Paper 269-2013

Predicting Application Review Rating with SAS® Text Miner

Tianxi Dong, Rawls College of Business Administration, Texas Tech University
Jonghyun Kim, Rawls College of Business Administration, Texas Tech University
Zhangxi Lin, Rawls College of Business Administration, Texas Tech University

ABSTRACT

With the wide proliferation of text-based data on the Internet, there is a need for dealing with the information overload. The large amount of online user reviews may present an obstacle to developers who want to know users' feedback and potential customers who are interested in applications. Here we employ text analysis provided in SAS® Text Miner to predict the overall and feature-based ratings for online application reviews. We use examples from Android Market and Apple Store, the real world of online application stores. The findings may aid in promoting the sales of applications by better satisfying customer demands.

Keywords: online reviews, application store, text mining, Android Market and Apple Store

INTRODUCTION

Recently with the popularity of smart phones and tablet personal computers, thousands of application developers, working as additional service providers, have emerged and a so called producers-consumer network (Toffler 1980) has flourished. The sale of applications is important for both developers and application providers such as the Android Market and Apple Store. For the hardware manufacturers, applications attract end-users and thus drive hardware sales. The business model for application sales operates in this way: application developers receive 70% of the application price, with the remaining 30% distributed among carriers and payment processors(Duan, Gu, & Whinston, 2008). Therefore, both developers and application providers are deeply concerned with ways to improve online application stores so that sales and customer satisfaction remain high.

Generally, to learn customer demands and feedback, developers and end-users can communicate through user forums or the websites of application stores. For example, on the websites of application stores, users can rate and write reviews for different applications. Based on the ratings, application websites provide the rankings of applications in different categories. The rating information, however, is not informative enough for developers to improve the applications. Developers must also understand the textual content of user reviews. Many users provide the reasons why they do not like the application, for example, "Game is awesome! However, needs more levels and lags quite often.[1]" But the textual content of user reviews has at times been ignored (Cao et. al., 2010). The developers can improve the application by increasing levels and reducing the lags. If we only focus on the star rating, no one will know how to improve this application or why this application is so popular.

Although reading online reviews may help developers to improve applications, in many cases the large quantity of textual reviews available for an application can be overwhelming, thereby impeding the developers' ability to track the key information. For example, the number of reviews for any average application on the Android Market and Apple Store can exceed several hundreds. As of December, 2011, the number of reviews for the popular application *Angry Birds*, had reached 959,140 (cited from Android Market website).

To address the problem of information overload, websites have added a helpful voting mechanism that may encourage users to evaluate the helpfulness of user reviews by simply asking anyone who has read the review to vote on the question, "Was this review helpful to you?." Yet, the degree to which this mechanism may aid the reader remains limited. Even for the most popular online application stores, such as the Apple Store and the Android Market, a large portion of application reviews do not receive any helpfulness votes. Therefore, with a lack of helpfulness votes, reader decision-making facilitated by the helpfulness votes will be limited in effect.

In order to assist the developers to more easily find helpful review information to improve the application, a text mining based method of summarizing online reviews was investigated. Our objective is to help developers to better master customer demands and thus promote the sales of applications by employing the text mining approach.

The present study contributes to the eBusiness research discussion by proposing a text mining analysis method with SAS® Enterprise Miner to identify the sentiment words related to each topic and predict the feature rating and overall rating based on reviews. In this way, it may be easier for developers to find relevant and meaningful information to improve the application, thus increasing the application's sale. This research also contributes to the text-mining research literature by using a text-mining approach in the study of application reviews. To the best of our knowledge, extant studies (Ghose & Ipeirotis, 2011; Li, Li, & Lin, 2008) of reviews on commercial websites used the data mainly

---

[1] Retrieved May 10, 2011, from

https://play.google.com/store/apps/details?id=com.rovio.angrybirds&feature=search_result&hl=en

from commercial websites such as Amazon.com and Ebay.com. However, we noticed that few prior studies drew on the data from application stores such as the Android Market and Apple Store, which are newly emerging provider-user networks. To fill this research gap, the proposed approach in this paper empirically examined real data from the Android Market and Apple Store.

