Using Text Mining of Amazon Reviews to Explore User-Defined Product Highlights and Issues

L. Jack and Y.D. Tsai
Intel Corporation, Santa Clara, CA, USA

Abstract- *Advances in technology have made user-generated content ubiquitous. This includes user reviews of products which are publicly available on the internet and has led to an increase in the use of text mining to analyze consumer behavior. This paper presents a framework for using text mining to gather customer feedback. Text mining techniques are used to aggregate the top attributes associated with groups of devices, laptops and tablets, as well as individual devices. A case study comparison of three devices compares and contrasts positive and negative aspects mentioned by the users, which is useful to improve future generations of products. Manufacturers can incorporate and review product attributes when a product is launched and over time correct product issues, understand customer requirements, and maintain customer satisfaction.*

Keywords: text mining, customer, reviews, R

1. Introduction

User reviews on E-commerce websites like Amazon.com have a large influence on product reputation as they are heavily viewed by prospective buyers before they decide to make purchases. Text mining tools and algorithms can help uncover customer attitudes and sentiments on products they have purchased and used. This paper reviews a method of applying text mining techniques to compare and highlight top customer opinions of a product (in this case, laptops and tablets) as a means to provide feedback to enhance future products. Understanding the overall positive and negative perceptions of a product enables manufacturers to be in tune with the reception of their products. It also enables them to identify, fix, and resolve issues uncovered in user reviews.

Flanagin and colleagues [1] found that product ratings are used as a barometer of product quality, where higher perceived quality is associated with greater purchasing intentions. User ratings are considered a credible source of information about products consumers are intending to buy though users may only attend to average product ratings when making purchasing decisions. When a product reached a certain level (4.4 stars), a ceiling effect was found and ratings above the level did not result in perceptions of enhanced product quality. Potential buyers may only look at top level information to make purchasing judgments, however, the review text gives insight as to what contributes to their overall rating.

Several studies on review helpfulness suggest that extreme reviews are the most helpful. Chen and Tseng [2] found that high-quality reviews are those that subjectively comment on several product features. There is greater ambiguity in positive product assessments than in negative product assessments when comparing extreme reviews (4 and 5-star reviews vs. 1 and 2-star reviews). Mudambi, Schuff, and Zhang [3] call attention to the text of the review to get accurate details on the user's view of product quality. They compared the rating differences between feature-based goods and experience goods. Users interested in buying feature-based goods, such as music players, prefer reviews that outline pros and cons of the product and contain mainly objective inform mation with only few subjective statements on the product. In contrast, users rely on personalized, sentimental reviews, not captured by the product description, for experience goods like a movie DVD [5]. Review characteristics such as subjectivity and readability were also key feature categories that determine review helpfulness to the potential buyer [6]. These studies emphasize the need to not only pay attention to the summary characteristics of the product reviews, but also the detailed nuances of a user's likes and dislikes of the product.

2. Research Questions

There are many possible ways to collect feedback from users about products they have purchased and used. One typical way to do so is to ask users to complete surveys. Another is to do experimental research or an observational study of users interacting with devices. In this study, text analyses were completed using Amazon review data. Amazon reviews are considered a good source of data for capturing consumer perceptions primarily because of the large number of data points. In addition, customers are able to post their reviews after they have used the product and know its pros and cons, and while they are in a low-to-no-pressure setting, usually sharing their thoughts and feelings at home versus being in a timed lab setting.

These research and analyses were conducted as part of a larger project to identify, understand, and evaluate the basics of system performance, specifically in tablets and laptops. Essentially, the research question was, what basic features are most important to users of the product and most influential in molding their perceptions? Similar to collecting feedback, there are several ways one could go about assessing "importance." Some possible ways are to set a threshold for how many people share the same perception or what percentage have similar views. The determinant of importance in this research was frequency; the topics that came up the most were prioritized and used as a standard to assess how important all other opinions of a product were when considering the group of users as a whole.

The end goal of the project was to have assessments of both individual products (e.g. a specific laptop) and groups of products (e.g. laptops in general) that provided information about what features were important to excel at or improve upon to create a better customer experience with the product. Lastly, these assessments would become recommendations to the manufacturers for future product development and improvement. The focus of this paper is to detail the methodology and results of said investigation of Amazon reviews.

