What in Consumer Reviews Affects the Sales of Mobile Apps A Multifacet Sentiment Analysis Approach

Liang, Ting-Peng; Li, Xin; Yang, Chin-Tsung; Wang, Mengyue

Published in:
International Journal of Electronic Commerce

Published: 01/01/2015

Published version (DOI):
10.1080/10864415.2016.1087823

Publication details:
Liang, T.-P., Li, X., Yang, C.-T., & Wang, M. (2015). What in Consumer Reviews Affects the Sales of Mobile Apps: A Multifacet Sentiment Analysis Approach. *International Journal of Electronic Commerce*, *20*(2), 236-260. https://doi.org/10.1080/10864415.2016.1087823

This is an Accepted Manuscript of an article published by Taylor & Francis in INTERNATIONAL JOURNAL OF ELECTRONIC COMMERCE on 30 Nov 2015, available online: http://www.tandfonline.com/10.1080/10864415.2016.1087823.

What in Consumer Reviews Affects the Sales of Mobile Apps: A Multi-Facet Sentiment Analysis Approach

Ting-Peng Liang, National Chengchi University, Taipei, Taiwan*
Xin Li, College of Business, City University of Hong Kong, Hong Kong, China
Chin-Tsung Yang, National Sun Yat-Sen University, Kaohsiung, Taiwan
Mengyue Wang, College of Business, City University of Hong Kong, Hong Kong, China

* Corresponding author: Ting-Peng Liang, Email: tpliang@nccu.edu.tw.

Ting-Peng Liang is the Director of Service Innovation and Electronic Commerce Research Center at the National Chengchi University in Taiwan. He is also a Fellow of the Association for Information Systems. He received his doctoral degree in Information Systems from the Wharton School of the University of Pennsylvania and had taught at the University of Illinois, Purdue University, Chinese University of Hong Kong, and City University of Hong Kong. His primary research interests include electronic commerce, intelligent systems, decision support systems, knowledge management, and strategic applications of information systems. His papers have appeared in journals such as *Management Science, MIS Quarterly, Journal of MIS, Operations Research, Decision Support Systems, and Decision Sciences*. He also serves on the editorial board of Decision Support Systems and several academic journals.

Xin Li is an assistant professor in the Department of Information Systems at the City University of Hong Kong. He received his Ph.D. in Management Information Systems from the University of Arizona. He received his Bachelor's and Master's degrees from the Department of Automation at Tsinghua University, China. His work has appeared in the *Journal of MIS, Decision Support Systems, Journal of the American Society for Information Science and Technology, ACM and IEEE Transactions,* among others.

Chin Tsung Yang is a master student of the Department of Information Management, National Sun Yat-sen University, Taiwan.

Mengyue Wang is a PhD student in the Department of Information Systems at City University of Hong Kong. She received the M. Eng. Degree from University of Science of Technology of China. Her research interests include recommendation systems and electronic commerce. Her work has been published in the *Decision Support Systems, International Conference on Information Systems, and other international conferences.*

What in Consumer Reviews Affects the Sales of Mobile Apps: A Multi-Facet Sentiment Analysis Approach

Abstract:

With the rapid adoption of smartphones, developing mobile apps has become an attractive arena for entrepreneurs. Many factors drive the sales of mobile apps, one of which is online word-of-mouth (eWOM). This research examines the impact of textual consumer reviews on the sales of mobile apps. Noting the inconsistent findings on textual reviews' impact in previous literature, this study inspects how the sentiments of different topics in online reviews affect app sales. We develop a multi-facet sentiment analysis (MFSA) approach to measure the dimensions in consumer reviews. Specifically, we are interested in the comments on product quality and service quality in this research. Employing a real-world data set of 79 paid and 70 free apps from an IOS app store, we found that although consumers' opinions on product quality occupies a larger portion of consumer reviews, their comments on service quality have a stronger unit impact on sales rankings. The empirical analysis illustrates the value of our proposed MFSA approach for better understanding of textual consumer reviews' impact on mobile app success.

Keywords:

Mobile app, word-of-mouth, sales, text mining.

Introduction

Smartphones and mobile applications (apps) are increasing in use and popularity. IDC's 2015 survey reported that the global smartphone vendors shipped a total of 334.4 million smartphones worldwide in the first quarter of 2015, up 16.0% from the 288.3 million units in the first quarter of 2014 [34]. Mobile apps are software applications that run on smartphones to achieve certain purposes. Since its invention in 2008, a huge market for mobile applications has emerged [56] and become an attractive arena for entrepreneurs. By mid-2014, Apple announced that users had downloaded 75 billion applications, users visited the App Store 300 million times per week, and there were 9 million registered developers, up 47% from the previous year [53]. Successful mobile apps become important assets for companies. For example, in 2014 Facebook spent $19 billion (US) to acquire WhatsApp, a popular mobile app for online chatting [13].

Many factors affect the success of apps. One factor that differs from those that impact traditional software systems is online word-of-mouth (eWOM). Apps are often distributed through app stores that allow consumers to post comments about apps. As a result, consumers often consult customer reviews in making their purchase decisions.

Several marketing studies have recognized the impact of eWOM on consumers' purchases beyond inherent product and brand effects and other marketing tactics [16, 27, 73]. Chen and Xie [7] argued that consumer reviews provide product-matching information for consumers to find products that match their needs. Such supplementary information helps consumers reduce uncertainty about products and facilitates sales. In fact, previous research reported that consumer comments are more trustable than expert opinions in many cases [6].

Despite such theoretical arguments, prior studies on sentiment in textual reviews have conflicting results [72] and the effect may be negligible when controlling for numerical ratings [44]. Inconsistent findings may be due to inadequate or inaccurate sentiment analysis techniques. We argue that previous studies often investigate the overall aggregated sentiments of comments. Such an approach cannot provide more information from textual content than numerical ratings provided by users, as the aggregation process may lose some important information on different aspects of the product that are considered in consumers' decision making.

To fill this gap, we propose a multi-facet sentiment analysis (MFSA) of consumer feedback to deepen our understanding of textual consumer reviews and examine their impact on app sales. Particularly, based on previous studies, we focus on two dimensions of sentiments in app eWOM: product quality and service quality. This differentiation is important for mobile apps, since many apps rely on a server (or, cloud) to store user information and distribute product information. Service is an inherent part of apps and plays a quite different role than customer service for traditional software applications. We extract such sentiments from consumer reviews and conduct an empirical study on a data set of IOS apps. Findings from the data set support our conjecture and indicate that eWOM sentiment is a good predictor of IOS app sales, given that sentiments are differentiated to product quality and service quality.

The contribution of this article is three-fold. First, we argue that sentiment analysis on consumer reviews needs to consider different aspects of consumer concerns. An aggregative view of sentiment analysis is too coarse to fully reveal the value of consumer reviews. Second, we find that both product and service reviews have effects on app sales rankings, which differ for free and paid apps. Free app users rely more on online reviews and care more about service quality of apps. Third, we propose a framework to help identify multi-facets of sentiments from consumer reviews, which can be applied to other business applications.

The rest of this paper is organized as follows. Section 2 reviews previous literature on eWOM and text mining. Section 3 presents the theoretical basis and our proposed framework for analyzing eWOM sentiment related to product quality and service quality. Section 4 describes the data source and research methodology for evaluating the method. Section 5 shows the findings from our empirical analysis. Section 6 concludes the paper and summarizes the limitations and future research opportunities.

Related Work

eWOM in Electronic Commerce

In marketing literature, WOM has been well recognized as influencing consumers’ purchasing behavior [59]. Cunningham [14] pointed out that consumers are likely to generate conversations related to products and to request information from friends and relatives if they are not sure about a purchase. Bone [5] found that WOM influences short-term and long-term product judgments, especially when a customer faces uncertainties.

The development of Web 2.0 and the rapid growth of electronic commerce allow customers to share opinions about products and services online, which creates lots of eWOM [31]. Many scholars consider eWOM as a determinant of product success [16, 27, 30, 73] that is moderated by the characteristics of products [73] and consumers [71]. External WOM sources have been found to have significant effect on retail sales [29]. Recent studies also analyzed the interplay between online consumer reviews and recommender systems in consumers' decision making [3] and the formation of helpfulness of online product reviews [41].

As argued by [2], eWOM can convey the reputation of the product, the brand, and complementary goods. Such reputation can be conveyed in both the volume and valence of eWOM [1]. Volume means the amount of eWOM, such as the number of online reviews. It reflects the popularity of the product. In prior studies, Godes and Mayzlin [26] found that the volume of eWOM has a positive impact on TV show viewership. Liu [44] and Duan, Gu and Whinston [21] both showed that the volume of eWOM has a significant impact on movie box office revenue.

Valence is an affective indicator to show whether the reviewer’s sentiment is positive or negative. Since eWOM is often anonymous, consumers are more comfortable sharing both negative and positive opinions. Recent research reports that affective factors have significant influence on the adoption of eWOM (e.g., [23, 60]). Review sentiment can often be reflected through numerical ratings provided by users. Previous research has drawn consistent conclusions about the impact of rating valence. Chevalier and Mayzlin [8] showed that consumer ratings significantly influence book sales at Amazon.com. Dellarocas, Zhang and Awad [17] added online review ratings to a basic forecasting model and found its accuracy was significantly improved. Through experiments in a mobile app setting, Huang and Korfiatis [33] found that review valence and consistency alter the emotional process during trial attitude formation but do not affect the cognitive process. They identified the moderating role of online reviews on product trial experience, which in turn influences the formation of product attitudes.