The remainder of this study is presented as such: Section 2 reviews literature relevant to the presented thesis. Data collection techniques and other methodologies are discussed in Section 3. Empirical results are presented in Section 4. Section 5 considers the limitations of current research while proposing meaningful avenues for further exploration.

RELEVANT WORK

ANALYSIS OF ONLINE REVIEW BY TEXT MINING

There are several studies that examine the detailed textual information contained within the online reviews (Snizek et al. 1979; Danescu-Niculescu-Mizil, et al.

2009). Content analysis was utilized to identify the factors affecting the evaluative content of book reviews in sociology (Snizek et al. 1979). As this technique is extraordinarily time-consuming when dealing with a large amount of data, such as the online book reviews, text mining is gaining attention in information systems research. Opinion-mining and sentimental-analysis have previously been used to analyze the textual information found in Amazon book reviews (Danescu-Niculescu-Mizil, et al. 2009). Danescu-Niculescu-Mizil, et al. found that the perceived helpfulness of a review depended on its content as well as how the evaluation related to other evaluations of the same product. Cao et al. (2011) employed latent semantic analysis to uncover the semantic factors that may influence the number of helpfulness votes given to a particular review.

Several studies have extracted users' opinions from online reviews to predict products' sales (Forman, et al. 2008) and examined the effect of semantic characteristics on the helpfulness votes (Danescu-Niculescu-Mizil, et al. 2009, Cao et al. 2011). However, few studies have realized the usefulness and importance of identifying the topics contained in user reviews in the application stores by automatically extracting the semantic information of reviews, thereby possibly helping developers better understand users' responses. This research explores the potential ability to utilize a text-mining technique to quantify the semantic information contained in user reviews for applications.

TOPIC MODELING

Since the introduction of the LDA model (Blei, et al. 2003), various extended LDA models have been used for automatic topic extraction from large-scale corpora. In the context of social tagging systems, where multiple users are annotating resources, the resulting topics reflect a shared view of the document; and the tags of the topics reflect a common vocabulary. As for community detection, the most representative approaches include centrality or betweenness-based approaches and graph partitioning-based approaches. Girvan and Newman extended the betweenness measure to edges and designed a clustering algorithm which gradually removes the edges with the highest betweenness value (Girvan, et al. 2002). This algorithm has been improved through modularity; and the complexity is reduced from *O (m2n)* to *O(mdlogn)* where *d* is the depth of the dendrogram of the community structure (Clauset, et al. 2004). Many studies provide various models and algorithms for topic mining and community detection; yet, few of them have integrated those models and algorithms, performed topic mining for detected communities.

Although much work has been done in the topic modeling, there remains a need to apply these topic models empirically. The abundance of information created by the online review system provides a significant obstacle to identifying the information concerned. This paper take the initiative to investigate how to identify the latent topics contained in the application review data, potentially helping developers to more easily identify useful user demand information.

RESEARCH METHODS

The research conceptual model of this research is illustrated in the Figure 1.

As shown in the Figure 1, our study is divided into three steps:

Step 1: Latent features identification in the review data: In this step, we identify the latent application features of Angry Birds by employing LDA topic model. Detailed methods can be found in Li et al. (2010).

Step 2: Manual Coding: There are two coders who independently coded each review and gave the scores for each feature. The score scale is from 1 to 5, which is consistent with the overall rating scale. We presume that 1 and 2 refer to negative review while 4 and 5 refer to positive review. 3 means the review does not mention the feature.

Step 3: Prediction with SAS® Enterprise Miner: We employ SAS® Enterprise Miner to establish the overall rating and feature rating prediction model.

