A Study of Extracting Demands of Social Media Fans

Chih-Chuan Chen
Interdisciplinary Program of Green and Information Technology
National Taitung University
Taitung, Taiwan, R.O.C.
e-mail: ccchen@nttu.edu.tw

Chien-Wei He
Institute of Information Management
National Cheng Kung University
Tainan City, Taiwan, R.O.C.
e-mail: simulatedmsn@gmail.com

Hui-Chi Chuang
Institute of Information Management
National Cheng Kung University
Tainan City, Taiwan, R.O.C.
e-mail: huichi613@gmail.com

Sheng-Tun Li
Institute of Information Management, Department of Industrial and Information Management
National Cheng Kung University
Tainan City, Taiwan, R.O.C.
e-mail: stli@mail.ncku.edu.tw

***Abstract*—With the boom of the Internet era, people spend more and more time on social media, such as Facebook, Twitter, and Tumblr. How to get people's attention is becoming a critical issue for companies and celebrities, since it is an era of distractions. In the past, if a company wanted to become popular, it simply spent money on traditional media, like newspapers or TV commercials. Now, one has to know audiences' needs and then utilize the new social media platforms to reach those specific audiences. How to know the demand of customers (audiences) is an unavoidable challenge? To answer that question, most commonly used methods are conducting market surveys, including questionnaires and focus groups. However, it is not only time wasting but also effort consuming. In this paper, we combine text mining techniques and Kansei engineering to analyze audiences' demand. Firstly, we collect data from Facebook Fan Pages, including numerical data (number of likes, shares, comments) and text data (postcontent). Secondly, we extract the topics by using Latent Dirichlet Allocation (LDA). Thirdly, experts will give eight pairs of Kansei words that are most relevant to the articles. Finally, we produce a semantic differential questionnaire to find the relationship between topics and Kansei words. The relationship can give helpful insights into the demands of the audience. Moreover, a supervised LDA is incorporated in this approach to predict the popularity of posts.**

Keywords—Kansei engineering; topic model; text mining; demand analysis

I. INTRODUCTION

How can I become famous? This is a question that every author asks himself or herself. Traditionally, a writer signs a contract with a publishing house, and then the publishing house harnesses their own channels to promote the writer's books. Nowadays, with the boom of the Web and social media, writers must cultivate online channels, as well. According to a recent survey by the Market Intelligence and Consulting Institute (MIT) in Taiwan [1], the top five most used social media platforms, in order, are Facebook (95.8%), Google+ (24.7%), Pixnet (20.7%), Xuite (12.7%), Plurk (8%). The top three discussion groups are Mobile01 (51.4%), NTU PTT (51.2%), Yahoo Knowledge+ (46.2%). Accordingly, there are both opportunities and challenges for writers. If writers can devote themselves to managing a social-media platform, they will gain popularity, which may boost their next publication sales. Conversely, the next book sales may not improve if the writers do not pay attention to their online platform.

In this respect, some proactive writers take advantage of social media, and post some parts of their publications to the Web, then track the audiences' reactions. If the audience enjoys the post's content, they might write more about it, and vice versa. Some famous examples of such an approach include Giddens Ko, Hiyawu and Riff Raff Tsai. Once the online popularity has accumulated, they will publish a paper book via the offline channel.

Among online publications, one branch has been rising, which is called "healing essay". In the contemporary era, people feel more pressure than before; thus, "healing" became a popular topic. On Google Trends, a search for the keyword 'healing' shows that the search volume has increased year after year since 2011. This trend has resulted in many products, such as the "healing picture book", "healing music" and "healing app". The healing essay also fits within this healing category. Moreover, the healing essay can make people feel better or inspired when they are struggling with depression and pressure.

Although we are well aware of the importance of social media, its operation is still difficult for writers and publishing houses. The main challenge is that people's demands are not particularly clear. Data collected by social media are unstructured and text-based, which creates difficulty when analyzing people's demands. In order to better analyze the demand, some people will use Kansei engineering, but doing so is time-consuming. As such, in this study, we use Latent Dirichlet Allocation (LDA), a probability language model, to reduce the time needed to obtain factors for why some essays are so attractive [2].

In this paper, we endeavor to build a framework to characterize the demand of the audience and the popularity of articles. We introduce Kansei Engineering and LDA to extract the Kansei words and topics. After experts turn the identified Kansei words into topic names, we conduct semantic differential questionnaires to highlight the relationships between topics and Kansei words.