Online reviews convey more information than reputation. Chen and Xie [7] argued that consumer reviews provide product-matching information that helps consumers find products that match their needs. Such supplementary information helps consumers reduce uncertainty about products and facilitates sales. As a result, the extent of subjectivity, informativeness, and readability of reviews are found to influence sales [25]. Berger, Sorensen and Rasmussen. [4] show that even negative reviews may have positive effects on sales, since they may increase product publicity, especially for lesser-known products. In addition, the variances in review ratings [12] and the diversity of textual reviews [72] are found to positively affect product sales since they provide extra information. The effect of rating dispersion varies on hedonic vs. utilitarian products, where highly dispersed ratings are perceived more positive on hedonic products [11].

<Table 1 approximately here>

However, the findings on how textual comments affect sales have not been fully illustrated through textual valence. For instance, Liu [44] studied the influence of WOM on movie box office revenue based on manually coded textual review sentiments. Liu et al. [45] further studied review sentiment using text-mining techniques. Both studies found that the textual-based measures did not affect product sales. In Table 1, we summarize previous studies related to the sales impact of eWOM, which show inconsistent findings on textual valence.

We project one reason is that previous studies investigate the overall aggregated sentiments of comments without differentiating the multiple dimensions their content. Such an approach cannot provide significantly more information from textual content than numerical ratings provided by users. Previously, Park and Kim [51] and Park and Lee [52] classified eWOM into attribute-centric and benefit-centric. The attribute-centric reviews provided additional information on product features, while benefit-centric reviews focus on emotional and subjective recommendations. There may be different emotions associated with different aspects of a product in product reviews. The aggregation process may lose some important supporting information in consumers' decision making. Analysis of eWOM at a finer granularity could provide new insights, which is the research gap we want to fill in this research in the context of mobile apps.

Text Mining and Opinion Analysis

Text mining is an effective and efficient method to automatically process the large number of textual comments that consumers read (and write). Product features can be extracted from online comments (e.g., [64]). Opinion analysis based on natural language processing [50] has been applied to summarize opinions from textual content.

There are two primary methods for opinion analysis: learning-based and lexicon-based methods. Learning-based methods use machine-learning techniques such as probabilistic models and SVM to build classification models. Many previous studies focused on positive/negative sentiment classification [15]. For example, Turney [63] presented a PMI-IR algorithm to calculate the semantic orientation of phrases based on their association with the two human-selected seed words (poor and excellent) in a large corpus. Pang and Lee [48] used a graph-cut approach to classify document sentiments. Some studies, such as Pang and Lee [49], inferred people's attitude to a multi-point scale. Yang et al. [66] applied association rules and a naïve Bayes classifier to identify the sentiment of consumer reviews and found good accuracy.

The learning-based method has a major constraint. It needs manually coded training data to build learning-based models, which is often unavailable in eWOM studies. Thus a lexicon approach, which uses lexicons and predefined rules to annotate sentiments of terms in text, is often easier to apply [72]. For example, Hu and Liu [32] built a seed list with a set of common adjectives (e.g., positive adjectives include great, fantastic, nice; negative adjectives include bad, dull, tardy), then used WordNet [10] to determine the semantic orientation for each opinion word. Demers and Vega [20] used a lexicon approach to measure the tone of news for firm valuation.

Another issue related to this research is that Chinese language processing is different from that of English. One major challenge of processing Chinese texts is to segment sentences to words [68]. In computational linguistics, several text segmentation methods have been proposed to address this problem, such as Chinese Knowledge Information Processing [46]. Another challenge for Chinese opinion analysis is the lack of established Chinese sentiment lexicons. In order to handle these challenges, Zagibalov and Carroll [69] proposed an unsupervised classification method to build sentiment lexicons from a small seeding lexicon. Taking a learning-based paradigm, Ku, Huang and Chen [38] employed morphological and syntactic structures to analyze opinions in Chinese words and sentences. Xu et al. [65] proposed a procedure to annotate opinions from online Chinese product reviews. These studies provide basic techniques to handle Chinese eWOM in our study.

Theoretical Basis and Hypotheses

To study the multiple facets of eWOM sentiments, we need to find a perspective to differentiate topics and illustrate the impact of their related comments on app sales. In this research, we take a theory-driven approach and focus on the differentiation of product quality (such as the functional correctness and usability of the app) and service quality [35, 40] (such as service reliability, information product quality, monetary value, customer service responsiveness, etc.). The two aspects of IT artifacts have been widely studied in previous research [35, 40]. There are a few theoretical lenses that support the differentiation of product and service as two distinct dimensions that affect consumer decisions.

First, product-based view and service-based view are two viewpoints on software [62]. A traditional view of software, especially consumer software, is that it is an off-the-shelf product, where firms design and develop applications and consumers hold, use, and maintain the software together with associated data. With the development of the Internet, the concept of software-as-a-service (SaaS) is increasingly popular in the IT industry, where firms maintain some parts of software modules and/or user data on the server side for users' remote access and use. In such a context, the concept of "service" is quite different from software maintenance and customer service for traditional software applications. In this study, we want to assess the impact of eWOM sentiments of the two aspects on app sales.

In *ex ante* literature, product quality is a key construct in modeling consumer utility and behavior [55]. For regular products, one important role of online reviews is to provide information about the product to support sales [6]. Viewing a mobile app as a type of software product, its product quality comes from the development process and directly affects user satisfaction and experience [37], which need to be carefully managed [58]. In the literature on information system success in organizations, DeLone and McLean [19] also employed system quality and information quality to measure characteristics of information systems and examined their impact on users' satisfaction and intention to use. Obviously consumers' reviews on software systems could affect potential consumers' perceptions of product quality. The user comments with positive sentiments on product quality should affect people's assessment and increase their purchase intention. Thus, we hypothesize that:

H1: The sentiment of comments on product quality is positively correlated with app sales.

Service is another important dimension in marketing literature that affects consumer behavior [70]. Traditional IS literature also values service quality. In research on information system success in organizations, Pitt, Watson and Kavan [57] argued that service quality (such as user support) affected IS success, which was later included in the IS success model [18]. In IT outsourcing, service quality is one important factor affecting consumers' decisions [28]. In the context of mobile apps, many apps take such a model, either storing the user data on the "cloud" or distributing information products (such as weather broadcasts) from the backend to the users. This makes it difficult for users to separate the service element from the value they receive from the product, which is different from the traditional customer service quality often measured in IS literature [36, 54]. Thus, service quality plays a more important role in decisions on mobile app purchases than on traditional software applications. The existence of online reviews on service actually provides a rating of the service, which complements the terms of the service level agreement (SLA) [43]. The user comments with positive sentiments on service quality should affect people's assessment and increase their purchase intention. Thus, we hypothesize that:

H2: The sentiment of comments on service quality is positively correlated with app sales.

Research Methodology

In order to test the hypotheses, we develop a multi-facet sentiment analysis (MFSA) approach and collect real data from the Apple Store to evaluate the approach.

Data Set

We collected a data set on IOS app sales in Taiwan from App Annie (www.appannie.com) to test the aforementioned hypotheses. App Annie has archived information on apps, such as sales rank, prices, reviews, and version changes, in different geographic locations since 2009. The IOS app store only reports the sales rank of the top 500 bestselling apps. In this research, we collected information on the bestselling free and paid apps in Taiwan for each week in 2011. However, many apps only appeared on the leaderboard for a couple of weeks, leaving the other weeks with missing data. We thus kept apps appearing more than 35 weeks (i.e., 2/3 year) in the data set for a relatively longer panel for our study. We collected reviews that appeared in 2011 and examined their impact on sales rank. Note that in the first couple of weeks, some app's reviews were actually published before 2011, which caused some missing data. We removed such app-weeks from the data set. After cleaning, our data set contains 79 paid apps and 70 free apps.

According to FaberNove [22], IOS store app rankings are based on the weighted sum of the last four days' sales, i.e., *day k's ranking sales = day k's sales *8+ day k-1's sales *5+ day k-2's sales *5+ day k-3's sales *2.* Thus, we can safely measure the impact of reviews (and other information) from Sunday on the sales (represented by sales rank) of the next Thursday. In this paper we label them with week subscript *t-1* and *t*, respectively. We collected these average ratings and the number of reviews to capture the effect of existing numerical measures' impact on consumers. The data set is an unbalanced panel. We report the descriptive statistics of the data set in Section 5.1.

MFSA for Opinion Analysis

A key component of MFSA is an appropriate opinion analysis to differentiate the two types of facets, product quality and service quality. In our data set, user comments are in Chinese (mixed with English). There is no comprehensive lexicon of Chinese sentiment words. Therefore, we adapted a procedure as shown in Figure 1 to build lexicons and then code customers' online textual comments. This approach allows us to process a large amount of reviews after building the lexicon. Then, the coding of lexicons requires substantial effort, especially in differentiating between similar concepts. Thus, we choose to only differentiate product quality and service quality and do not get into multiple, more detailed sub-dimensions of each construct. We leave the examination of other sub-dimensions of lexicon coding to future research.

