A Case Study on Mining Social Media Data

H. K. Chan[1], E. Lacka[2], R. W. Y. Yee[3], M. K. Lim[4]
[1]Nottingham University Business School China, University of Nottingham, Ningbo, China
[2] Department of Marketing, Strathclyde Business School, University of Strathclyde, Glasgow, UK
[3] Institute of Textiles and Clothing, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[4]Derby Business School, University of Derby, Derby, UK
(hingkai.chan@nottingham.edu.cn)

Abstract – **In recent years, usage of social media web-sites have been soaring. This trend not only limits to personal but corporate web-sites. The latter platforms contain an enormous amount of data posted by customers or users. Without a surprise, the data in corporate social media web-sites are normally link to the products or services provided by the companies. Therefore, the data can be utilized for the sake of companies' benefits. For example, operations management research and practice with the objective to make decisions on product and process design. Nevertheless, little has been done in this area. In this connection, this paper presents a case study to showcase how social media data can be exploited. A structured approach is proposed which involves the analysis of social media comments and a statistical cluster analysis to identify the inter-relationships among important factors.**

Keywords - **Social Media, text mining, content analysis, cluster analysis**

I. INTRODUCTION

Web technology and applications have been penetrating to end users in the last two decades. Social media is one of such examples. Usage of such applications has been soaring at an unprecedented rate. Despite the origin of non-web-based platforms, the term "social media" is mainly referred to online applications that allow users to exchange their comments electronically, which form the development of Web 2.0 [1]. By definition, these platforms have the unique characteristic: "content and applications are no longer created and published by individuals, but instead are continuously modified by all users in a participatory and collaborative fashion" [2]. In this research, we also adopt this definition of "social media".

Social media data are essentially secondary data, which normally "have not been collected with a specific research purpose" [3]. In most cases they are not well-structured because of the aforementioned reason. Therefore, extracting value from social media data requires the transformation of unstructured data to structured data. Additionally, users' comments are very biased in many cases, especially there is no structured way for them to post their comments. Subjectivity of these data adds additional uncertainty regarding the reliability of the information being used [4]. In spite of the abovementioned issues, social media data can still be a good source of information. This is magnified by the fact that social media data are free of charge generally. Companies can download the data from their web-sites freely.

Typical Operations Management (OM) research involves making decision. With such unstructured and imprecise social media dataset, it is difficult, if not impossible, for OM researchers to fully utilize the value of social media data. It is therefore not surprising that this area is under-researched. This research aims to overcome the challenges by combining qualitative content and statistical cluster analysis in order to reveal the factors and their inter-relationship from the social media data. This approach can convert the unstructured dataset into a structured hierarchy based on the statistical approach, which help to reduce the negative impacts associated with the subjectivity of the data.

The contribution of this paper is twofold. First, to help identify the factors/themes/issues from the social media data through content and cluster analysis. The main concern is from the OM perspective so applications could be linked to product development, process design, and also supply chain management. Second, to explore a structured approach to analyze social media data associate to and facilitate decision-making research. The focus of this paper is put on product development with respect to different OM performance indicators. Data collection and analysis are facilitated by the latest version of NVivo, a content analysis tool, which incorporates a new web browser plug-in called NCapture capable of capturing social media data (in raw format). This plug-in provides a channel to download social media data for further analysis by the software NVivo.

The next section reveals the research method of this paper, including how to access social media data, and the procedures to analyze the data. Section III then summarizes the results from the content analysis and cluster analysis. It also presents the findings. Section IV concludes this paper.

II. METHODOLOGY

Facebook is selected as the source of social media data in this research. This is partly because Facebook is the most popular social media platform, and the case company has its own Facebook page for data collection. More specifically, data were accessed and downloaded for analysis from the SAMSUNG Mobile Facebook page [5] due to the launch of the Samsung smartphone. "Data"

978-1-4799-6410-9/14/$31.00 ©2014 IEEE

refer to the comments posted by Facebook users on the captioned Facebook page. They are captured using NCapture for NVivo 10. Overall 128371 comments from 10 June to 10 September 2013 were downloaded. To retain focus on product development as the subject of this research, posts in relation to Samsung Galaxy S4 that was launched in late April 2013 was selected for analysis. S5 is not used as when the research was conducted, the model was not available yet. Only comments posted in English language were considered for analysis. Then, content analysis was carried out using conceptual analysis and then relational analysis with the help of statistical cluster analysis, as visualized in the flow diagram below (Figure 1).