II. Literature Review

A. Text Mining

In the previous data mining study, researchers focused on numerical data that came from the relational database. Nowadays, we find that around 85~90% of data are stored in unstructured format [3], such as emails, customer's comments, and PDF files. In addition, this data usually contains a large amount of important information; yet understanding it by computer is challenging. As a result, text mining has become increasingly common in recent years. Text mining is also known as knowledge discovery in textual databases, and can be applied to many areas.

For instance, Jin, Ji, Liu, and Johnson Lim [3] translated online customer opinions into engineering characteristics, which is an important step in Quality Function Deployment (QFD). In the traditional QFD method, to digest customer reviews into useful information is a time-consuming and labor-intensive process. In Jin's study, 770 printer reviews were collected from Amazon and Epson's website, and it took more than two weeks to translate the reviews.

The second example is Popescu's research [4], in which epochs of human history were automatically defined. Conventionally, the definition of an epoch relies on historians' knowledge and observations of extended time periods. However, it is difficult to define such epochs objectively and no standard measurement to support their opinions exists. Therefore, the research team decided to use Google n-gram to find evidence of epochs changing. Google n-gram is a part of the Google Books project, and counts any word or short sentence yearly in sources printed from the 6th century to the 21st century, where n represents the length of segmentation. For example, 1-gram means to separate each word in the sentence; 2-grams mean two words will become a group (e.g., “what do”, “do you”, “you need”). The researchers employed many statistic measurements to identify significant changes in word frequencies.

B. Kansei Engineering

Kansei Engineering (KE) is a method to convert consumers' feelings and images for products into design elements [5]. Kansei is an ergonomics and consumer-oriented approach for producing new products [6] and is defined as “translating the customers' Kansei into the product design domain” [5]. In other words, Kansei is applied to translate the feelings and images of customers regarding what they want, need, and demand into the product design field, including product mechanical function [6]. KE can be applied to many areas, such as door design [7], kitchen design [8], housing design [9], [10]

However, to the best of our knowledge, there are no studies that have investigated the transiting audiences' Kansei to article design. In recent years, social media growth has been phenomenal, with more and more people sharing their status and feelings on the Web. When they share their bad situations, they are actually searching for someone to give a positive reply. Sometimes, this reply will have healing effects to those people in need. Such responses are called ‘healing essays’ in this paper.

C. Topic Model

Topic model, which is widely used in machine learning and in Natural Language Processing (NLP) areas, is a generative model in statistical theory. The purposes of the topic model are to discover the latent semantics (latent topics) within a corpus, and to build a generative model through those latent topics. Topic model applies various probability distributions to constructing the generative model, which could deal with the semantic problem better, like synonym and polysemy. The two most common topic model are Probabilistic latent semantic analysis (PLSA) and LDA [2].

PLSA extends the concept of latent semantic analysis (LSA) using statistical view [11]. Instead of SVD, PLSA employs aspect model as its main structure. Aspect model is a latent variable model which represents the latent semantic relation within observed data by probability function. Then maximum likelihood estimation is used for inferring the parameters of PLSA model.

Latent Dirichlet allocation is a probabilistic generative model of a corpus, which can solve the problems that PLSA suffers from. In addition, PLSA is a special case of LDA [2]. In LDA, each document is considered as a mixture model that is constructed by random latent topic, and each latent topic is characterized by a distribution over words. LDA applies a three layers' representation to a corpus, and employs different probabilistic distributions between layers. Recently, some research has shown that LDA performs well in natural language model [12], [13] and machine learning areas [14].

Most topic model are unsupervised, just like the LDA. However, in a realistic world, some corpus is labeled, which means that each document belongs to one category. Mcauliffe and Blei proposed an improved version of LDA in 2008 [15]. This method can let us deal with labeled data. We use response to denote the labeled value, and the value can be categories, ordered class label, and real values.

The generative process of each document is similar to LDA, but it adds a response variable. It also uses E-M algorithm to maximize the likelihood function. The Supervised Latent Dirichlet Allocation (SLDA) method has better performance than traditional LDA method, since the response will help the process of LDA.