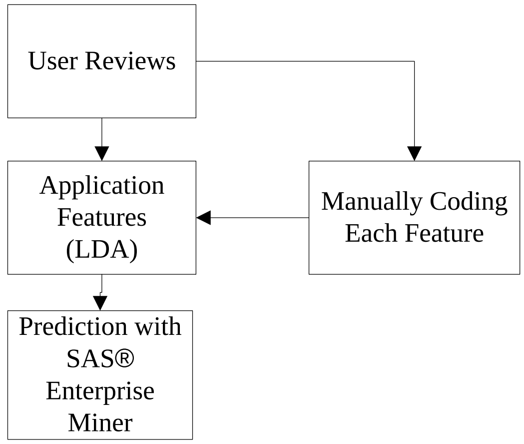


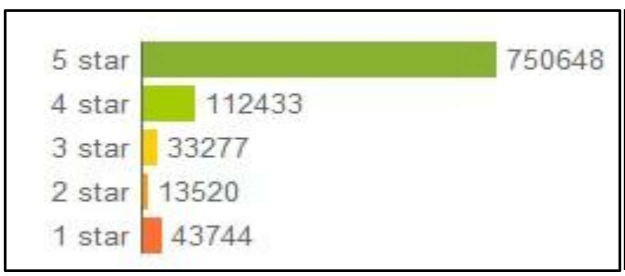
Figure 1. Research Conceptual Model

DATA COLLECTION

Data was collected from the Android Market and Apple Store, online application providers for smart phones and tablet personal computers running the Android system and iOS respectively. The Android Market and Apple Store provide an ideal study environment for this paper because the competition in the market is fierce and developers must understand users' demands.

The *Android Market and Apple Store* and *Apple Store* offer users a friendly feedback system to share their opinions and experiences about the applications. The user review system includes detailed comments and a five-star user rating system. One of the advantages of the Android Market and Apple Store's online user review system is that many users can note the shortcomings of the applications so that developers can improve them.

To collect review data, we chose Angry Birds, one of the most popular applications in the Android Market and Apple Store, as our target application. We collected the entire history of review data up to December 15, 2011 for this application (Cited from Android Market and Apple Store). The total number of user reviews was 953,619 (from Android Market) and 817,913 form Apple Store. The rating distributions of this application from two websites are shown in Figure 2.



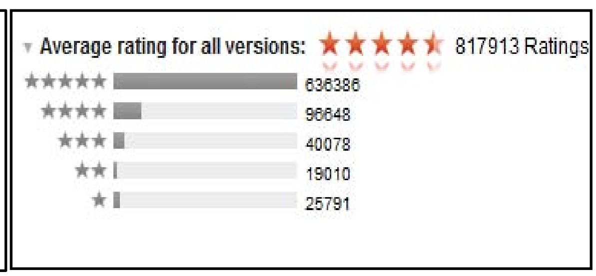


Figure 2a. Rating distribution of Android

Figure 2b. Rating distribution of iOS

*Data period: May 2010 to December 2011

As Figure 2a shows, the rating choice range is from one star to five stars, and the number of ratings below five stars reached 202,974. This illustrates that many users are not totally satisfied with this application despite the average rating being 4.6. Even for the reviews with five stars, some users still pointed out the shortcomings of the application (The example illustrates this point was introduced in the first section). For the data collected, each record consisted of reviewer's ID, the title of the review, overall rating and the detailed reviews.

TOPIC MODELING METHOD

Statistical topic modeling has been proposed as an unsupervised method to summarize the contents of large document collections. The classic model is called Latent Dirichlet Allocation (LDA) (Blei, et al. 2003). These models and their extensions use simple surface features such as word occurrences within documents to reveal the semantic content of documents. In this study, we employed LDA to identify the primary topics contained in the online review data set. LDA is a model for topic discovery (Blei, et al., 2003). In LDA, each document (here "document" refers to individual reviews) is viewed as a mixture of various topics, where a topic is a probability distribution over words.