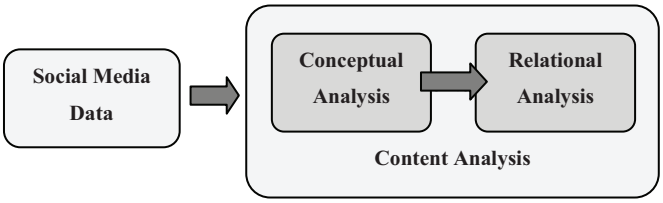

Fig. 1. Research model.

The conceptual analysis is to convert the qualitative comments into manageable quantitative parameters. They are mainly the frequency of occurrences of the codes and concepts of concern, and how they link to each comment (each comment can be mapped to more than one concepts or codes). Therefore, definition of the coding strategy is imperative in this stage. Below outline the procedures to define the coding strategy.

First the concepts/codes were clearly defined based on the objectives of operation performance indices (first column of Table I). Next, each concept/code was allocated an individual item in order to reduce subjectivity while analyzing the data. Finally, sample comments were provided in order to avoid further possibility of confusion. This serves additional definition of the concepts/codes. Table I presents all concepts/codes, allocated items, the label attached to each item, as well as the sample comment.

Consequently, the coding process can be carried out with the help of the clearly defined concepts/codes. Overall 1800 items were selected for final analysis (see Table II).

As mentioned before building on conceptual analysis, the relational analysis was conducted in order to examine the relationships among concepts/codes with respect to the posted comments. The relational analysis aided examination of the relationships among concepts/codes in the data by statistically examine these relationships, cluster analysis was conducted and Pearson correlation coefficient test was run. The results of cluster analysis and Pearson correlation coefficient test are discussed in next section.

After this stage, the relationship of the factors can be extracted based on the comments posted by the Facebook users. Since this step aggregate the users' comments for statistical analysis so individual subjectivity can be reduced. However, this approach cannot remove the subjectivity completely, and cannot detect collective bias, if any.

TABLE I
DEFINITION OF CONCEPTS/CODES

Concept/ Code	Item	Label	Comment
Speed	Delivering the product to the consumer as soon as possible	S1	Questions regarding product introduction date
Dependability	Doing things on time as promised	D1	Questions regarding the delivery update of update
	Developing trustworthiness	D2	Comments regarding consumers' willingness/ unwillingness to purchase the product
	Using effective equipment	D3	-
	Developing effective communication	D4	All kinds of questions asked by consumers and help requests
Flexibility	Being able to change operations to fulfil new requirements	F1	Comments suggesting introduction of the product and its features
	Being able to introduce new products or modify existing products	F2	Comments suggesting improvement of the product and its features
Quality	Meeting expectations	Q1	Consumers' statements outlining their satisfaction/ dissatisfaction with the product related to consumers' expectations
	Fulfilling requirements	Q2	Comments regarding product's features and problems encountered due to faulty features
	Maintaining effective communication	Q3	-
	Doing things right	Q4	Comments concerning satisfaction/ dissatisfaction (e.g. I like S4)
Cost	Doing things economically at low price	C1	Questions regarding product price and cost of product repair
Lead time	Production time	L1	Questions regarding product update

TABLE II
NUMBER OF CLUSTERS

Concept/ Code- Item	Label	No of References
Speed- Delivering the product to the consumer as soon as possible	S1	93
Dependability- Doing things on time as promised	D1	25
Dependability- Developing trustworthiness	D2	148
Dependability- Using effective equipment	D3	0
Dependability- Developing effective communication	D4	456
Flexibility- Being able to change operations to fulfil new requirements	F1	13
Flexibility- Being able to introduce new products or modify existing products	F2	106
Quality- Meeting expectations	Q1	139
Quality- Fulfilling requirements	Q2	366
Quality- Maintaining effective communication	Q3	0
Quality- Doing things right	Q4	287
Cost- Doing things economically at low price	C1	127
Lead time- Production time	L1	40