<Figure1 approximately here>

To build the lexicon, we took a semi-manual approach; we first identified the candidate words and then invited human subjects to code. We employ the 3,284 reviews of the top 10 most popular apps to build the lexicon. First, we fed the reviews to CKIP API to conduct text segmentation and POS tagging [46]. We then extracted the candidate sentiment words according to their POS. Note that the symbolic POS system developed in Taiwan is different from that developed in Mainland China or in English. Many words that are considered as adjectives in those two systems are deemed intransitive verbs (Vi), which are considered to carry most of the sentiments in a sentence [47] and employed as our candidate sentiment words. If the consumer reviews contained English terms, we considered the English adjectives as reflecting sentiment.

Then we added negation prefixes (e.g., 不/non-/ir-/dis-) to the extracted candidate sentiment words.

<Table 2 approximately here>

It is possible a word's sentiment or product/service classification could not be judged by itself. Thus, we used the heuristics [47] in Table 2 to identify phrases as the context of sentiment words to help the human subject's judgment. Note that the rules in Table 2 were used recursively, where identified noun or Vi phrases are then used to identify longer terms. According to [47], the use of contextual information significantly improves sentiment assessment in Chinese.

The process provides us a list of extracted candidate sentiment words (including the ones after adding or removing negation words). Some of the terms are associated with contexts that were extracted from the reviews. We invited 10 college and postgraduate students to code the polarity of these words (positive vs. negative) and whether they were related to product quality or service quality. In Taiwan, college students are a major consumer group of apps. We consider that their judgments of app reviews are reliable enough and reflect the perception of reviews by most app users. The coding results naturally contain some ambiguous words with inconsistent understanding across coders. To address this concern, we required mutual agreements from 7 out of 10 coders to decide the polarity and/or product/service classification. For example, if 7 out of 10 coders agree that a term is positive (while the others consider it negative or neural), the term is labeled as positive. By keeping only the high-agreement terms, we consider the majority of college student coders are able to reveal the majority of public perceptions when seeing a term.

In our coding process, the candidate sentiment words are in their original language, mostly Chinese. The small amount of English words is generally simple and easy terms that were embedded in the Chinese comments. All 10 coders are native Chinese speakers with sufficient education to understand the simple English terms in our coding process; they expressed no concerns about understanding terms. The users coded the terms independently, without exchanging ideas on the coding offline. We first asked them to judge terms without context. Then context information was provided for the terms that could not be judged. (When using the coded lexicons to assess reviews in our experiments, the terms identified with context are applied before the ones identified without context, since they represent more specific meaning of a term in a particular sentence.) Noticing the difference in term usage between hedonic (i.e., game) and utilitarian apps, we built separate lexicons for these two types of apps for the subjects to code.

From these, the coders identified 50 positive terms and 30 negative terms on utilitarian app product quality (Appendix Table 1) and 9 positive terms and 14 negative terms on utilitarian app service quality (Appendix Table 2). By providing contexts of words as a reference, the coders further identified 17 positive terms and 19 negative terms on utilitarian app product quality (Appendix Table 3) and 6 positive terms and 13 negative terms on utilitarian app service quality (Appendix Table 4). For the hedonic apps, the coders identified 53 positive terms and 14 negative terms on product quality without context (Appendix Table 5), 12 positive terms and 14 negative terms on service quality without context (Appendix Table 6), 16 positive terms and 15 negative terms on product quality with context (Appendix Table 7), and 5 positive terms and 15 negative terms on service quality with context (Appendix Table 8).

After creating the lexicons, we coded the sentiments of reviews for each app each week. After segmentation and POS tagging, the sentiment words without context can be directly annotated. For the words that need to be determined based on context, we checked if the context of the word in the review contains any of the listed phrases in our coding tables. We then aggregated the sentiment words by extending Demers and Vega [20] semantic score measures.

As argued by Lee, Ku and Chen [39] and Yao and Lou [67], adverbs strengthen the effect of positive and negative sentiments in Chinese. Thus, we considered the number of adverbs used before the sentiment words as in the following formula:

$$Score=p+0.5*adv_p-n-0.5*adv_n, \tag{1}$$

where p is the number of positive words appearing in a product's reviews; adv_p is the number of adverbs before positive words; n is the number of negative words in a product's reviews; and adv_n is the number of adverbs before negative words. We get both scores for product quality and service quality, represented as *ProdScore* and *ServScore,* respectively.

<Table 3 approximately here>

Table 3 shows the 10 online reviews for Angry Birds retrieved on August 7, 2011. These reviews were treated as one piece to measure the WOM information in this research. The text contains 10 sentiment word occurrences on product quality that can be determined without context: 好玩/Fun (2 times), 欲罷不能/can't help myself from playing it, Fun (3 times), 棒/great (2 times), good (2 times), 糟/bad. The term 動腦/think hard needs to be determined with sentence context, which is retrieved using the rule (Vi+N) from our lexicon. The phrase is a positive term related to product quality. These words are associated with 3 adverbs: 太好玩/so interesting, Very fun, 很棒/so great, which adds an extra 1.5 positive score to the total. We determined the sentiment words and their associated adverbs for service quality in a similar manner. Eventually we calculated the product quality and service quality scores for Angry Birds' reviews on August 7, 2011 as:

$$ProdScore=11+0.5*3-0-0.5*0=12.5$$

$$ServScore=1+0.5-2-0.5=-1$$

Our proposed method essentially is a lexicon-based method. Its cost mainly comes from the labor of manual coding, which also ensured the high quality of the lexicon. The general process of our approach is based on previous research that has been validated. In this study, our major purpose is to differentiate comments on product quality and service quality, which is not a common task in text mining. No gold standards exist that can directly support a learning-based method. Although it is possible to cluster the terms before coding or employ a semi-supervised approach to train machine-learning models employing a small set of coded data, lexicons built from such methods still need human validation to ensure their quality. We left such technical improvements to future research.

The Econometric Model

In this research, we collected panel data in a weekly manner and used econometric models to control the confounding factors and assess the impact of two types of textual reviews on app sales.

We employed the logarithmic transformation of sales rank as an indicator of app sales (downloads). A higher value in sales rank indicates a lower sales volume. As shown in several previous studies [24], there is a power relationship between sales rank and sales. Thus we can use *LogRank* to replace *LogSales* as a dependent variable.

The two independent variables of the studies are two types of textual sentiment measures: *ProdScore* and *ServScore*. Since the two measures are calculated based on the last 10 reviews from each week, they reflect the impact of recent reviews. In 2011, app store reviews were sorted by chronological order. It is natural to expect recent reviews had a higher chance of being referenced and influencing consumer decisions.

In this research, we control the variables commonly used in marketing literature for eWOM. First, most economic literature supports that price affects the sales of products; a higher price will reduce sales. Since not all apps cost the same, it is also important to include variable *Price* in our model. Second, we control the valence of consumer numerical ratings *AvgRating* (i.e., the average number of stars), which reflect the reputation of the product and is often controlled in previous studies [8]. Third, we control the volume of consumer ratings/reviews. Since the number of reviews is countable data, we take a log transformation (*LogNumRev*) to make its distribution closer to normal distribution. We project the volume of ratings/reviews will have two effects in our model. 1) As a number reported in the App store interface, it directly affects consumers who view it. 2) The number of ratings reflects the popularity of apps. However, as a cumulative measure, directly using number of ratings for this purpose may combine past and current popularity. From this perspective, the change of number of ratings better represents the popularity of the app in a short period. Thus, in this research, we include both *LogNumRev* and *ΔLogNumRev* as control variables. Since a logarithmic transformation is conducted, *ΔLogNumRev* also reflects the growth rate of number of reviews. Last, considering that some apps have upgrades during the year, we included a binary variable, *Upgrade,* to capture its effect.

In the basic model, we applied a two-way fixed effect model:

$$LogRank_{i,t}=\alpha+ \beta_1\, ProdScore_{i,t-1} + \beta_2\, ServScore_{i,t-1} + \gamma_1 Price_{i,t-1}+ \gamma_2 Upgrade_{i,t-1}+ \gamma_3 AvgRating_{i,t-1}+ \gamma_4 LogNumRev_{i,t-1}+ \gamma_5 \Delta LogNumRev_{i,t-1}+ \mu_i+\eta_t+\varepsilon_{i,t}. \qquad (2)$$

where i represents an app and t represents a week. μ_i represents the product characteristics that do not vary across time. η_t represents the time-variant factors that influence the entire market. $\varepsilon_{i,t}$ is the random noise left. Since free apps and paid apps are ranked separately, their relations between sales rank and sales may be different. We conduct separate analyses for free and paid apps.

One may argue that the variables related to reviews in equation (2) may be endogenous. We conduct two types of robustness checks to address this concern. First, we employ a random effect model, adding app categories as control variables to enrich the possible omitted missing variable bias. Second, we employ the lags of these variables as instrumental variables and apply 2SLS (using Stata's xtivreg2 package) to estimate the coefficients. Such variables are likely to have serial correlations and will not be correlated with the error terms at time t, since their appearances are replaced by the variables in the time point t when a consumer read the reviews.

In order to show the effect of differentiating user comments to product quality and service quality, we build a measure of overall sentiment topics by using the lexicon built in Section 4.2. Since we do not need to differentiate product quality from service quality in this model, we only used the sentiment words and ignore the context words in doing the sentiment coding. We replicated the analysis with the developed *TotalScore* variable.