III. Research Method

A. Data Collection

Facebook was launched in 2004 by Mark Zuckerberg and was opened in Taiwan in 2006. After one decade, Facebook has become the most popular social network in Taiwan. People use Facebook to share opinions, check in, update their recent status, and mo*st importantly, read news feed.*

Facebook also has Pages for companies and celebrities. Businesses have found it to be a good channel to broadcast their ideas and promote their products. In recent years, some bloggers and popular writers have used Pages as a platform to communicate with their fans.

B. Data Preprocessing

In this paper, the collected corpus was Mandarin Chinese (hereafter referred to as simply Chinese). How to correctly segment sentences into words is an important task. In English, white space can be used as an indicator to cut sentences into words; however, white space do not exist in Chinese sentences. Thanks to efforts of previous researchers, there are many tools that can deal with this problem, such as CKIP (provided by Academia Sinica), Stanford Parser (provided by Stanford NLP group), and Jieba (provided by Sun Junyi).

The preprocessing process can be divided into four steps. The first and second steps involve building the Netizen-words list and Stop-words list, respectively. Then, the third step is segmentation, which is followed by the final step of P-O-S (part of speech) tagging. A more detailed explanation is given below.

Netizen-words list: Many Chinese segmentation tools are based on their training corpus (term dictionary) to obtain better results. As time goes by, more and more special terms are used by netizens. Such terms usually have a particular meaning, such as “BJ4” , which means ‘no need to explain’. If those terms can be collected as a list and provided to a segmentation tool, the accuracy of the results can be greatly enhanced. In other words, we can obtain the right terms from sentences.

Stop-words list Commonly-used words, punctuation and meaningless words (such as “of”) should be removed during the segmentation step. If those words do not get filtered, the segmentation results will contain many useless terms, which will unnecessarily increase the computation loading. Thus, a stop-words list must be built beforehand.

Segmentation: Among the many segmentation tools, we chose Jieba in this paper, the main reason for which is because it is an open source tool that allows code modification if needed. This tool can load users own dictionaries, stop-words list, etc. Although Jieba can interpret unknown words by using Hidden Markov Model (HMM), offer dictionary (netizen-words list and stop-words list) can enhance the performance.

P-O-S tagging: P-O-S tagging is an abbreviation for part-of-speech tagging. In the following Kansei engineering step, we need to distinguish adjectives from sentences.

C. Topic Extraction

In order to extract the topics of every article, we use the LDA algorithm. The LDA input can be divided into four levels, as illustrated in Figure 1. The Documents level is the set of every separate document. We segmented the documents into words to create the Words level. The Dictionary level is the set of all words in each document, for which each unique word was given a unique ID. The final level is creating the corpus, for which the word frequency of each document was counted.

After processing the documents, we needed to set the initial parameter of the LDA algorithm. Following Steyvers and Griffiths, we set α=50/k & β=0.01 [16].

The next step of the LDA involved using the expectation-maximization (EM) algorithm to optimize the parameters. The LDA outputs are two matrices, where one is a document-topic matrix and the other is a topic-word matrix. The values of the document-topic matrix denote the probability distribution between documents and latent topics. The values of the topic-word matrix denote the probability distribution between latent topics and words.

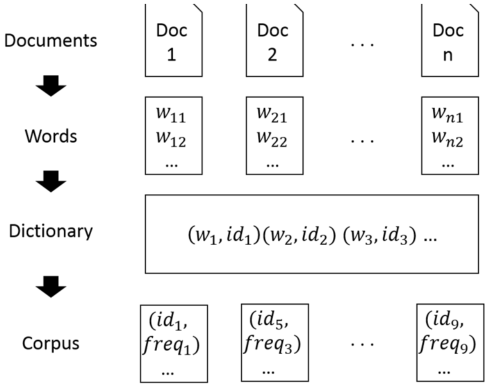

Figure 1. Input of LDA

Experts still need to define the topics’ names. After reading a couple words from each topic, experts give a name that can represent the general concept of each topic.

D. Semantic Differential

Semantic differential is a general method for measuring people's attitudes or feelings. This method was proposed by C. E Osgoods and G. J. Suci in 1957 [17], and often uses a questionnaire to characterize respondents’ emotions through a series of polarization scales. Semantic differential comprises three parts, namely concept, scale and subject. Concepts are for evaluating targets. In this paper, the concepts are topics extracted in the previous step. The scale is composed of two different adjectives, such as important and unimportant. The final part is subjects, which refers to sample size.