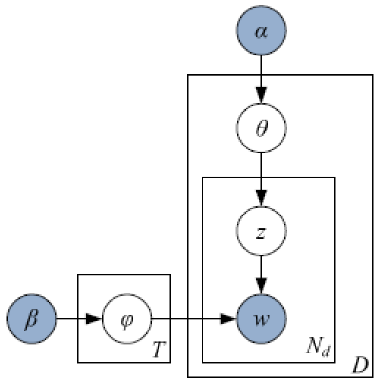


Figure 3. LDA Model

where α is the parameter of the Dirichlet prior on the per-topic word distributions; β is the parameter of the Dirichlet prior on the per-document topic distributions; θ is the topic distribution for document D; φis the word distribution for topic T; Z is the topic for the word W in the document D.

As shown in Figure 2, the existing framework of LDA has three hierarchical layers, where topics are associated with documents, and words are associated with topics. Latent Dirichlet Allocation maintains a statistical foundation by defining the topic-document distribution, which allows inferences on new documents based on a previously estimated model and avoids the problem of over fitting, which is known as a deficit of Probabilistic Latent Semantic Indexing (pLSI). The procedure of generating each word in a document under LDA can typically be broken down into two stages. One first chooses a distribution over a mixture of K topics. Following that, one acquires a topic randomly from the topic distribution and draws a word from the topic according to the topic's word probability distribution.

Our study attempts to identify the latent topics contained in each review by employing the LDA model. The process is introduced below.

Preprocessing

Preprocessing was performed on the Android Market and Apple Store review data before the subsequent text analysis. Punctuation, numbers and other non-alphabetic characters were removed first. Then for reducing the vocabulary size and addressing the issue of data sparseness, stemming was performed using Porter's stemmer algorithm. Stop words were also removed based on a stop word list.[2]

Latent Features

We employ the LDA method in (Li et. al, 2010). The latent topics identified from our dataset are listed below:

Entertainment: This feature is about the fun that is brought by the game *Angry Bird.* The key words related to this feature include: *fun*,

Quality: This feature talks about the game's graphics and sound quality. The key words related to this feature include: *graphics, screen,* sound.

Performance: This feature refers to the game's speed and update situation. The key words related to this feature include: *update, speed, space.*

Interface: This feature is about how the game interacts with users, for example, if users complain with the advertisements popping up during the game. The key words related to this feature include: *advertisement.*

TEXT MINING WITH SAS® ENTERPRISE MINER

Manual Coding

Two people coded scores of each feature independently based on reviews. The final result of the scoring is the average of two independent scoring. Score 1 and 2 indicate negative scores, 4 and 5 indicate positive scores for each feature. Score 3 indicates that review does not mention the feature. In addition, for increasing accuracy of scoring, the two coders discussed and got consensus on records that have larger difference than the threshold.

[2] http://ir.dcs.gla.ac.uk/resources/linguistic utils/stop words/

	A	B	C	D	E	F	G	H	I
1	ID	OS	Ver	Entertainment	Quality	Performance	Interface	Rating	Customer_Reviews
2	1	iOS	2.3.0	3	3	1	3	2	Why would you bother re leasing an update that doesn't support the new phone?
3	2	iOS	2.3.0	3.5	3	1	3	1	What exactly are the developers doing? The biggest game on the iPhone and still no support for the iPhone 5.
4	3	iOS	2.3.0	3	1	2	3	3	I don't have the iPhone 5 so ... But I just downloaded the new update and it made the whole game very low quality so I don't s
5	4	iOS	2.3.0	4	1.5	3	3	3	The new bad piggies update just came outfor Angry Birds! I was exited! Until I updated it and it toteld my graphics! Fix it!
6	5	iOS	2.3.0	4	2	2	2	3	C'mon guys ... one of the most popular apps on the App Store and several weeks after the iPhone 5 is released Angry Birds sti
7	6	iOS	2.3.0	3.5	3.5	1	3.5	1	Will give five stars when updated
8	7	iOS	2.3.0	3.5	3.5	1.5	3.5	4	5 stars when you can update for new iPhone 5 screen.