Results

Descriptive Statistics

<Table 4 approximately here>

We applied the MFSA framework as discussed in Section 4.2 on the weekly app reviews and obtained the product score and service score. Table 4 reports the descriptive statistics of the data. Our data set contains 79 paid apps and 70 free apps. The average sales rank of paid apps is 133 and the average sales rank of free apps is 199. The price of paid apps varied from 0 (for special promotions) to 15.99 USD with an average of 2.64 USD. The average rating was 4.14 stars in the 10 most recent reviews. The apps on average have about 200 reviews, with an average rating of about 3.9. The paid apps have higher textual scores than free apps. The *ProdScore* is about 8~9, the *ServScore* is about 0.2~0.8, and the *TotalScore* is about 17~20. Note that the *TotalScore* does not equal the sum of *ProdScore* and *ServScore* since it covers more information than the product and service dimensions that are restricted in our coding process. The probability for an app to have an upgrade during a week is 0.13. With our limitation that the app has to appear in more than 35 weeks, on average paid apps appeared in 42 weeks and free apps appeared in 45 weeks.

<Table 5 approximately here>

Table 5 reports the correlation coefficients among the variables. In general, there is no strong concern on the collinearity problem. The correlation between *AvgRating* and *ProdScore* is about 0.6, which is still acceptable for regression analysis and shows that the ratings of reviews to a large extent reflect product quality.

Regression Results

<Table 6 approximately here>

Table 6 shows the results on the paid and free apps using the fixed effect model, random effect model, and instrumental variable regression. (For the instrumental variable regression, after under-identification and over-identification tests, we choose Lags 1 to 3 and Lags 1 to 2 of *ProdScore, ServScore, AvgRevRating, LogNumRev*, and *ΔLogNumRev* as instrumental variables for the paid and free apps, respectively. In such a setting, the endogeneity test is significant.) The overall results are consistent across the two data sets. Reviews on both product quality and service quality generally show a significant negative correlation with the sales rank, i.e., a positive correlation with app sales. More positive comments on the product and service quality lead to higher sales. The results show the significant impact of textual reviews, particularly the two dimensions of the textual reviews on app sales. Hence, Hypotheses 1 and 2 are confirmed.

The results on our control variables also fit previous findings and our intuition. App price is positively related to app sales rank and negatively related to app sales; i.e., a higher price leads to lower sales. The appearance of new versions of apps significantly improves app sales (and reduces sales rank). The average rating of the reviews also influences the app's sales rank, where a higher rating leads to lower sales rank and higher sales. Furthermore, our results show that both the existing number of reviews and changes of number of reviews have a negative impact on sales rank of apps, which indicates the existence of interface effect (to show a total number of reviews) and the network effect caused by the current popularity of apps.

Table 6 allows us to compare paid apps and free apps[1]. A major difference we can see is the much higher R-square on the models of free apps. Note that the variables reported in this research are generally available on app stores. For free apps, a user may make quick decisions based on such available information. However for paid apps, one may pursue information from other channels, such as third party review Websites, before making a purchase. As a result, the prediction power of our model is higher on free apps than paid apps.

We also notice that review volume and valence have different effects on paid and free apps. In general, the average rating's impact is much larger on free apps, while the review volume's impact is much larger on paid apps. Our results indicate that free app users are more influenced by ratings, which may be due to the low cost of downloading, trying, and switching. However, it is not easy to switch paid apps. The large volume of reviews indicates the large user base. Knowing that many people have purchased the app will increase the users' confidence in purchasing.

On the two hypotheses we want to test, free app and paid app users are generally consistent. We also notice free app users generally have slightly higher concerns about the app service

[1] We thank the anonymous reviewer for raising this important point to discuss.

(ServScore). We believe this is because app service quality occupies the major part of the free app users' cost. In general, it is relatively easy to figure out an app's product quality after a quick trial. However, service quality takes longer to determine. If there is any service problem, free app users are usually less protected. Then, the time they spend to find fix a service quality problem becomes a major portion of their cost. Therefore, free app users tend to focus more on service quality.

<Table 7 approximately here>

To further illustrate the value of differentiating user comments to product quality and service quality, we experiment using the overall sentiment measure *TotalScore*. Model (IV) in Table 7 reported the use of *TotalScore* in a fixed effect model. The results on the random effect model and instrumental variable model (2SLS) are consistent. It is found that the *TotalScore* measure can help predict sales rank on paid apps but not on free apps. We further conducted a non-nested *J*-test between model (I) and model (IV). In a *J*-test, the predicted value of one model is included as an independent variable in another model to investigate whether it can bring further prediction power as compared with existing variables. Models (V) and (VI) in Table 7 report the results of the *J*-test. Consistent with model (IV), *TotalScore* provides extra prediction power over *ProdScore* and *ServScore* on paid apps. *ProdScore* and *ServScore* provide extra prediction power over *TotalScore* on free apps. To a certain extent, the results show the value of our proposed MFSA approach that decomposes online reviews into multiple facets for sentiment analysis.

Discussion

<Figure 2 approximately here>

To understand the implications of our findings, we visualize the data set's product quality and service quality sentiments in Figure 2. The two measures have a very small correlation (correlation coefficient=0.210 and 0.096 respectively in Table 5). The product quality sentiments (on average 9.27 and 8.12 respectively) are much stronger and more positive than those on service quality (on average 0.84 and 0.16, respectively). In other words, if we do not differentiate these two types of sentiments, the positive product quality sentiments will dominate the overall sentiments and overshadow consumers' opinions on service quality. Using an overall sentiment measure of textual reviews will cause biased prediction of app sales.

The larger scale of the product quality score shows that a majority of consumer comments are on product quality, which is the basis for the success of mobile apps. However, the service score coefficient is 3 to 4 times larger (in absolute value) than that of the product score in Table 6, suggesting that negative comments on service hurt sales more and mobile app companies definitely need to put emphasis on improving their service quality. Previously, Li and Hitt [42] argued that showing multidimensional numerical ratings could reduce the bias of ratings caused by product price, which can provide consumers with better references for purchase decisions. In this study, we further illustrate the importance of differentiating the multiple facets of textual reviews to help us better understand consumers' opinions and predict demand. This finding should be considered in future text-mining studies.

Based on our data set, consumers take the comments on both product quality and service quality of mobile apps seriously. This finding highlights the importance of service in the mobile app business, in addition to the product quality measure valued in e-commerce. To succeed in the highly competitive mobile app business, it is necessary to improve both product and service quality. We found that service quality comments have a slightly stronger impact on free app users than paid app users. The non-nested *J*-test shows that it is more valuable to differentiate product quality and service quality when predicting free app sales. Mobile app companies should carefully maintain their online reputation to succeed in the mobile app business.

Conclusion

This research explores the relationship between textual reviews and mobile app sales. We developed a multi-facet sentiment analysis method for analyzing textual sentiments from the perspective of product quality and service quality. Using a data set from the IOS app store in Taiwan, we found that after differentiating product quality and service quality, consumer textual reviews show a significant impact on app sales rankings. Our study shows that app users care about service quality in addition to product quality, especially for free apps. To succeed in the mobile app market, it is necessary to improve both product and service quality.

Our study provides support for using multi-facet sentiment analysis to understand eWOM. Previous research on understanding of textual eWOM is often conducted in an aggregative way. Existing studies failed to show the additional value of textual reviews over numerical ratings, which may be due to the cancelling of (conflicting) information in the aggregation. In this paper, through the analysis of app reviews, we show that sentiment analysis needs to get into the details of different aspects of consumer opinions. Analysis of eWOM at a finer granularity could provide new insights and may improve the theoretical explorations on e-commerce applications.

Among the different types of concerns in consumer reviews, this study focuses on product and service quality based on previous studies on IT artifacts. This study provides a textual analysis method to code eWOM and assess the two types of information from customer reviews. This method can be employed in future studies on other types of IT artifacts, especially in languages without extant lexicons or sentiment analysis models. For example, in SaaS ERP systems, users are concerned about both the service security and system functionality (user experience). In intelligent transportation solutions, users care about both the quality of (map-based) software functionality and interface and the accuracy of the provided traffic information. In such scenarios, a similar differentiation of product quality and service quality can be conducted, where our proposed approach can be applied.

In this research, we took a theory-driven approach and only explored two dimensions in app reviews. It is possible to extend the research to aspects other than product and service qualities along with the development of IS theories. In this research, we build lexicons for each of the two quality measures. It is possible to extend our framework to more detailed sub-dimensions of product and service qualities. To do so, it is necessary to develop multiple lexicons and further clarify where terms belong in each of the sub-dimensions of a construct, which will require substantial coding efforts of both IS and CS researchers. However, such fine development of lexicons may empower us with more business insights from product reviews. From a text-mining perspective, it is also possible to incorporate richer textual features from reviews depicting consumer concerns related to IT artifacts' characteristics. Exploring richer features and more advanced sentiment analysis models could lead to more effective predictions. Although it is not the intention of this paper to build text-mining models, we believe the advances in text-mining techniques may also lead to improvements from a theoretical development perspective.