3. Methods

This analysis of Amazon product reviews focused on evaluating a variety of user reviews of tablet and laptops. "Two-in-one" devices were not included because they straddle between the two device types and could make category comparisons more difficult. To cover an assortment of products, 40 devices were chosen; 20 laptops and 20 tablets. The devices were chosen to create a diverse set of products, to ensure variety in operating systems (OS), price, brand, popularity, and to be representative of what was available and purchased by consumers on the (Amazon) marketplace.

Amazon review data was web scraped and text mined using R, a statistical software. R is an open-source programming language commonly used for statistical computing. R has both data mining (web scraping) and data analyses (statistical and text analysis) capabilities and the analyses are scripted, customizable, and repeatable. An R script was developed in this research to pull Amazon reviews of the devices of interest. In total, the number of reviews collected across all devices was 19,080.

The number of reviews per device ranged from 50 - 4,100. The prices ranged from $46.99 - $2,249.99. Table 1 (a) lists the price and number of reviews associated to the laptops; Table 1 (b) lists the tablets used in the study. Thirty-seven out of the 40 products had over 100 reviews. Using calculations based on power analysis, 100 reviews was estimated to be a large enough sample size to begin finding significant relationships [7] [8]. The three products with less than 100 reviews were chosen because of their low ratings. This was a potential confound as consumers do not often choose to buy poorly rated devices. Therefore, less people purchased these devices and they had less reviews,

TABLE I (a)

LAPTOP PRICES AND NUMBER OF REVIEWS

Laptop ID	Product Price	Number of Reviews
A	249.99	230
B	199.99	2,470
C	249	280
P	1779	130
D	247	550
V	265	100
F	299	720
U	736.59	100
G	283.38	390
H	387.99	370
J	1139.99	240
K	1229	150
S	1299	90
T	2249.99	50
Q	1099.99	100
E	378	110
R	588.96	50
L	439	220
M	294	140
N	159.94	110

TABLE I (b)

TABLET PRICES AND NUMBER OF REVIEWS

Tablet ID	Product Price	Number of Reviews
UU	92.99	140
ZZ	47.99	570
EE	169	100
FF	49.95	270
GG	149.99	550
HH	51.99	640
JJ	89.5	100
LL	353	2140
XX	559.99	1770
WW	309	260
MM	379	160
NN	306.83	170
PP	374.95	250
QQ	79.95	260
RR	198.99	4100
YY	299	190
SS	46.99	120
TT	299	380
VV	124	170
KK	59.95	140

which results in a skew of the total product ratings toward higher scores. Also, the character counts of lowly rated reviews of 1 and 2 was significantly higher (Median = 251.5, SD = 699.16) than those of highly rated reviews (Median = 163 SD = 786.07), $p < 0.05$. All of these reviews were included in the analyses, however, as they aid in understanding what elements of the product caused the users to rate the device well or poorly.

4. Analyses

The research was primarily focused on understanding what was really important to users, what positively or negatively affected product reviews, and what specifically users choose as highlights or pain points when reviewing laptop and tablets. This then, in terms of ratings, translated into which reviews were the best and the worst. The analyses focused on subsets of the data: the lowest ratings (ratings of 1 and 2) and the highest ratings (ratings of 5). Ratings of 1s and 2s were grouped together because the positive ratings outweighed the negative ones. As mentioned previously, products that are highly rated (and therefore have a good reputation), have more positive reviews than negative ones as potential buyers are unlikely to purchase a product that many people rated badly and has a poor reputation. Therefore, in order to get a large enough sample of lowly rated reviews, both 1 and 2 rated products were combined into a single category. In the end, there were 11,730 5-rated reviews and 1,678 1- or 2-rated reviews (Figure 1).