III. RESULTS

Content analysis is able to extract empirical information from the datasets to formulate a theory, generate propositions and so on. Normally, large scripts with less number of samples are involved. Social media data, however, are short and involve many users (i.e. samples) so the size of which is simply too big to be handled. More importantly, it is not easy to generate empirical findings like interviews, in traditional channel for gathering data for content analysis. Therefore, in this research statistical cluster analysis is employed to help aggregate the social media data based on the output of the conceptual analysis outlined in previous section. This allows researchers to classify a large dataset into a number of subsets, which are sometimes referred to as objects [6]. This has been applied in some disciplines such as marketing [7]. The approach can also reduce the number of factors of our concern [8].

The question really is how the clusters can be formed. Pearson Correlation Coefficient [9] is adopted in this research. The coefficient measures the similarity of a pair of "objects". The closer the coefficient is to 1, the higher is the similarity of the pair. In contrast, negative value refers to dissimilarity so -1 means the pair is not similar at all. Zero value means the pair is not correlated to each other (i.e. they have no linear relationship).

Following this line of thought, clusters of important OM criteria (as shown in Table 2) in relation to the consumers' comments on the chosen product can be calculated. Table 3 lists the coefficients and the corresponding pair of items (i.e. labels in the table). For example, Q1 and Q2 pair has a high value of the coefficient (0.971369), which means they are very similar. This is not surprising as these two factors are all related to quality. On the contrary, Q4 and F1 pair has a value of 0.670879, which means they are not "close". Q1 and F1 pair is even worse and has a value of 0.629375. Highly correlated factors can be grouped together. As a consequence, a hierarchy of clusters can be constructed that can facilitate the later decision-making process.

TABLE III
PEARSON CORRELATION COEFFICIENT OF THE ITEMS

Label	Label	Pearson correlation coefficient
Q2	Q1	0.971369
Q4	Q2	0.970377
Q4	Q1	0.968907
Q2	D4	0.948166
F2	D4	0.939958
Q4	C1	0.937702
D4	D2	0.935226
Q4	D4	0.933664
Q2	C1	0.928996
Q1	C1	0.923178
Q4	F2	0.905635
D4	C1	0.901911
Q4	D2	0.899569
F2	D2	0.896237
Q2	F2	0.895921
S1	L1	0.894889
Q1	D4	0.887127
Q2	D2	0.872725
D2	C1	0.872062
S1	D4	0.857865
F2	C1	0.854281
S1	F2	0.853988
S1	D2	0.853459
Q1	F2	0.840803
Q1	D2	0.834816
L1	F2	0.829576
L1	D4	0.803419
L1	D2	0.800131
S1	Q4	0.767503
D4	D1	0.755551
S1	C1	0.751035
S1	Q2	0.733648
Q4	D1	0.732828
F2	F1	0.722472
Q2	D1	0.722408
D1	C1	0.721811
F2	D1	0.719562
D2	D1	0.718563
F1	D4	0.712495
Q1	D1	0.709018
Q4	L1	0.707582
S1	D1	0.698561
F1	D2	0.682489
Q2	L1	0.682228
S1	Q1	0.670933
Q4	F1	0.670879
L1	C1	0.66919
Q2	F1	0.665662
F1	C1	0.662722
S1	F1	0.644967
Q1	F1	0.629375
L1	D1	0.621831
Q1	L1	0.614663
F1	D1	0.577654
L1	F1	0.570809

Using the coefficients (i.e. similarity), a straightforward benefit of the analysis is that researchers can formulate or verify (potential) hypotheses among the criteria. For example, Q1 and Q2 are highly related as mentioned before but this is not surprising. Another pair is that production lead time (L1) is highly related to another measure, namely, speed to deliver the product to the consumer as soon as possible (S1), despite the fact that they are under different dimensions (Lead time and Speed respectively). Their Pearson Correlation Coefficient is 0.894889, which is very high. The implication is that L1 and S1 can be grouped as one cluster and hence a positive relationship between these two factors can be hypothesized for further analysis such as in a quantitative questionnaire survey. To generalize this at a higher level, one may wonder if the six dimensions are homogenous measures. From Table 3, it is safe to conclude that the items under the dimensions can correlate to the items under other dimensions. Quality is one of such dimensions.