For the empirical findings we derived on mobile apps, it should be noted that our data is on IOS apps in Taiwan. It is possible that the effects of eWOM on sales vary across regions. It is also possible that Android apps (another type of popular mobile apps) may show different characteristics from IOS apps. In future studies we plan to extend this research to more platforms and regions. We only collected sales rank data on the top 500 apps due to limitations of the IOS app store. It is possible the effects of eWOM differ on less popular apps, which is worth studying. In this study, we only consider product reviews of app stores. It would also be valuable to examine eWOM in other channels, such as blogs, web forums, Twitter, etc., and examine their impact on mobile app sales, which is also deferred to future research.

Acknowledgement

The work described in this paper was partial supported by CityU SRG 7004287 and SRG 7004142. All opinions are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. Amblee, N., and Bui, T. Can brand reputation improve the odds of being reviewed on-line? *International Journal of Electronic Commerce*, 12, 3 (2008), 11-28.

2. Amblee, N., and Bui, T. Hardnessing the influence of social proof in online shopping: The effect of electronic word of mouth on sales of digital microproducts. *International Journal of Electronic Commerce*, 16, 2 (2011), 91-113.

3. Baum, D., and Spann, M. The interplay between online consumer reviews and recommender systems: An experimental analysis. *International Journal of Electronic Commerce*, 19, 1 (2014), 129-161.

4. Berger, J., Sorensen, A.T., and Rasmussen., S.J. Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*, 29, 5 (2010), 815-827.

5. Bone, P.F. Word-of-mouth effects on short-term and long-term product judgments. *Journal of Business Research*, 32, 3 (1995), 213-223.

6. Chen, Y., Fay, S., and Wang, Q. Marketing implications of online consumer product reviews. *Business Week*, 7150 (2003), 1-36.

7. Chen, Y., and Xie, J. Online consumer review: Word-of-mouth as a news element of marketing communication mix. *Management Science*, 54, 3 (2008), 477-491.

8. Chevalier, J.A., and Mayzlin, D. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43, 3 (2006), 345-354.

9. Chintagunta, P.K., Gopinath, S., and Venkataraman, S. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, 29, 5 (2010), 944-957.

10. Christiane, F. *Wordnet: An electronic lexical database*. Cambrige: MIT Press, 1998.

11. Chu, W., Roh, M., and Park, K. The effect of the dispersion of review ratings on evaluations of hedonic versus utilitarian products. *International Journal of Electronic Commerce*, 19, 2 (2015), 95-125.

12. Clemons, E.K., Gao, G.G., and Hitt, L.M. When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems*, 23, 2 (2006), 149-171.

13. CNN. Facebook buys whatsapp for $19 billion. http://money.cnn.com/2014/02/19/technology/social/facebook-whatsapp/, 2014, (Assessed on May 10th 2014).

14. Cunningham, S.M. Perceived risk as a factor in the diffusion of new product information. *Science, Technology and Marketing* (1966), 698-722.

15. Dave, K., Lawrence, S., and Pennock, D.M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the International Conference on World Wide Web*: ACM, 2003, pp. 519-528.

16. Dellarocas, C. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49, 10 (2003), 1407-1424.

17. Dellarocas, C., Zhang, X.M., and Awad, N.F. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 21, 4 (2007), 23-45.

18. Delone, W.H. The delone and mclean model of information systems success: A ten-year update. *Journal of Management Information Systems*, 19, 4 (2003), 9-30.

19. DeLone, W.H., and McLean, E.R. Information systems success: The quest for the dependent variable. *Information Systems Research*, 3, 1 (1992), 60-95.

20. Demers, E., and Vega, C. Linguistic tone in earnings announcements: News or noise? , *FRB International Finance Discussion Paper*, 2011.

21. Duan, W., Gu, B., and Whinston, A.B. Do online reviews matter?—an empirical investigation of panel data. *Decision Support Systems*, 45, 4 (2008), 1007-1016.

22. FaberNove. The iphone app store's ranking algorithm. http://www.readwriteweb.com/start/2010/02/iphone-appstore-ranking-algorithm.php, 2010, (Assessed on Apr 21st 2013).

23. Fang, Y.H. Beyond the credibility of electronic word of mouth: Exploring ewom adoption on social networking sites from affective and curiosity perspectives. *International Journal of Electronic Commerce*, 18, 3 (2014), 67-101.

24. Garg, R., and Telang, R. Inferring app demand from publicly available data. *MIS quarterly*, 37, 4 (2013), 1253-1264.

25. Ghose, A., and Ipeirotis, P.G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23, 10 (2011), 1498-1512.

26. Godes, D., and Mayzlin, D. Using online conversations to study word-of-mouth communication. *Marketing Science*, 23, 4 (2004), 545-560.

27. Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., Libai, B., Sen, S., Shi, M., and Verlegh, P. The firm's management of social interactions. *Marketing Letters*, 16, 3 (2005), 415-428.

28. Grover, V., Cheon, M.J., and Teng, J.T.C. The effect of service quality and partnership on the outsourcing of information systems functions. *Journal of Management Information Systems*, 12, 4 (1996), 89-116.

29. Gu, B., Park, J., and Konana, P. The impact of external word-of-mouth sources on retailer sales of high-involvement products. *Information Systems Research*, 23, 1 (2012), 182-196.

30. Harrison-Walker, L.J. The measurement of word-of-mouth communication and an investigation of service quality and customer commitment as potential antecedents. *Journal of Service Research*, 4, 1 (2001), 60-75.

31. Hennig‐Thurau, T., Gwinner, K.P., Walsh, G., and Gremler, D.D. Electronic word‐of‐mouth via consumer‐opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing*, 18, 1 (2004), 38-52.

32. Hu, M., and Liu, B. Mining and summarizing customer reviews. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: ACM, 2004, pp. 168-177.

33. Huang, G.H., and Korfiatis, N. Trying before buying: The moderating role of online reviews in trial attitude formation toward mobile applications. *International Journal of Electronic Commerce*, 19, 4 (2015), 77-111.

34. IDC. Smartphone os market share, q1 2015. http://www.idc.com/prodserv/smartphone-os-market-share.jsp, 2015, (Assessed on Jun 30th 2015).

35. Jiang, J.J., Klein, G., and Carr, C.L. Measuring information system service quality: Servqual from the other side. *MIS quarterly*, 26, 2 (2002), 145-166.

36. Kettinger, W.J., and Lee, C.C. Zones of tolerance: Alternative scales for measuring information systems service quality. *MIS quarterly*, 29, 4 (2005), 607-623.

37. Kitchenham, B., and Pfleeger, S.L. Software quality: The elusive target. *IEEE Software*, 13, 1 (1996), 12.

38. Ku, L.-W., Huang, T.-H., and Chen, H.-H. Using morphological and syntactic structures for chinese opinion analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Volume 3*: Association for Computational Linguistics, 2009, pp. 1260-1269.

39. Lee, C.-Y., Ku, K.-W., and Chen, H.-H. A study on identification of opinion holders. *Proceedings of the Conference on Computational Linguistics and Speech Processing*, Taichung, Taiwan, 2009, pp. 101-114.

40. Lepmets, M., Cater-Steel, A., Gacenga, F., and Ras, E. Extending the it service quality measurement framework through a systematic literature review. *Journal of Service Science Research*, 4, 1 (2012), 7-47.

41. Li, M.X., Huang, L.Q., Tan, C.H., and Wei, K.K. Helpfulness of online product reviews as seen by consumers: Source and content features. *International Journal of Electronic Commerce*, 17, 4 (2013), 101-136.

42. Li, X.X., and Hitt, L.M. Price effects in online product reviews: An analytical model and empirical analysis. *MIS quarterly*, 34, 4 (2010), 809-831.

43. Limam, N., and Boutaba, R. Assessing software service quality and trustworthiness at selection time. *IEEE Transactions on Software Engineering*, 36, 4 (2010), 559-574.

44. Liu, Y. Word-of-mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70, 3 (2006), 74-89.

45. Liu, Y., Chen, Y., Lusch, R., Chen, H., Zimbra, D., and Zeng, S. User-generated content on social media: Predicting market success with online word-of-mouth. *IEEE Intelligent Systems*, 25, 1 (2010), 75-78.

46. Ma, W.-Y., and Chen, K.-J. Introduction to ckip chinese word segmentation system for the first international chinese word segmentation bakeoff. *Proceedings of the SIGHAN Workshop on Chinese Language Processing-Volume 17*: Association for Computational Linguistics, 2003, pp. 168-171.

47. Ning, K.-c., and Lai, K.-c. Sentiment analysis in scenic spot: Experiences from web community using senmantically relevant contextual information. *Proceedings of the Conference on Computational Linquistics and Speech Processing*, 2010, pp. 320-334.

48. Pang, B., and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the Annual meeting on Association for Computational Linguistics*: Association for Computational Linguistics, 2004, pp. 271.

49. Pang, B., and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the Annual Meeting on Association for Computational Linguistics*: Association for Computational Linguistics, 2005, pp. 115-124.

50. Pang, B., and Lee, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2, 1-2 (2008), 1-135.

51. Park, D.-H., and Kim, S. The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews. *Electronic Commerce Research and Applications*, 7, 4 (2008), 399-410.

52. Park, D.-H., and Lee, J. Ewom overload and its effect on consumer behavioral intention depending on consumer involvement. *Electronic Commerce Research and Applications*, 7, 4 (2009), 386-398.