Figure 4. Examples of Scored Dataset

Prediction Model

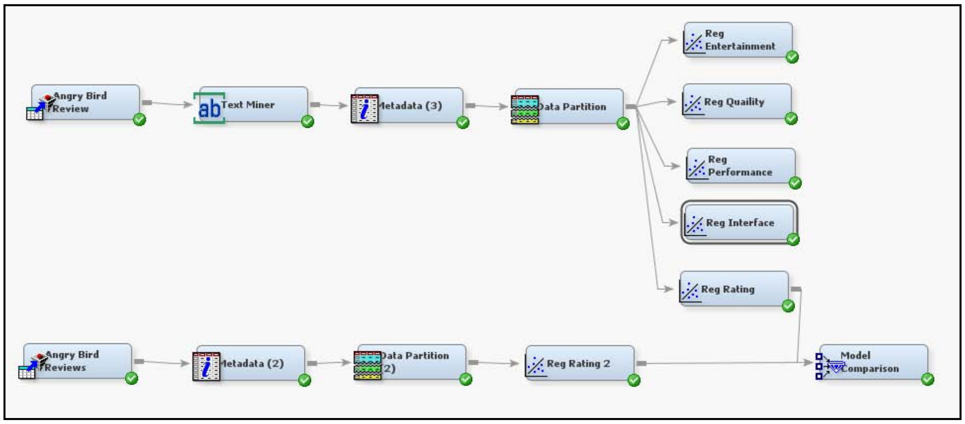


Figure 5. Prediction Model

SAS® Enterprise Miner is used to set up a text analysis project. The software component gives flexibility in terms of setup and storage so that you can identify the location of the project and the corresponding data sets. The software uses a java interface, allowing you to use point-and-click access once the project has been set up(Cerrito, 2011).

RESULTS AND DISCUSSION

The analysis of the latent topics identification from the user reviews of *Angry Birds* is listed in Table 1.

Table 1. Results Summarization of Text Mining Analysis

Application Features	Related Positive words	Frequency	Related Negative words	Frequency
Entertainment	love	49	awful	4
	great	16		
Quality	mighty	5	bad	31
			terrible	21
Performance			late	3
			sad	4
Interface	amazing	2	annoying	4
			difficult	2

As shown in Table 1, four functional features and their related sentiment words have been identified. It seems that users are really satisfied with the Entertainment brought by this application. There are such positive words as “Love” and “Great” for this feature. The frequency for each word is 49 and 16 respectively. There is only one negative word “Awful” related to this feature. *Quality* refers to the game’s graphics and sound quality. Two main negative words relate to this feature are “Bad” with frequency of 31 and “Terrible” with frequency of 21. As for *Performance,* users are also not satisfied, which can be illustrated by the negative words such as “Late” and “Sad.” Users complained about the recent performance of the game since there is no new version for running on the new iOS. Most of them feel sad about the game. *Interface* refers to all the comments about advertisements on the application. We can see that more negative words are related to this feature than positive words, which means that users are annoyed by the advertisements on the application. One reviewer wrote, “I would quite happily pay to update and get rid of the *annoying ads* (ads is the abbreviation of advertisement based on the context).” The negative sentiment can be reflected by such words as “Difficult” and “Annoying”.

We can see that Quality*, Performance,* and *Interface* have more negative words than do *Entertainment.* Overall users talked more about the poor graphics, late update and the advertisements appearing on the application. This makes sense in that recently a new iOS has been issued but the *Angry Birds* is not compatible with the new operation system. Angry users tended to comment on the disadvantages of the application’s bad performance and poor graphics quality brought by the incompatibility. Users’ reviews are really helpful for developers to know how to improve the target application.

To further examine the words related to each feature, another feature of SAS Text Miner is to use the visualization of concept links to see how terms are linked in these short definitions. Figure 6a to 6d separately show the terms related to Game (Entertainment Feature), Graphics (Quality Feature), Update (Performance Feature) and Advertisement (Interface Feature). Figure 6e is the terms related to the game *Angry Birds*.