We use semantic differential to analyze whether the topics extracted by LDA are important to the audience or not.

IV. Experiment and Analysis

In this section, we record the results of each step during the experiment and conduct an analysis of the results

A. Experiment

Among the many popular writers, we chose Little Lifer as our study target. Little Lifer is a writer in Taiwan, and his posts are written in Chinese. Therefore, we implemented all the experiments by using Chinese articles. There are three reasons we chose Little Lifer. First, he rose to fame from the Web, therefore, the vast majority of his fans comes from the Web. Second, Little Lifer’s Facebook Page has been growing since January 2015, and as of April, 2016, has been liked by 68,000 people. And third, a netizen writer can provide more measurable data from the Web, such as the number of likes and the number of fans.

Among the data collected from Little Lifer, each post can be categorized into three types: Status, Photo and Link. Since this paper focuses on the articles part, which often comprises long articles to help with Netizen’s problems, we filtered all other posts. We deleted each data row except the post message containing “Reply”.

Textual data were collected within Little Lifer’s replies on PTT (the biggest forum in Taiwan). There are a total of 88 articles, 67 of which are also posted on Facebook. At first, we saved each article in a separate file. As mentioned above, Jieba was used to segment the words after loading the Netizen-words list and Stop-words list. The Netizen-words list contained 255 terms frequently used by netizens and Little Lifer. On the other hand, the Stop-words list consisted of 22 punctuation marks and meaningless words. During the segmentation process, we simultaneously performed P-O-S tagging.

TABLE I. ARTICLES SUMMARY OF TERMS COUNT

# Articles		All terms	Adj.	Noun
88 (PTT)	total	47214	1940	9560
	Average	537	22	109
67 (Facebook)	total	36061	1479	7249
	Average	538	22	108

After the process, 44,214 and 36,061 terms were obtained from PTT and Facebook, respectively. The term count summary is shown in TABLE In examining Little Lifer’ s articles more closely on Facebook, we found that 27% of his articles are composed by 301~400 terms, and 24% by 401~500 terms. In other words, about half of his articles contained less than 500 terms, which indicates that Little Lifer tends to write short articles to help his readers.

In this section, we use LDA to automatically generate article topics. Initially, we put each term into the corpus to generate topics, for which the number of topics was set at 4. In doing so, a list of topics was obtained. We generated topics by using all terms, adjectives, nouns, and a mix of nouns, verbs, and adjectives.

After examining the topic composition presented above, we decided to use the adjective topics as our corpus input. There are two primary reasons for choosing adjectives as our input are as follows. Firstly, adjectives are more representative to people’s Kansei attributes; if we read an adjective, we can interpret the meaning immediately. Secondly, adjectives are easier for experts to label; knowing the general concept behind those grouped adjectives is more straightforward.

Then, we changed the number of topics from 4 to 8, which means that totally 5 topics were produced. We sent those results to Little Lifer, who played the role of expert in our study. He suggested that we should try to set 4 as our number of topics, since he thought it would be easier to label each composition. After labeling, we obtained our topic model for Little Lifer’s articles (Figure 2). In the next section, we utilize the Kansei engineering process to gain more insight into our topic model.

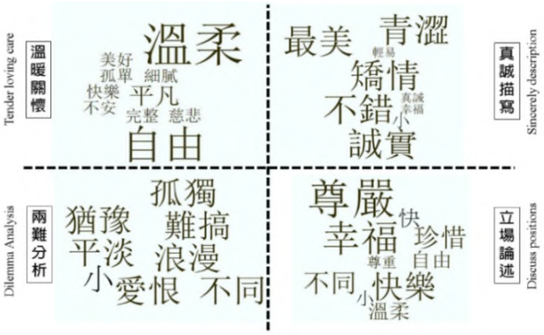

Figure 2. Topic Model of articles

In order to better understand readers’ Kansei feeling toward the different topics listed in the prior section, a Kansei semantic differential questionnaire was conducted. The questionnaire consisted of two parts.

The first part was the design object. In normal Kansei engineering, this can be furniture, residences, cars, etc. In this study, we treated articles as design objects. Then, we took each article’s topic composition as the object elements. Because we set the number of topics at four, we selected four articles in each topic cluster.