53. Perez, S. Itunes app store now has 1.2 million apps, has seen 75 billion downloads to date. http://techcrunch.com/2014/06/02/itunes-app-store-now-has-1-2-million-apps-has-seen-75-billion-downloads-to-date/, 2014, (Assessed on Jun 30th 2015).

54. Petter, S., DeLone, W., and McLean, E. Measuring information systems success: Models, dimensions, measures, and interrelationships. *European Journal of Information Systems*, 17, 3 (2008), 236-263.

55. Phillips, L.W., Chang, D.R., and Buzzell, R.D. Product quality, cost position and business performance - a test of some key hypotheses. *Journal of Marketing*, 47, 2 (1983), 26-43.

56. Pitt, L.F., Parent, M., Junglas, I., Chan, A., and Spyropoulou, S. Integrating the smartphone into a sound environmental information systems strategy: Principles, practices and a research agenda. *The Journal of Strategic Information Systems*, 20, 1 (2011), 27-37.

57. Pitt, L.F., Watson, R.T., and Kavan, C.B. Service quality: A measure of information systems effectiveness. *MIS quarterly* (1995), 173-187.

58. Ravichandran, T., and Rai, A. Quality management in systems development: An organizational system perspective. *MIS quarterly*, 24, 3 (2000), 381-415.

59. Richins, M.L., and Root-Shaffer, T. The role of evolvement and opinion leadership in consumer word-of-mouth: An implicit model made explicit. *Advances in Consumer Research*, 15, 1 (1988), 32-36.

60. Shih, H.P., Lai, K.H., and Cheng, T.C.F. Informational and relational influences on electronic word of mouth: An empirical study of an online consumer discussion board. *International Journal of Electronic Commerce*, 17, 4 (2013), 137-165.

61. Sun, T., Song, T., and Huang, J. Effects of microblogging and third-party word of mouthon product sales: Empirical study based on movie box office revenue. *Proceedings of the Pacific Asia Conference on Information Systems*, 2014.

62. Turner, M., Budgen, D., and Brereton, P. Turning software into a service. *Computer*, 36, 10 (2003), 38.

63. Turney, P.D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the Annual Meeting on Association for Computational Linguistics*: Association for Computational Linguistics, 2002, pp. 417-424.

64. Wei, C.P., Chen, Y.M., Yang, C.S., and Yang, C.C. Understanding what concerns consumers: A semantic approach to product feature extraction from consumer reviews. *Information Systems and e-Business Management*, 8 (2010), 149-167.

65. Xu, R., Xia, Y., Wong, K.-F., and Li, W. Opinion annotation in on-line chinese product reviews. *Profeedings of the Language Resources and Evaluation Conference*, 2008, pp. 26-30.

66. Yang, C.C., Tang, T., Wong, Y.C., and Wei, C.P. Understanding online consumer review opinions with sentiment analysis using machine learning. *Pacific Asia Journal of the Association for Information Systems*, 2, 3 (2010), 73-89.

67. Yao, T., and Lou, D. Research on semantic orientation distinction for chinese sentiment words. *Proceedings of the 7th International Conference On Chinese Computing*, 2007, pp. 221-225.

68. Zagibalov, T., and Carroll, J. Automatic seed word selection for unsupervised sentiment classification of chinese text. *Proceedings of the International Conference on Computational Linguistics-Volume 1*: Association for Computational Linguistics, 2008, pp. 1073-1080.

69. Zagibalov, T., and Carroll, J. Unsupervised classification of sentiment and objectivity in chinese text. *Proceedings of the International Joint Conference on Natural Language Processing*, 2008, pp. 304-311.

70. Zeithaml, V.A. Service quality, profitability, and the economic worth of customers: What we know and what we need to learn. *Journal of the Academy of Marketing Science*, 28, 1 (2000), 67-85.

71. Zhang, J.Q., Craciun, G., and Shin, D. When does electronic word-of-mouth matter? A study of consumer product reviews. *Journal of Business Research*, 63, 12 (2010), 1336-1341.

72. Zhang, Z., Li, X., and Chen, Y. Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews. *ACM Transactions on Management Information Systems (TMIS)*, 3, 1 (2012), 5.

73. Zhu, F., and Zhang, X. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74, 2 (2010), 133-148.

Figures

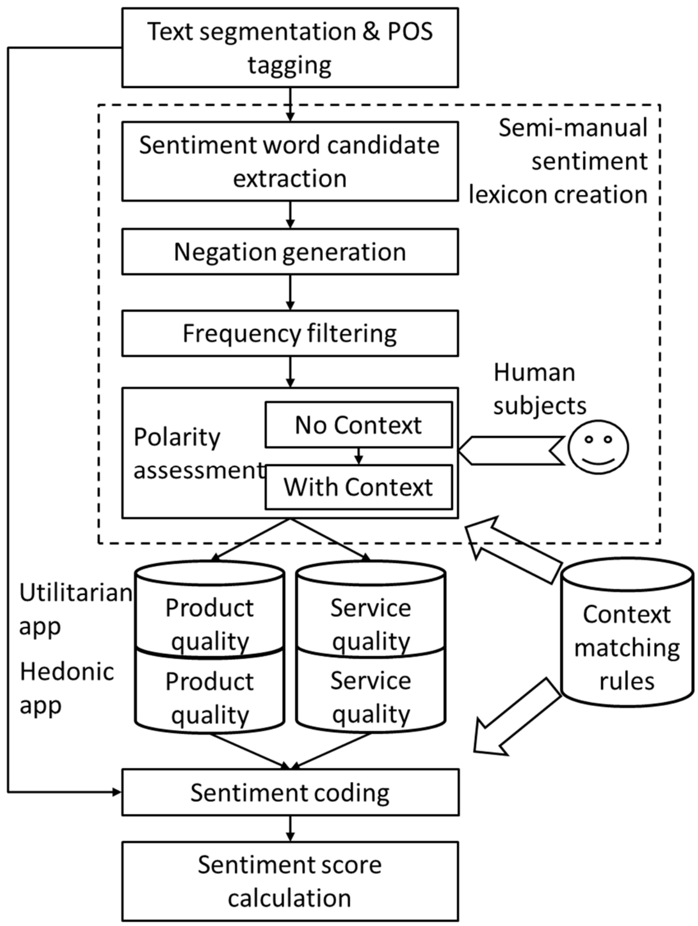
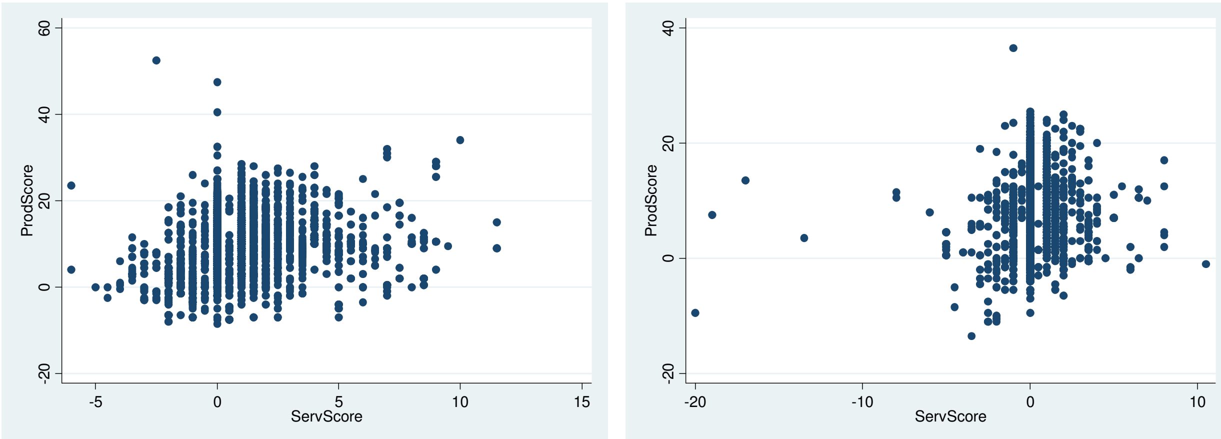


Figure 1. Procedure for Sentiment Analysis



a) Paid Apps b) Free Apps

Figure 2. Scatter Diagram of *ProdScore* and *ServScore*

Tables

Table 1. Summary of Previous Research on WOM's Impact on Sales

Paper	**Volume**	**Numerical Valence**	**Textual Valence**	**Variance**	**Others**	**Application**	**Results Related to Textual Valence**
Chevalier and Mayzlin [8]		√				Book	
Godes and Mayzlin [26]	√					TV show	
Clemons, Gao and Hitt [12]				√		Beer	
Liu [44]	√		√			Movie	Not significant
Dellarocas, Zhang and Awad [17]	√	√				Movie	
Duan, Gu and Whinston [21]	√					Movie	
Chintagunta, Gopinath and Venkataraman [9]	√	√		√		Movie	
Liu et al. [45]	√	√	√			Movie	Not significant
Zhu and Zhang [73]	√	√		√		Video game	
Zhang, Craciun and Shin [71]		√			√	Book	
Amblee and Bui [2]	√	√				Book	
Ghose and Ipeirotis [25]	√	√			√	HI-FI, Camera, DVD	Textual subjectivity is related to HI-FI sales
Gu, Park and Konana [29]	√	√			√	Camera	
Zhang, Li and Chen [72]	√	√	√	√		Book & Movie	Not significant
Sun, Song and Huang [61]	√	√	√			Movie	Not significant

