for Pinot (figure 5), Merlot (figure 6), Cabernet Sauvignon (figure 7), and

Shiraz (figure 8). These SOMs can be viewed as an organic form to be used in its place of the dimensions enumerated in LIWC scale or as tags that indicate key components in the conversation taking place in Twitter for the particular varietal(s) in the dataset.

Figure 5 shows a SOM for Pinot and the following observations can be noted. Pinot is a very picky grape which requires just the right conditions and thus the inclusion of the word “young” – the only derogatory comment in any of the SOMs may reflect this. This figure also contains the tag “wine spectator”. Both of these tags and the relative absence of non-taste related tags may confirm the stereotype that Pinot is a connoisseur’s variety. The tag for ‘fake’ seems puzzling here but a search into the dataset shows numerous tweets referring to a scandal involving the Pinot Noir varietal and reported in the news from Dec. ’09 to Feb. ’10. For example one tweet in February was “Fake Pinot Noir Scandal Ends In Convictions In French Court: Twelve French wine industry figures were convicted We... <http://bit.ly/cigxCp>”. We also see that a conversation about Manohara Pinot an Indonesia-American socialite, is dominating the map. There were a large number of conversations in Twitter relating to a kidnapping that was reported on this subject. This illustrates a fundamental property of the Self-Organizing Map algorithm, - it is a topologically preserving mapping from a high-dimensional space to a lower-dimensional viewable space [17].

The types of food mentioned in each of the varietal SOMs matched the wine-food pairings that are common, especially with Merlot to steak, shown in Figure 6, and Malbec to lamb, salmon, and ribs in figure 9. In fact it is noteworthy that there is very little overlap in food pairings among the different varietals with the exception of pasta which shows up for both Cabernet and for Shiraz. However due to the variety of flavor possibilities of ‘pasta’, this exception here seems to prove the rule. The pairings of certain varietals with certain foods shows that pairings are solidifying. Finally despite the mention of ‘African’ which may refer to the cuisine, mentions of food pairings were entirely with traditional western fare showing that wine pairings with non-western foods are still undeveloped. This reflects the growing interest with pairings of wine and food.

The variety of flavor related terms “brie, honey, flavors etc.” mentioned in the Cabernet SOM (figure 7) compared to Merlot or the Shiraz (figures 6 and 8) shows the greater range of the flavors and combinations possible for the winemaker with Cabernet, whereas the Merlot or the Shiraz is a more standardized flavor. Also Cabernet is the only one with a price related tag “deal” reflecting the fact that it’s usually the most expensive wine on the list. The inclusion of the term “finish” which refers to the final taste of the wine is thought to be what separates a good cabernet from inferior cabernet. Conversely, the

Shiraz has the tag “cheap” Also, the dominance of Cabernet Sauvignon in the market is reflected by the near absence of other types of wine tags whereas the Shiraz SOM is dominated by references to Cabernet Sauvignon. Cabernet’s strong association with a particular wine region Napa is initially very surprising given that Cabernet Sauvignon is grown in most wine growing regions. However an inspection of the tweet geo-references (figure 10) shows that this may reflect a much higher number of American tweeters than wine drinkers from other English speaking countries.

VII. CONCLUSIONS AND FURTHER WORK.

Data mining applied to social media can explain and reflect opinions and sentiment about consumer goods. In this study we employed LIWC, artificial neural networks, and domain expertise to summarize and investigate consumer sentiment towards 8 red wine varietals. The Linguistic Inquiry and Word Count shows that the social network conversation can be understood through algorithmic analysis. Specifically we found that leisure and social dimensions predominate in the discussion in this group of wines and that money seemed to be “tweeted about” less and this may indicate that the former dimensions are of more important to this audience (tweeters). Using artificial neural networks provides an organic multidimensional view of the tweet conversation. Although much of the analysis confirms domain expertise, this is significant as it supports the methodological approach suggested here. Novel information can also be reaped through this process as well as evidenced by the mined result that there was a significant media scandal involving Pinot wines this time period. This analysis could be done in real time to provide vendors and customers with up to the minute analysis. 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 products and topics.



Figure 4. SOM for all wines together.



Figure 5. SOM for the Pinot varietal



Figure 8. SOM for the Shiraz varietal



Figure 6. SOM for the Merlot varietal

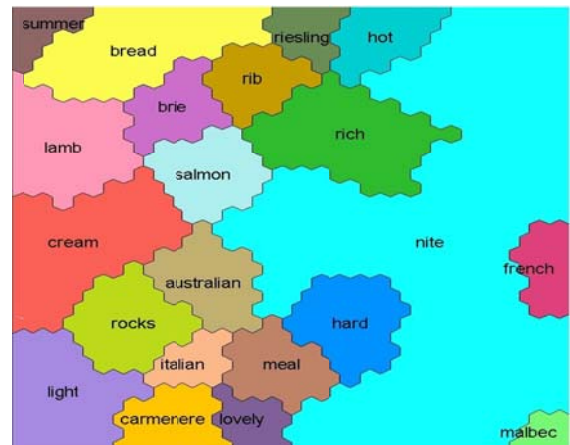


Figure 9. SOM for the Malbec varietal



Figure 7. SOM for the Cabernet Sauvignon varietal

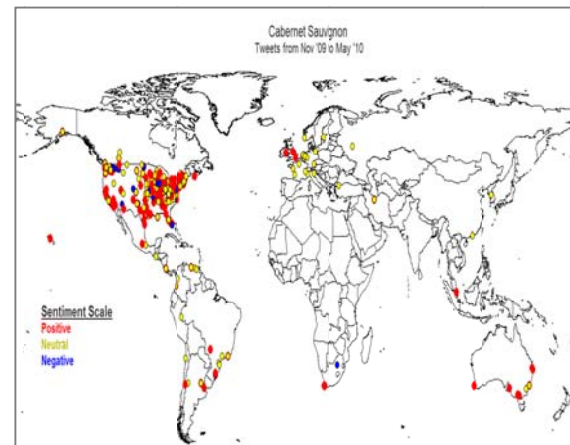


Figure 10. Geographical distribution of tweets on Cabernet Sauvignon. Also showing an approximate measure of positive/neutral/negative sentiment (red, yellow, blue respectively).

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 Recognition, 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] Pennebaker, J.; Chung, C.; and Ireland, M. 2007. The development and psychometric properties of LIWC2007. Austin, TX.
- [14] Tausczik, Y. R., and Pennebaker, J. W. 2009. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology.
- [15] Yu, B.; Kaufmann, S.; and Diermeier, D. 2008. Exploring the characteristics of opinion expressions for political opinion classification. In Proceedings of the 2008 international conference on Digital government research, 82-91. Montreal.
- [16] Salton &&&, G. and M. J. McGill (1983). Introduction to modern information retrieval. McGraw-Hill. ISBN 0070544840.
- [17] Timo Honkeda (1997) Using Self-Organizing Maps in Natural Language Processing, Helsinki University of Technology, Neural Network research Center, Finland.