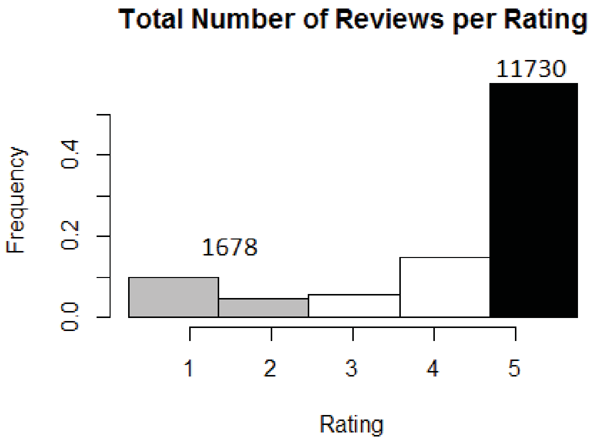


Fig 1. Histogram of the total number of reviews by rating

After collecting all the data, common text mining techniques [4] [5] were performed to prepare the data for analysis. Another R script was developed to analyze and understand patterns of discussion across groups of reviews. The scripts developed for the following analyses used the R libraries plyr, tm, RStem, stringr, ggplot2, and xml. The reviews were treated as documents and were aggregated into larger corpuses depending on what subset of data was being analyzed - all reviews, laptops, tablets, reviews from a single device, and review groupings that mentioned common key terms (for example, “battery life”).

Regular expressions were used to break the data into desired groups based on a word or words with similar meanings. Examples include words about the screen, touch, and resolution, and adjectives that describe the goodness of a product and its features - great, good, excellent, amazing, etc. Some of the words and general product features that were examined included battery life, display, touchscreen, touchpad, keyboard, and price. These product features were defined from previous work on the overall project which aggregated common attributes mentioned from technical press reviews. The key terms were also used to parse out specific sentences for further analyses.

After being turned into a corpus, punctuation and extra symbols were removed. Words were converted to lowercase, and stopwords were removed. Plots were created to ascertain correlations between numeric variables and significance testing was conducted on the proportions of times words and phrases appeared in different groups to see if there was a real difference in how many times one group mentioned a topic compared to another. Lastly, n grams were performed, which is a text mining technique for assessing the frequency of words and phrases (with n being the length of the phrase) in a corpus. Lastly, the three devices with the most reviews are presented in this paper as a case study for individual analysis.

5. Results

5.1 Overall user feelings

Assessing all 19,080 reviews across the 40 devices in aggregate, the most common aspects that were addressed overall by users were the: 1) battery life, 2) (touch) screen, 3) value/price and 4) generally (positive) feelings about the device. In reviews rated 1 or 2, reviewers used negative adjectives like bad, poor, worse, worst, and horrible most often when they were describing the quality of the product or the customer service provided. In reviews rated as 5 stars, reviewers used the positive adjectives “nice”, “best”, “good”, “great”, “perfect”, “excellent”, and “amazing” most often when they were describing the battery life, screen, and keyboard. These positive adjectives were also often associated with describing the sound (audio quality), value, and physical traits of the device, but to a lesser extent. Positive emotive words such as like, love, and happy were most often used by reviewers to describe a general love for the device and to lesser extent, to describe the screen and backlit keyboard. In an interesting contrast, the words “like” or “love” were some of the most common words whereas the word “hate” rarely if ever appeared, even in the reviews of poorly rated devices.

5.2 Frequent words and phrases

When analyzing the reviews, some words and phrases were more popular and mentioned more frequently than others. These words and phrases will be addressed in general and a few have been selected to be addressed in detail to provide more context of the analyses conducted.

Reviewers often mentioned the purpose of buying the product and wrote about key features that were surprising when encountered. Only a small percentage of reviewers (3.2%, 609 reviewers) mentioned that a product was “bought for” someone else, not for themselves. For example, this could be a parent purchasing the device for their children. Thus, the majority of reviewers have personally used their devices and are capable of talking about its specific features, their likes, and dislikes. In addition, users revealed that their “time” is valuable to them by negatively reviewing the length of time it took to accomplish tasks for setup or troubleshooting a new device. Some key words also highlighted specific features that were important to customers. For example, SD card slots were an unexpected feature that customers seemed to appreciate and miss when not present. The power button differentially appeared in low versus high ratings for two reasons: 1) In low ratings, the power button received negative reviews for being broken, and 2) In high ratings, reviewers mostly disliked the placement of power and volume buttons. Lastly, one operating system (OS) present in several devices received many mentions for varied reasons. For example, reviewers voiced frustration with the difficulty of interacting with their device due to the software, the incompatibility of the software version loaded on their device with another similarsoftware version, and the limited content of the app store. On a positive note, for those who mentioned this OS, the ability to multi-task on tablet devices was also mentioned in 10.3% of the user reviews as an appreciated feature. Example text from two reviewers, one that gave a low rating and one that gave a high rating, is shown with some of the most common concepts bolded in Figures 2 &3.