Table 2. Rules for Identifying Phrases as Contexts of Words*

Tagging System	Rule	Type	Examples (English Translation)
CKIP (For Chinese)	N+N	N	應用(utility) + 軟體 (software)
	N+N+N	N	遊戲 (game)＋音樂 (music)＋音效 (sound effect)
	Vi+Vt	Vi	努力(endeavor) + 嘗試 (try)
	Vi+Vi	Vi	暢快 (smooth) + 過癮 (enjoyable)
	Vt+Vi	Vi	覺得 (feel) + 有趣 (interesting)
	N+Vi	Vi	印象 (impression) + 深刻 (vivid)
	Vi+N	Vi	好玩 (fun) + 遊戲 (game)
	Vi+ADV+Vi	Vi	覺得 (look) +太 (too) + 醜 (ugly)
	Vt+ADV+Vi	Vi	期待 (expect) + 好 (so) + 久 (long)
	N+ADV+Vi	Vi	遊戲 (game) +太 (so) + 爛 (bad)
	ADV+Vi	Vi	太 (too) + 爛 (bad)
	Vi+ T+ Na	Vi	糟糕 (terrible) + 的(-) +中文化 (Chinese localization)
Yahoo (For English)	A+N	N	Good game

* The POS tags used by CKIP are explained in http://ckipsvr.iis.sinica.edu.tw/papers/category_list.doc. N: Noun; Vi: Verb-intransitive; Vt: Verb-transitive; A: Adjective; ADV: Adverb; Na: Generic noun; T: Interjection.

Table 3. An Example of Computing Semantic Score (English translation in parentheses)

1. *太* **好玩** 了，令人**欲罷不能** (It's so interesting that I can't help myself from playing it.)
2. **Fun**
3. 好刺激的破關方式，實際玩時需要動**動腦** (What an exciting way of finishing the mission. One needs to think hard to play.)
4. *很* **棒**，比周杰倫還**棒** (Very great, greater than Jay Chou.)
5. 這遊戲感覺到*非常* **物超所值** (This game is very much worth the price.)
6. **good**
7. **中文化字體糟**透了，讓人糾結 (The Chinese font is terrible, which makes people frustrated.)
8. 娛樂性滿分，**good** (Full score in entertainment, good)
9. **Fun**，**中文化字體很醜**，不如不要 (Fun. The Chinese font is ugly, which is better to throw away.)
10. *Very* **fun**，**好玩** (Very fun, interesting)

* Bold words are sentiment terms and italic words are adverbs.

Table 4. Descriptive Statistics of the Data in the Last Week

	Paid Apps					Free Apps				
Variables	**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**	**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**
SalesRank	3321	133.00	107.10	1	500	3098	199.32	112.13	2	500
LogRank	3321	4.48	1.05	0	6.21	3098	5.09	0.71	0.69	6.21
Price	3321	2.64	2.86	0	15.99	3098	0	0	0	0
NumRev	3321	194.59	316.03	1	3249	3098	223.79	431.14	1	5585
LogNumRev	3321	4.47	1.29	0	8.09	3098	4.76	1.11	0	8.63
AvgRating	3321	3.96	0.91	1	5	3098	3.90	0.76	1.5	5
ProdScore	3321	9.27	6.90	-8.5	52.5	3098	8.12	5.94	-13.5	36.5
ServScore	3321	0.84	1.97	-6	11.5	3098	0.16	1.32	-20	10.5
TotalScore	3321	19.81	11.72	-11.5	112.5	3098	17.20	11.88	-16.5	236.5
Upgrade	3321	0.13	0.33	0	1	3098	0.13	0.33	0	1
NumWeek	3321	42.61	4.91	35	51	3098	44.79	4.73	35	52

Table 5. Correlation Matrix

		Paid Apps									Free Apps						
		V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**		**V1**	**V2**	**V3**	**V4**	**V5**	**V7**	**V8**
LogRank	**V1**	1									1						
TotalScore	**V2**	-0.053	1								0.035	1					
ProdScore	**V3**	-0.033	0.533	1							0.048	0.376	1				
ServScore	**V4**	0.011	0.427	0.210	1						0.005	-0.071	0.096	1			
Upgrade	**V5**	-0.054	0.027	0.015	0.018	1					-0.018	-0.011	-0.009	-0.014	1		
Price	**V6**	-0.016	0.110	-0.146	0.045	0.032	1										
AvgRating	**V7**	-0.057	0.190	0.623	0.19	0.027	-0.094	1			0.089	0.040	0.595	0.025	-0.003	1	
LogNumRev	**V8**	-0.213	-0.054	0.171	-0.16	0.005	0.237	0.281	1		0.161	-0.100	0.072	-0.062	-0.025	0.283	1

Table 6. The Effect of Differentiated Sentiments on App Sales (* p<0.10, ** p<0.05, *** p<0.01; p value in parenthesis)

	Paid Apps			Free Apps		
	(I) Fixed Effect	**(II) Random Effect**	**(III) IV-2SLS**	**(I) Fixed Effect**	**(II) Random Effect**	**(III) IV-2SLS**
ProdScore	-0.005*	-0.005*	-0.004	-0.005**	-0.005**	-0.007**
	(0.056)	(0.068)	(0.271)	(0.014)	(0.014)	(0.033)
ServScore	-0.014*	-0.014*	-0.025**	-0.021***	-0.020***	-0.019**
	(0.061)	(0.068)	(0.026)	(0.001)	(0.001)	(0.027)
Price	0.168***	0.146***	0.130***			
	(0.000)	(0.000)	(0.000)			
Upgrade	-0.066*	-0.065*	-0.033	-0.047**	-0.046*	-0.051**
	(0.064)	(0.070)	(0.404)	(0.044)	(0.051)	(0.028)
AvgRating	-0.211***	-0.163***	-0.138	-0.308***	-0.256***	-0.111*
	(0.001)	(0.006)	(0.137)	(0.000)	(0.000)	(0.052)
LogNumRev	-0.125***	-0.163***	-0.191***	-0.060**	-0.058**	-0.137***
	(0.000)	(0.000)	(0.000)	(0.030)	(0.026)	(0.000)
ΔLogNumRev	-0.802***	-0.831***	-2.125***	-0.669***	-0.674***	-1.770***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Time dummy	√	√	√	√	√	√
App id dummy	√		√	√		√
App category dummy		√			√	
Num Apps	79	79	79	70	70	70
Num Obs	3113	3113	2610	2973	2973	2767
R-square	0.1862	0.1854	0.1505	0.3552	0.3548	0.3158
Endogeneity test			0.0271**			0.0038***

Table 7. J-test: Total Sentiment vs. Separated Sentiment in a Fixed Effect Model (* p<0.10, ** p<0.05, *** p<0.01; p value in parenthesis)

	Paid Apps			Free Apps		
	IV	**V**	**VI**	**IV**	**V**	**VI**
TotalScore	-0.006***		-0.005***	0.000		0.001
	(0.000)		(0.002)	(0.578)		(0.203)
ProdScore		0.001			-0.006***	
		(0.661)			(0.005)	
ServScore		-0.007			-0.020***	
		(0.389)			(0.002)	
Price	0.166***	-0.003	0.138*			
	(0.000)	(0.961)	(0.069)			
Upgrade	-0.064*	0.000	-0.053	-0.046*	0.077	0.002
	(0.071)	(0.994)	(0.239)	(0.050)	(0.406)	(0.931)
AvgRating	-0.213***	-0.002	-0.175	-0.334***	0.598	0.011
	(0.001)	(0.983)	(0.148)	(0.000)	(0.364)	(0.898)
LogNumRev	-0.133***	0.001	-0.111*	-0.059**	0.102	0.005
	(0.000)	(0.989)	(0.087)	(0.034)	(0.399)	(0.873)
ΔLogNumRev	-0.805***	0.009	-0.671*	-0.674***	1.161	0.037
	(0.000)	(0.973)	(0.065)	(0.000)	(0.382)	(0.833)
Predicted Value from Model (IV)		1.016***			2.711	
		(0.001)			(0.168)	
Predicted Value from Model (I)			0.166			1.049***
			(0.707)			(0.000)
Time dummy	√	√	√	√	√	√
App id dummy	√	√	√	√	√	√
Num Apps	79	79	79	70	70	70
Num Obs	3113	3113	3113	2973	2973	2973
R-square	0.1887	0.1889	0.1887	0.3511	0.3557	0.3556

Appendix:

Table A1. The Context-independent Sentiment Lexicon for Utilitarian App Product Quality

Positive Terms	Negative Terms	Agreements
好用; 方便;易用; 美; convenient	難用; 當機; 不方便; 醜; 難看; 失焦; 不細緻	10
穩定; fun; 漂亮; useful; 好玩; popular; 愛不釋手; 別出心裁; 美麗; 標緻	不便; lag; 當掉; 過時; 模糊; vague); 不穩; 失靈; 不清晰; 曝光; 不清楚; 昏黃	9
簡單; 推薦; 便利; 專業; 可愛; 順暢; 有趣; 豐富; awesome; helpful; Beautiful; 流行; 清楚; 有創意的; 逼真; 感動; 絢麗; 五光十色; fashionable	不好用; ugly; 無聊; 灰暗; 不專業; 過氣	8
Good; 讚; 實用; 好看; Nice; great; 清晰; 齊全 ; perfect; 精彩; 美輪美奐; 美味; 厲害; 顯眼; 特立獨行; 標新立異	Bad; 膩; poor; 不真實; 錯置; 走調	7