The second part is Kansei words. After reviewing the topic composition, the expert gave us eight pairs of Kansei words, as displayed in TABLE II . Each pair of Kansei words is made up of two different words.

TABLE II. KANSEI WORD PAIRS

Pairs		Pairs	
Rational	Sensible	Be healed	Be reprimanded
Sincere	Pretentious	Touching	Funny
Humor	Rigid	Irrelevant	Straightforward
Constructive	Colloquial	Helpful	Unhelpful

This semantic differential questionnaire was posted on Little Lifer’ Facebook Pages. Finally, we obtained 549 copies of the questionnaire, the validity criteria of which is described below.

B. Analysis

Initially, we needed to remove some copies of the questionnaire due to responding time constraints. There were two criteria for removing questionnaires, as explained in the following. Firstly, our questionnaire respondent needed to read four articles sequentially. We assumed that each article required about two minutes; that is to say, each questionnaire required at least eight minutes to be finished. We used quartiles according to filling time to divide the questionnaire answers into four parts. Then, we removed the questionnaires for which filling time was less than Q1 (25% was removed). For the second criterion, we removed the questionnaires for which filling time was relatively longer than the normal situation (over one hour). After reviewing the filling time, we decided to delete the questionnaires for which the filling time was longer than one hour. After these two steps, 380 valid questionnaires remained, which constituted a validity rate of about 69%.

To determine whether the mean between the different topics is the same or not, we conducted an analysis of variance (ANOVA). We set eight hypotheses as below:

$$H_{0,i}\colon \mu_{1,i} = \mu_{2,i} = \mu_{3,i} = \mu_{4,i}$$

where μ denotes the mean of each topic, for which the range of each topic was between 1 to 4. The variable, i, denotes each question for different Kansei words, where the range of i is from 1 to 8. The value of significant of each hypotheses are below 0.05.

As displayed in Figure 3, we can find that the four topics have different Kansei feeling towards the audience. For example, topic 2 (sincerely description) is more touching than topic 1 (tender loving care). Namely, if the writer wants to give the reader a touching feeling, he or she can choose sincerely description as the main topic.

In order to predict whether the Facebook post would be popular or not, we used SLDA to predict the like rate. At first, we used the interquartile range to discretize the Y value (like rate) into four labels, as in TABLE III. Then, we used different numbers of topics and parts of speech in the corpus as the control variables, respectively. The results were validated by 10-fold cross validation, a common validation method adopted in the data mining area. The experiment setting is exemplified with SLDA Experiment Setting - 1 in TABLE IV.

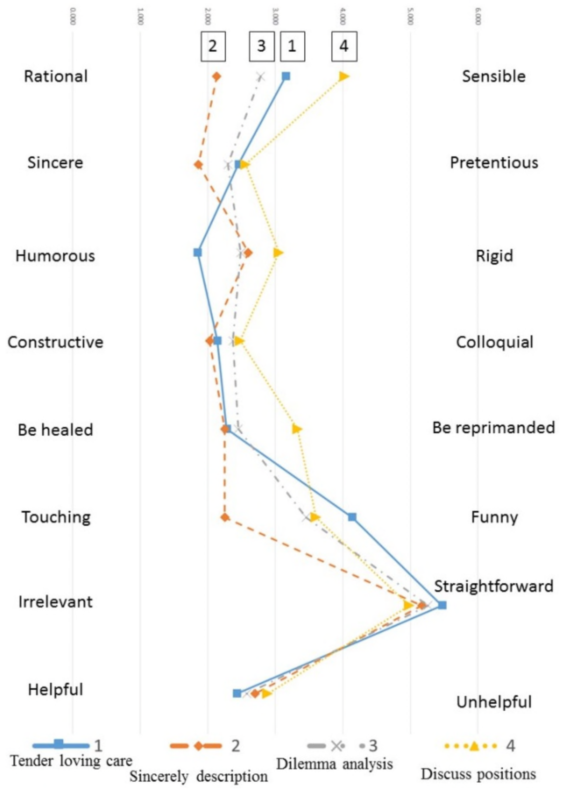

Figure 3. Topic-Kansei Relationship

TABLE III. LABELS OF Y VALUES

Interquartile range	Percentage	Labels
Q1	0%~25%	Very low
Q2	26%~50%	Low
Q3	51%~75%	High
Q4	76~100%	Very High

TABLE IV. SLDA EXPERIMENT SETTING - 1

Experiment name	adj_4type	all_4type
Number of topics	3 ~ 55	3 ~ 55
Part of speech	Adjective	Use all terms
Validation method	10-fold cross validation	10-fold cross validation

The results are presented in Figure 4. As can be seen, when the number of topics was relatively low, using adjectives can yield better performance. Moreover, as the number of topics increased when all terms were used, the value of R-square also increased. Accordingly, the more words we used, the greater the number of topics needed when using the SLDA model.