1 star rating

> *“I sent two of these back ...one decided to work off and on, and the other ones* **screen cracked** *without any impact involved. I don’t know if anyone else noticed, but there is a paper that comes in the box that basically says call the original seller, not Amazon. Fortunately I did not see it. I am only dealing with Amazon. Now, I am going to go ahead and* **spend the extra money** *to get [another device]... Do I want to* **pay that much** *for these, no, but I know what [it] does, and how much more sturdy they are. If I had one way to describe these ..., it would be flimsy, with poor technology. By the way, they take pictures, and they are poor quality. The* **battery life is too short***... I guess I have learned, once again,* **you get what you pay for***, with the exception of items you know are good, and only get a good sale.”*

Fig. 2 Example low rating review

4 star rating

> *“I bought this device mainly because the* **retail price was very low** *compared to competitors and the device looked attractive and had a* **nice feel***... if you don't always have an internet connection and use spreadsheets a lot then this device and OS is not for you. ..... that I saw was missing here was the speech input.... The build quality of this machine is one of the* **best I've seen at this price point***! I find it* **visually appealing** *.... The soft touch material being used for the outer body of this laptop* **feels nice to the hands***... This device is also* **very thin***. The device also has a nice selection of ports providing USB 2.0, USB 3.0, full size HDMI,* **SD card reader***, headphone/microphone combo jack, lock slot and an indicator light for sleep/use.:... but for $300 I guess you can't expect to get a* **1080p IPS display***...Some* **back-lit keys would be a nice** *addition in future models ... the trackpad had a silky smooth texture to it ...I* **love** *the addition of trackpad gestures ...So far from my testing the performance has been great.... I was able to get 4 days of use out of it before having to plug it in and combined use time was 7 hours. 1.75 hours of this was watching videos so* **battery life** *could have been better if just used for web browsing and document editing... I'm confident 9.5 hours is an accurate claim.... this machine should be able to* **last close to 6 hours** *for viewing back-to-back movies ....***Charging time** *from 5% took exactly 2 hours to fully charge back to 100% which is great...Overall I am pleased with the device but* **$300 might still be a little high** *for something that only serves as an internet browsing machine... It simply doesn’t outperform ... has more offline functionality on-the-go and the battery life of this machine comes up a little short....should allow add-ons or upgrades on their website for a* **backlit keyboard***...”*

Fig. 3 Example high rating review

5.3 Battery life

Battery life was mentioned most often across all devices, in 2,665 1-5 star reviews, across device types (laptop, tablet) and individual devices. Additionally, reviewers mentioned battery life statistically more often when they were giving high ratings than low ratings (162 out of 1,678 1-2 star reviews, 9.6%; 1,539 out of 11,730 5 star reviews, 13%, $p = 0.0$). The reviews mentioning battery life were mined to further understand the context of why battery life was so prominent and to analyze battery life expectations in hours. For laptops, reviewers considered four hours of battery life with moderate use satisfactory, around five hours to be good or standard, and above eight hours to be noteworthy. A further exploration was conducted of highly rated reviews that explicitly mentioned both “hours” and “battery [life]”. Seven percent of those reviewers said that the battery lasted around five hours, and gave the product a rating of 5 (the other 93% did not have a unified voice about battery life duration). As for tablets usage, many users spoke of watching TV and video streaming from their device. Reviewers assumed that tablets had a longer battery life compared to laptops. Over seven hours of battery life was considered satisfactory for a tablet, and 10 or more hours was considered excellent. In addition, how long the device took to charge seemed just as important as how long the charge lasted; this was especially true for tablets.

5.4 Screens

For reviewers who gave low ratings, tablet (touch) screens were most often mentioned in reviews, 21% of the time. Some reasons users cited were responsiveness issues with the touchscreen or the screen being cracked or broken (sometimes found upon unboxing the device). Example responsiveness issues include slow or lagging performance and inaccurate registration of touched screen locations. For laptop screens, some were said to have resolution mismatch issues, which meant that the capability of the device to display at a higher resolution made viewing lower resolution content that was not able to scale and adapt to a higher resolution unfavorable. In these cases, the product feature detracted from the overall experience of the product.