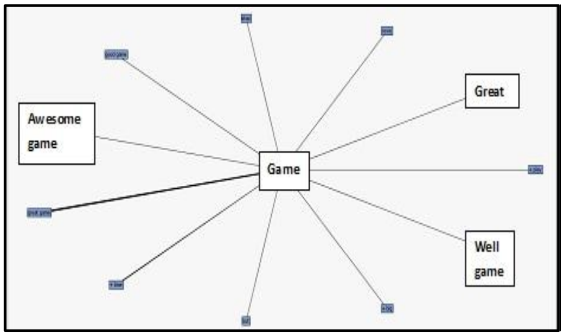


Figure 6a. Terms Related to Game

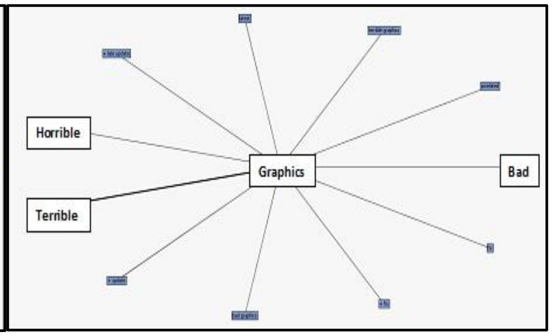


Figure 6b. Terms Related to Graphics

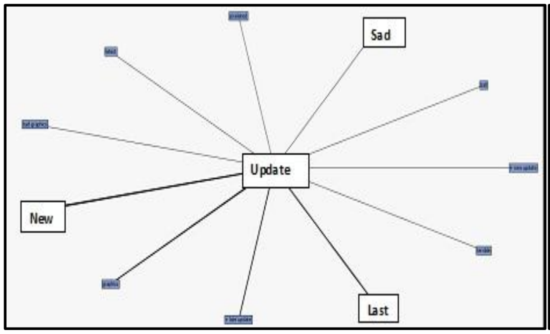


Figure 6c. Terms Related to Update

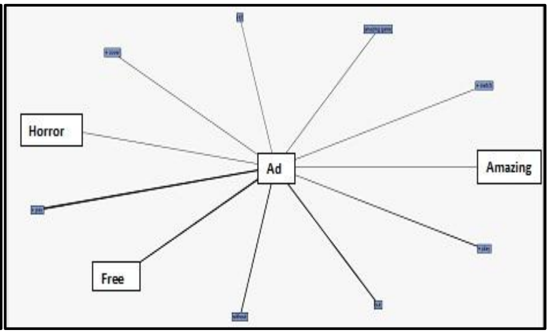


Figure 6d. Terms Related to Advertisement

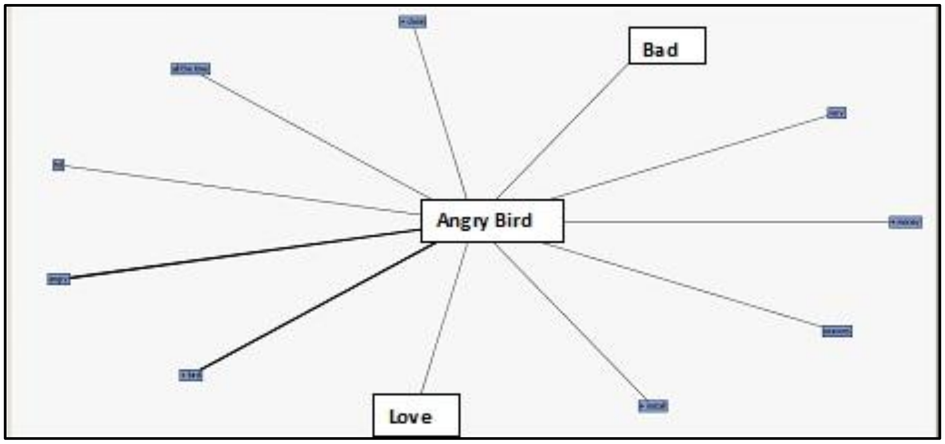


Figure 6e. Terms Related to *Angry Bird*

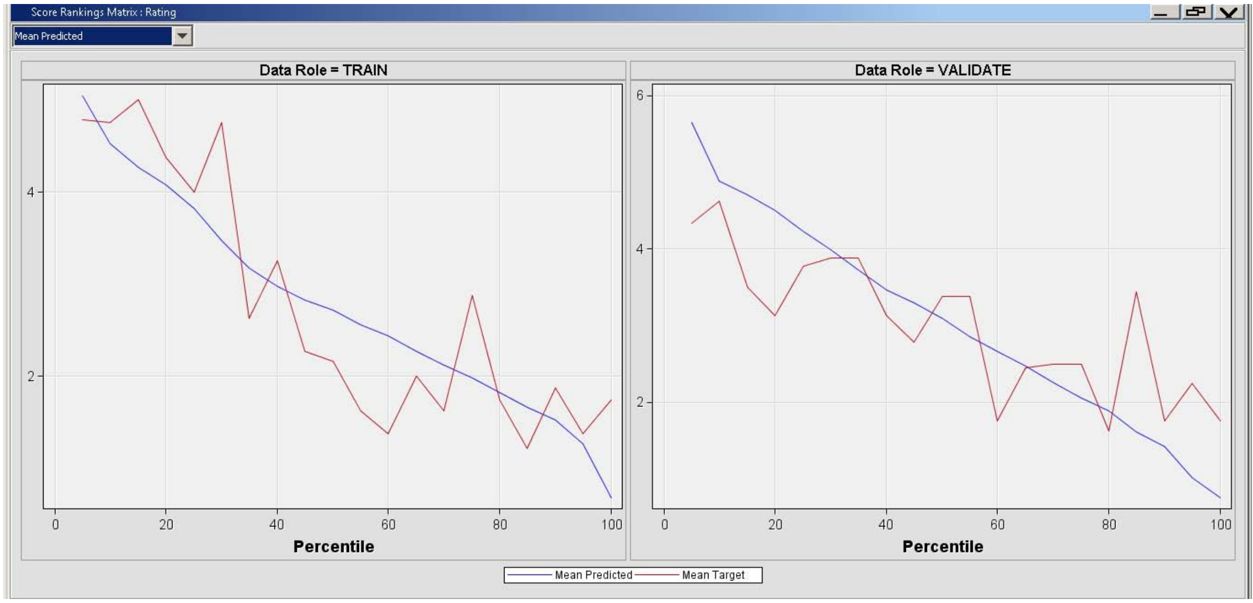


Figure 7. Performance of Prediction Model

Figure 7 shows the result of prediction model for overall rating using a linear regression model. SAS® Enterprise Miner analyzed reviews using a text mining method. Using the result of the text mining, our linear regression model predicts the overall rating of the review. The left matrix indicates mean average of predicted and target values of training data, and the right matrix indicates mean them of validation data. We can find the performance of this prediction model to calculate average square error between predicted and target values. The overall performance of our predication model is not as good as we expected. The possible reason for this could be the small dataset size. Some of the review is too short to catch the textual information.

CONCLUSION

In this paper, we examined a previously ignored yet important research question concerning the online user reviews: How can we enhance the communication between application developers and users? We addressed this question by investigating a text mining based method for mining and summarizing the users' opinions about the applications, which are contained in the online reviews. We first identified the features users mostly mentioned in the online reviews by employing LDA algorithm. Then the feature-related sentiment words are captured by SAS® Enterprise Miner and we further established a model to predict the feature rating and overall rating based on reviews with SAS® Enterprise Miner. A number of practical and research implications can be derived from this study.