Table A2. The Context-independent Sentiment Lexicon for Utilitarian App Service Quality

Positive Terms	Negative Terms	Agreements
值得; 便宜	浪費; 後悔; 貴	10
買; 敗; 物超所值; 不後悔	騙人; 不值得; 欺騙; 重載;還錢; expensive;後悔莫及	9
超值; Worth; 有誠意	重灌; 投訴; 沒誠意	8
	商業化	7

Table A3. The Context-dependent Sentiment Lexicon for Utilitarian App Product Quality

Positive Terms: Context	**Negative Terms:** Context	Agreements
好: 觸控, 光影, 畫質, 特效, 顏色, 效果, 操作, 工具	**爛:** 畫質, 特效, 顏色, 打光, 效果, 操作, 對焦, 介面, 攝影, 傳輸, 存取, 感光, 線, 設計, 光線, 傳輸, 平台	10
強: 連線, 特效, 顏色, 效果, 感光, 功能, 色調	**糟糕:** 設計, 介面, 畫質, 特效, 顏色, 速度, 連線	
佳: 感光, 鏡頭, 畫素, 色彩, 特效, 質感, 影像, 色調, 操作, 傳輸, 存取, 層次	**慢:** 連線, 傳輸, 對焦, 攝影, 拍照, 感光, 調整, 套用, 調節, 同步, 傳送	
及時: 簡訊, 傳輸, 連絡, 通訊	**弱:** 實用性, 娛樂性, 層次, 圖庫, 趣味, 感覺, 顏色, 效果, 操作	
深: 全景	**失敗:** 顏色, 傳輸, 讀取	
	笨拙: 觸控	
貼心: 功能, 介面, 設計	**差:** 實用性, 感覺, 亮度, 圖層, 顏色, 效率, 觸控, 介面, 連線, 效能	9

體貼: 功能, 設計	**錯誤:** 訊息, 傳輸, 讀取, 對話, 通知	
	小: 瑕疵, 問題	
	難: 拍照, 使用	
	可怕: 效果	
有: 智慧, 實用性, 娛樂性, 笑點, 層次, 圖庫, 趣味, 感覺, 設計	**不會:** 使用, 拍照, 同步	8
高: 實用性, 實用價值, 畫素, 畫質, 亮度, 解析度, 彩度, 質感, 享受	**失靈:** 觸控, 鍵盤, 讀取, 特效	
快: 傳輸, 存取, 照相, 感光, 影像, 對焦, 操作, 連線	**惡劣:** 印象	
完美: 鏡頭, 畫素, 色彩, 圖片, 特效, 質感, 生活, 影像, 色調, 工具, 經典	**不爽:** 心情	
打發: 時間		
深刻: 印象, 記憶		
消磨: 時間, 時光	**高:** 難度, 風險	7
喜歡: 介面, 色調, 畫質, 顏色, 效果	**異常:** 功能, 傳輸, 訊息, 對話, 讀取, 連結, 使用, 通知	
多: 功能, 特效, 效果, 層次, 朋友, 表情符號, 回憶, 功效	**沒有:** 感覺, 印象, 傳輸	
消除: 無聊	**有:** 問題	

Table A4. The Context-dependent Sentiment Lexicon for Utilitarian App Service Quality

Positive Terms: Context	**Negative Terms:** Context	Agreements
快: 更新, 客服	**多:** 廣告, 訊息, 垃圾訊息	10
好: 服務, 客服, 價格	**慢:** 更新, 退款, 流程, 寄送	
快速: 更新	**辛酸:** 更新	
週全: 服務		
親切: 客服		
及時: 反應, 更新	**爛:** 更新, 服務	9
	失敗: 更新, 安裝, 升級	
	高: 收費, 價格	
	煩: 廣告, 訊息	
	不滿意: 中文化, 更新, 新版本, 客服	
	不需要: 中文化, 新版	8
	差: 中文化, 更新	
	不會: 更新, 回信	7
	不要: 中文化, 新版	
	奇怪: 字體, 中文字	

Table A5. The Context-independent Sentiment Lexicon for Hedonic App Product Quality

Positive Terms	Negative Terms	Agreements

好玩；簡單；順暢；Great；簡易；Simple；	不好玩；難玩；不便；不便；不穩；bad；差劲的	10
耐玩；有趣；fun；逗趣；exciting；易用；精美；刺激；好聽；享受；容易 easy	膩；無聊；不方便；無趣；Bored；醜；憤怒	9
Cool；挺；便利；不膩；欲罷不能；有創意的；輕鬆；愛不釋手；amazing；豐富；酷；悅耳；popular	當機；lag；難聽；不順；當掉	8
讚；good；棒；可愛；爽；funny；nice；上癮；熱血；成功；活潑；awesome；愛不釋手；brilliant；熱鬧；special；perfect；精彩；炫；幸福 happiness；感動；流行；beautiful	生氣；難解；頭暈；過時；無言；不爽	7

Table A6. The Context-independent Sentiment Lexicon for Hedonic App Service Quality

Positive Terms	Negative Terms	Agreements
值得；買；便宜；worth	後悔；騙人；重灌	10
下載；物超所值；敗；超值；不後悔	浪費；不值得；欺騙；貴；詐欺；投訴	9
有誠意；worthy；不悔	假；重載；欺瞞；沒誠意	8
	Expensive	7

Table A7. The Context-dependent Sentiment Lexicon for Hedonic App Product Quality

Positive Terms: Context	**Negative Terms:** Context	Agreements
優秀: 畫面, 音樂, 聲音, 音效, 劇情, 設計	**爛:** 特效, 顏色, 效果, 操作, 關卡, 介面, 攝影, 傳輸, 存取, 感光, 介面, 設計 design)	10
高: 娛樂性, 樂趣, 娛樂價值, 質感, 享受, 層次	**糟糕:** 設計, 畫面, 效果, 特效, 劇本, 速度, 連線	
強: 遊戲, 排名, 聲光, 效果, 設計, 劇本, 情節, 音效	**錯誤:** 實用性, 娛樂性, 層次, 圖庫, 趣味, 感覺, 顏色, 效果, 操作	
消磨: 時間, 時光	**糟:** 畫面, 流暢度	
爽快: 遊戲, 節奏, 破關, 戰鬥	**困難:** 移動, 滑動, 瞄準	
	沒有: 文字, 音樂, 音效	
有: 娛樂性, 喜感, 經典, 趣味, 感覺, 設計, 笑點, 創新, 意思	**差:** 遊戲性 gameplay, 愉樂性, 效果	9
殺: 時間	**弱:** 畫面, 設計, 關卡, 顏色, 觸控, 介面, 特效	
卓越: 遊戲, 經典, 質感, 配音	**慢:** 觸控, 發射	
多: 關卡, 關, 娛樂性, 角色, 玩法, 朋友, 獎勵	**異常:** 難度, 風險, 訊號	
深刻: 印象		
消除: 無聊, 慾望		
打發: 時間	**退:** 流行	8
完美: 遊戲, 破關, 闖關, 通過, 色彩, 圖片, 局勢, 特效, 質感, 影音, 餘興, 詮釋	**失敗:** 訊息, 傳輸, 讀取, 對話, 通知, 修復, 影片, 照片	
	不佳: 瑕疵, 問題	

	假： 觸控, 遊戲, 特效	
	不悅： 心情	
好： 遊戲控制, 觸控, 關卡, 音樂, 背景, 特效, 經典, 效果, 操作, 心情, 夥伴, 音效, 印象		7
動： 腦		
喜歡： 主角, 遊戲, 玩, 寶物, 音樂, 畫面, 效果, 破關, 音效, 對話, 水果, 鳥, 豬, 蛋, 鳥蛋, 花, 彈弓, 植物, 殭屍, 星星, 寶箱, 闖關, 過關, 聲音, 水, 道具, 鞋子		

Table A8. The Context-dependent Sentiment Lexicon for Hedonic App Service Quality

Positive Terms: Context	**Negative Terms:** Context	Agreements
迅速： 儲值, 更新	**爛：** 更新, 服務	10
及時： 公司反應, 更新	**錯誤：** 更新, 退款	
好： 服務, 價格	**慢：** 更新, 流程, 儲值, 客服	
快速： 更新	**差：** 更新, 服務	
多： 抽獎, 優惠	**高：** 收費, 價格, 價錢, 花費, 訂價	
	不滿意： 中文化, 更新, 新版, 客服	
	遜： 服務	
	差勁： 更新, 新版	
	嚇人： 價格	
快： 更新, 客服, 公司處理, 排除錯誤	**多：** 廣告, 垃圾訊息	9
	失敗： 更新, 安裝, 升級	
	不會： 更新, 退錢, 退款	
	醜： 中文化, 更新	8
	煩： 更新, 廣告, 垃圾訊息	
	糟糕： 中文化, 中文字, 更新, 新版	7