In contrast, a potentially new relationship between an item under dependability (developing effective communication, D4) and flexibility (being able to introduce new products or modify existing products, F2). The corresponding Pearson Correlation Coefficient is 0.939958, which is even higher than the L1-S1 pair. The relationship is not obvious, but this finding is worth investigating and may lead to a new theoretical development and contributions.

Above examples demonstrate how the cluster analysis provides preliminary evidence to help researchers identify potential hypotheses (i.e. relationships among the factors). Although this paper makes use of Facebook comments in the research, the method itself is not restricted to this type of social media data only. Data from different social media platforms can be analyzed using a similar approach and there is no specific limitation on the input data. Nevertheless, pre-processing of the data may be required to facilitate content analysis due to the diversified nature of social media data. The concern is mainly the format of the data, rather than the content.

V. CONCLUSION

This research demonstrates a structured and practical approach for mining social media data. The focus is put on the OM perspective. The proposed procedures help quantify the qualitative social media data and to group them into clusters with similar characteristics for later applications, such as decision-making. In the future, OM researchers can collect data from this new channel together with the traditional channels (for example, expert judgment, interviews with production people, and so on).

The main contribution of this paper is to outline the approach to extract social media for later analysis. As mentioned above, this involves the quantification of social media data. This outcome can then be utilized in many applications, to name a few, empirical questionnaire survey, design of decision-making systems, and so on. Social media data can be available readily from the Internet. This convenience could be an additional advantage to OM research.

Nevertheless, the authors take a snapshot view on the data (to be precise, four months of data) in this research project. This is a limitation since the social media web-sites are kept updating and the corresponding dataset keeps growing. To address this, a real time data crawling decision-support systems coupled with the corresponding decision-making tools is required in order to monitor the dynamic comments on a real-time basis.

ACKNOWLEDGMENT

A pilot study of this paper was presented at the 2013 International Conference on Social Science and Education [10]. The authors thank the audiences who provided comments, which are very useful in this full study.

REFERENCES

[1] B. W. Wirtz, R. Piehler, and S. Ullrich, “Determinants of social media website attractiveness.” *Journal of Electronic Commerce Research*, vol. 14, no. 1, pp. 11–33, 2013.

[2] A. M. Kaplan, and M. Haenlein, “Users of the world, unite! The challenges and opportunities of social media,” *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[3] H. T. Sørensen, S. Sabroe, and J. Olsen, “A framework for evaluation of secondary data sources for epidemiological research,” *International Journal of Epidemiology*, vol. 25, no. 2, pp. 435–442, 1996.

[4] J. Q. Zhang, G. Craciun, and D. Shin, “When does electronic word-of-mouth matter? A study of consumer product reviews,” *Journal of Business Research*, vol. 63, no. 12, pp. 1336–1341, 2010.

[5] Samsung Mobile, available at: https://www.facebook.com/SamsungMobile

[6] D. J. Ketchen, and C. L. Shook, “The application of cluster analysis in strategic management research: an analysis and critique,” *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, 1996.

[7] G. Punj, and D. W. Stewart, “Cluster analysis in marketing research: review and suggestions for application,” *Journal of Marketing Research*, vol. 20, no. 2, pp. 134–148, 1983.

[8] J. F. Hair, W. C. Black, B. J. Bahin, and R. E. Anderson, *Multivariate data analysis: A global perspective*. London: Pearson Education, 2010.

[9] P. Ahlgren, B. Jarneving, and R. Rousseau, “Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient,” *Journal of the American Society for Information Science and Technology*, vol. 54, no. 6, pp. 550–560, 2003.

[10] H. K. Chan, and E. Lacka, “Using social media data for operations management research: an example of product development,” in *Proc. 2013 International Conference on Social Science and Education* (published in *Advances in Education Research*), Hong Kong, 2013, vol. 26, pp. 522–525.