This study is complementary to the previous work on online user reviews. Previous work on online reviews made an effort to predict the relationship between online reviews and product sales. This study examined how to identify the critical information contained in the online reviews by employing text mining method. The major contribution of this study is to add to understanding on how to identify the features users most frequently mentioned in the reviews and the corresponding sentiment towards these features.

This study also has significant implications for website designers in that it can guide them in designing multi-dimensional rating mechanisms that may satisfy users' diversified demands. The application websites may provide multi-dimensional rating mechanisms based on the most frequently mentioned application features in the reviews. For example, if users frequently mention "Advertisement" in the reviews, it will help users make a decision when there is a rating on the advertisement. Users who do not like advertisements popping up during the games may not choose this application if they find the rating on this dimension is low (which means there are too many advertisements during game). Actually, the application of the proposed approach is promising, including words association rule identification in linguistics, topics identification in dataset and so on.

REFERENCE LIST

1. Blei, D.M., Ng, A.Y., & Jordan, M., I (2003) Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
2. Blei, D. M. & Lafferty, J. D.(2006). Dynamic topic models. The Proceedings of the 23rd international conference on machine learning (ICML2006), page 113-120, Pittsburgh, Pennsylvania, USA.
3. Cerrito, P. (2011). Sentiment Mining Using SAS® Text Miner®. Paper presented at the SAS Global Forum 2011, Las Vegas.
4. Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large network. Physical

Review E, 70, 066111
5. Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? — An empirical investigation of panel data. Decision Support Systems, 45(4), 1007-1016. doi: 10.1016/j.dss.2008.04.001
6. Erdos, P., & Renyi, A (1959) On random graphs. I, *Publicationes Mathematicae, 6*, 290-291.
7. Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. Knowledge and Data Engineering, IEEE Transactions on, 23(10), 1498-1512.
8. Girvan, M., & Newman, M. E. J (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 9912, 7821–7826.
9. Hu M.Q. and Liu B. (2004), Mining and Summarizing Customer Reviews, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177.
10. Iwata, T., Yamada, T, Sakurai Y, & Ueda, N. (2010). Online Multiscale Dynamic Topic Models. The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 7.
11. Li, D., He, B., Ding, Y., Tang, J., Cassidy, S., Qin, Z., Yan, E., Li, J., and Dong, T. Community-based Topic Modeling for Social Tagging. In Proceedings of the Nineteenth Conference on Information and Knowledge Management (CIKM'2010). (short paper).
12. Li, D., Li, J., & Lin, Z. (2008). Online consumer-to-consumer market in China – A comparative study of Taobao and eBay. Electronic Commerce Research and Applications, 7(1), 55-67. doi: 10.1016/j.elerap.2007.02.010
13. Milgram, S (1967) The small world problem, Psychology Today, 2, 60–67.
14. Newman, M. E. J (2001) The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the United States of America, 982, 404-409.
15. Newman, M. E. J (2003) The structure and function of complex networks. Siam Review 45, 167–256 .
16. Nikolay Archak, Anindya Ghose, Panagiotis G. Ipeirotis, “Deriving the Pricing Power of Product Features by Mining Consumer Reviews,” *Management Sciences*, Vol. 57, No. 8, August 2011, pp. 1485–1509.
17. Shirky, C. (2005). Ontology is overrated: Categories, links, and tags. Retrieved May 20, 2010 from http://shirky.com/writings/ontology overrated.html.
18. Tan, C., Tang, J., Sun, J., Lin, Q and Wang, F. Social Action Tracking via Noise Tolerant Time-varying Factor Graphs. In Proceedings of the Sixteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2010). pp. 1049-1058.
19. Toffler, A. (1980): The Third Wave, London et al. (Collins).
20. Torunski, L. (2009). Smart and simple webcrawler. Retrieved May 20, 2010 from https://crawler.dev.java.net.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name	Tianxi Dong
Enterprise	Texas Tech University
Address	Rawls College of Business Administration, Lubbock, TX, 79415
E-mail	tianxi.dong@ttu.edu