The second experiment compared SLDA, support vector regression (SVR), and linear regression, the experiment settings of which are shown in TABLE V.

As shown in Figure 5, SLDA outperforms SVR and linear regression. As such, we conducted a hypothesis test between SLDA and SVR, where μ represents the mean value of R-square. Analysis results showed that the p-value was smaller than 0.05 in TABLE VI and so H_0 was rejected. In other words, SLDA had better performance than SVR.

$$H_0\colon \mu_{slda} \le \mu_{svr}$$
$$H_1\colon \mu_{slda} > \mu_{svr}$$

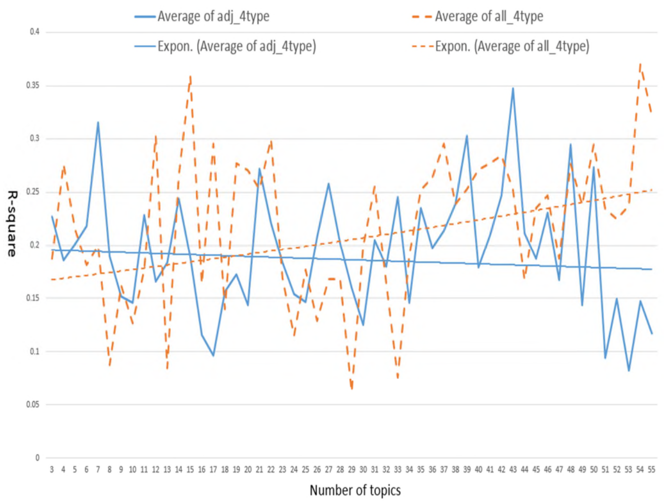

Figure 4. Results of SLDA experiment setting - 1

TABLE V. SLDA EXPERIMENT SETTING - 2

Experiment name	SLDA	SVR	LR
Predicting method	Supervised LDA	Support vector regression	Linear Regression
Number of topics	3 ~ 55	3 ~ 55	3 ~ 55
Part of speech	Adjective	Adjective	Adjective
Validation method	10-fold cross validation	10-fold cross validation	10-fold cross validation

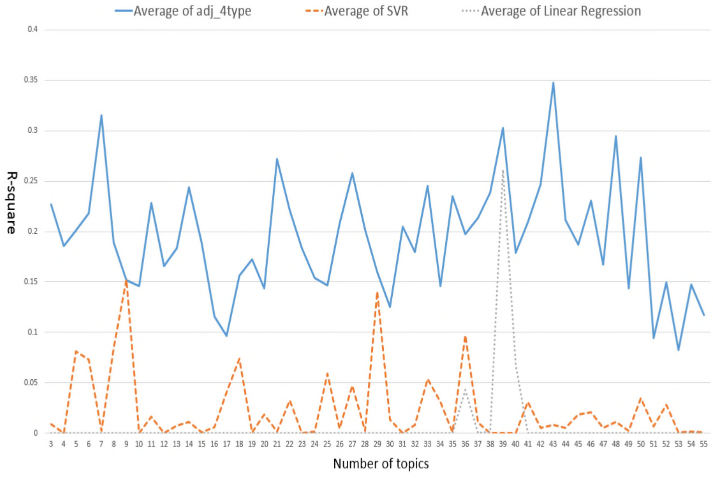

Figure 5. Comparison between SLDA, SVR, and Linear Regression

TABLE VI. ANOVA TABLE OF SLDA AND SVR

Source of Variation	SS	df	MS	F	P-value
Between Groups	0.771	1	0.771	3.932	**8.250E-35**
Within Groups	0.232	104	0.002		
Total	1.003	105			