5.5 Case study comparison between individual devices

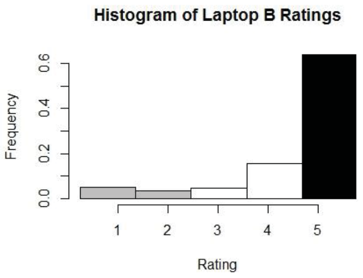
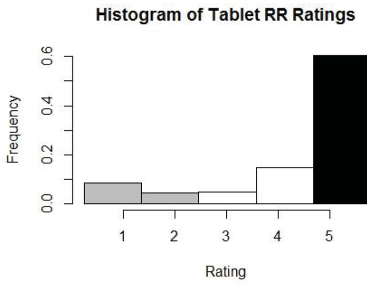
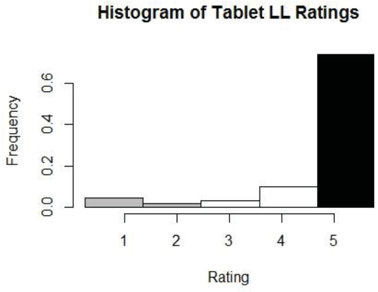
This analysis delved into the three devices with the most reviews: Laptop B, Tablet RR, and Tablet LL (Table 2). All product names and some specifications have been changed to protect brand privacy. For each of the three devices, analyses were done to find the positive aspects and negative aspects of the device. For example, one of the negative aspects mentioned about Tablet LL was a general warning from past customers to not buy the product from a well-known website. Delving more deeply into the reviews, surfaced the reason: customers experienced differences in device quality and customer returns when interacting with the original supplier of the device compared to a third party vendor; discouraging reviews were posted as a result. The comparison also highlighted interesting trends such as, people use “love” twice as often in reviews of Tablet LL than Tablet RR.

In addition to the results listed in the table, some data mining was done to ascertain user defined problems with the devices. Specifically, reviews of rating 1 or 2 that mentioned the words “issue” or “problem” were scrutinized. The results from this exercise were then compared to published articles (e.g. technical press articles) on device issues for verification of issues data mined from the reviews. When Tablet RR, for example, was compared to the top web search results for issues pertaining to that device, all items that arose from data mining (battery charging, freezes, random reboots, touchscreen responsiveness, and Wi-Fi connectivity) were addressed and verified as indeed being widespread problems.

6. Discussion

There are some key takeaways from this research. First, if one can create a feeling of love for a product, that could improve product ratings. The Amazon review data showed that people use “love” very often when describing their product interaction but rarely use “hate” when talking about the product, even for devices that received a poor rating. In addition, love was used to describe a general feeling toward

TABLE II
EXAMINING THE THREE DEVICES WITH THE MOST REVIEWS

Device ID, number of reviews	**Laptop B** 2470	**Tablet RR** 4100	**Tablet LL** 2140
Distribution			
Popular topics in reviews rated 5	•Specific preloaded software and operating system •Everything they need •Battery life •Love it •Fast	•Comparing it to a similar tablet •App availability •Screen •Price •Fast	•Would recommend it •App availability •Easy to use •Love it •Price
Popular topics in reviews rated 1 & 2	•Reboots and deaths •Keyboard and trackpad •Operating system and its applications •Printer & Wi-Fi connectivity	•Touchscreen •Battery life •Customer service	•Do not buy it from a specific dealer •Wi-Fi Connectivity

the device, not any particular aspects of it, but the device holistically. It is possible that customers are very susceptible to product branding, which indicates that product marketers stand to benefit from understanding why people love a product. Other studies have examined this in detail and have cited physical attributes (e.g. the object is beautiful, ultimate) and significance in personal value have been correlated with users having engrossing or transcending experiences with the object that leads to love [10]. The association of love for the product may be a key barometer of its success.

Second, users definitely take note of battery life. They have expectations around how good the battery needs to be and the data shows it is the most talked about topic in both positive and negative reviews. For tablets, users expected battery life to last over seven hours, for laptops, over 5 hours. Improving battery life can improve product ratings of average products to stand apart from other similar devices, but not when there are other glaring issues with the device. Thus, battery life is a feature that can be improved to get a better rating, but is not the only contributing feature for devices that have high ratings. Battery issues should not be overlooked as they directly affect how long users can use their devices.

Last but not least, manufactures need to make sure all devices are functional before shipping them out to customers. This includes making the touchscreen responsive and out of a good material. It is apparent that the level of quality control or functionality in some products needs to be addressed as many reviewers cited their devices breaking within the first few uses, or worse, already being broken upon arrival. Too many reviewers complained that their devices simply did not meet basic expectations, exhibiting issues such as being unable to turn the device on, connect to Wi-Fi, have working trackpads, have functioning power buttons, etc. Having a reliably functioning device or a means to quickly address these problems for the customer may help alleviate aggrieved customers with device issues. This should be an aim for all manufacturers.