V. CONCLUSION AND FUTURE WORK

A. Conclusion

Topic modeling is a widely applied technique in many areas for finding the latent topics of documents. Kansei engineering is a method aiming at the development or improvement of products and services by translating the customer's psychological feelings and needs into the domain of product design. In this study, we combined LDA and Kansei engineering to analyze the need of audience of social media fan page. Firstly, we collected numerical data and textual data from both Facebook and PTT. Secondly, LDA is used to generate topics with different contents. The contents were divided into four groups, including only nouns, only adjectives, and sets of nouns, verb and adjectives, and all terms. We found that using only adjectives generated the best topic models. After discussing the results with the expert, we included four topics in our corpus. The four topics were tender loving care, sincere description, dilemma analysis and discuss positions. Thirdly, we chose articles within different topics and conducted semantic differential questionnaires. In this step, a topic-Kansei relationship model was built, so we could know reader's Kansei feeling given any new articles. Finally, an SLDA model was built to predict whether a post would be popular or not. We also found that using only adjectives to generate topics is superior to using all terms when the number of topics is relatively low.

B. Future Work

The main purpose of this paper was to introduce a method that helps writers become more popular. However, some issues remain to be solved in the future.

1. The corpus of this study contained only 67 articles, which is relatively small compared to most text mining research. If a larger corpus is provided, better performance on topic modeling would be obtained.
2. The proposed model in this study was tested with only one writer. This model should be applied to other writers in the future to determine its validity.
3. The topic model was generated by one writer. In the future, a holistic topic model could be built using a group of different writers.

ACKNOWLEDGMENT

This study was supported in part by the Ministry of Science and Technology, ROC, under contract MOST 105-2410-H-143-020. The authors also thank Mr. Chien-Wei He for his help on the experimentation.

REFERENCES

[1] MIC., "96.2%台灣網友近期曾使用社交網站," Market Intelligence & Consulting Institute. Retrieved from http://mic.iii.org.tw/intelligence/pressroom/pop_pressfull.asp?sno=364, 2014.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation. the Journal of machine Learning research," vol. 3, pp. 993-1022, 2003.

[3] W. McKnight, "Text Data Mining in Business Intelligence," Information Management Magazine. Retrieved from http://www.information-management.com/issues/20050101/1016487-1.html, 2005.

[4] O. Popescu, and C. Strapparava, "Time corpora: Epochs, opinions and changes," Knowledge-Based Systems, vol. 69, pp. 3-13, 2014.

[5] M. Nagamachi, "Introduction of Kansei engineering," Japan Standard Association, Tokyo, 1996.

[6] M. Nagamachi, "Kansei engineering as a powerful consumer-oriented technology for product development," Applied ergonomics, vol. 33(3), pp. 289-294, 2002.

[7] Y. Matsubara, and M. Nagamachi, "Hybrid Kansei engineering system and design support," International Journal of Industrial Ergonomics, vol. 19(2), pp. 81-92, 1997a.

[8] Y. Matsubara, and M. Nagamachi, "Kansei analysis support system and virtual kes," Kansei Engineering I, Kaibundo, pp. 53-62, 1997b.

[9] C. Llinares, and A. Page, "Application of product differential semantics to quantify purchaser perceptions in housing assessment," Building and environment, vol. 42(7), pp. 2488-2497, 2007.

[10] C. Llinares, and A. F. Page, "Kano's model in Kansei Engineering to evaluate subjective real estate consumer preferences," International Journal of Industrial Ergonomics, vol. 41(3), pp. 233-246, 2011.

[11] T. Hofmann, "Probabilistic latent semantic indexing," Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.

[12] D. Mrva, and P. C. Woodland, "Unsupervised language model adaptation for Mandarin broadcast conversation transcription," Paper presented at the INTERSPEECH, 2006.

[13] Y. C. Tam, and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," Paper presented at the INTERSPEECH, 2005.

[14] D. M. Blei, and J. D. Lafferty, "Dynamic topic models," Paper presented at the Proceedings of the 23rd international conference on Machine learning, 2006.

[15] J. D. Mcauliffe, and D. M. Blei, "Supervised topic models," Paper presented at the Advances in neural information processing systems, 2008.

[16] M. Steyvers, and T. Griffiths, "Probabilistic topic models," Handbook of latent semantic analysis, vol. 427(7), pp. 424-440, 2007.

[17] C. E. Osgood, G. J. Suci, ,and P. H. Tannenbaum, "The Measurement of Meaning," Urbana, IL: University of Illinois Press, 1957.