6.1 Limitations & Next Steps

One limitation of this research is that the text mining was only conducted on Amazon reviews and it is possible that Amazon attracts a unique group of customers. By comparing the results of this research to a few technical review articles written by experts, the validity of the results was confirmed for one of the devices. However, there may still be some misjudgments across other devices. Future iterations of this research could include reviews from other databases and review websites. An advantage of this text mining methodology is that it could be repeatedly conducted to gain insight about how customers see products and how opinions may change over time with improvements to technology and to the products.

One major question that resurged repeatedly during this research was, “where does one draw the line?” When conducting surveys and questionnaires, or asking for opinions, deciding the point at which to start considering a person’s opinion or problem as an overall issue or problem is quite difficult. For example, when deciding to take action on customer feedback, does one consider: What are their top 10 concerns and highlights? Is the same issue brought up by 10% or more of the customers? Do certain highlights or concerns appear more often than would be “expected”? How does one define how often to expect a word, phrase, or concept to appear? So far, the literature in this area is either lacking or difficult to find. In this research, the decision was made to delve more deeply into the top (defined as the most frequent) concerns and opinions of the reviewers. However, other ways of assessing where the line is are equally valid.

In future iterations of this type of research, deeper machine learning type techniques could be applied to the conduct a predictive analysis. For example, one could attempt to predict what a product rating would be solely based on a review. That exercise in and of itself would help tease out what is important to customers and why they may rate a product in a particular way.

7. Conclusion

This paper outlines a method to apply text mining to understand consumer feedback about purchased products. Any person or business can use this framework to quickly gain insights about what customers in their field are saying about their products and customize the methodology to fit their needs. Information about customer preferences, key features, and encountered issues can then be used to improve upon the product. When this method is applied to review one product, the top features that are important to the user can be gleaned, as well as the main problem areas of the product. When applied to review a group of products, comparisons can be made across product types, comparing the overall features of importance and from there, generalizing to determine what areas need refinement across the entire product group. This method can also be used in a cyclical manner, to keep track of changes in opinion and product specifications over time as new products emerge on the market. Tracking of this process could essentially be delivered as reports, directly from the consumers to those who need this information - designers, product developers, etc., and ultimately result in the delivery of a better product. This is a practical way to use crowdsourced data in the form of online reviews to inform a company on how customers think about and react to products and what is most important to them and urgent to fix; it is a method of feedback to manufacturers.

8. References

[1] Flanigin, A.J., Metzger, M.J., Pure, R., Markov, A. and E. Hartsell. “Mitigating risk in ecommerce transactions: perceptions of

information credibility and the role of user-generated ratings in product quality and purchase intention," Electronic Commercial Research, 14, 2014, pp. 1-23.

[2] Chen, C.C., and Y.-D. Tseng. "Quality evaluation of product reviews using an information quality framework", *Decision Support Systems*, 5, 2011, pp. 755-768.

[3] Mudambi, S.M., Schuff, D., and Z. Zhang. "Why aren't the stars aligned? An analysis of online review content and star ratings", In *Proceedings of 47th Hawaii International Conference on System Science*, 2014, pp. 3139-3147.

[4] Hu, M. and B. Liu, "Mining and summarizing customer reviews," *Proceedings of the tenth ACM SIGKDD International conference on Knowledge discovery and data mining*, Aug. 22–25, 2004, Seattle, Washington, pp. 168-177.

[5] Dellarocas, C. and R. Narayan, "What motivates consumers to review a product online? A study of the product-specific antecedents of online movie reviews," In: Aberer, K., Peng, Z., Rundensteiner, E.A., Zhang, Y., Li, X. (eds.) *WISE 2006*. LNCS, vol. 4255. Springer, Heidelberg (2006).

[6] Ghose, A. and P.G. Iperiotis. "Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics," *IEEE Transactions on Knowledge and Data Engineering*, 23(10), 2011.

[7] Mooney, R. J. and R. Bunescu, "Mining knowledge from text using information extraction," *ACM SIGKDD Explorations Newsletter - Natural language processing and text mining*, 7(1), June 2005, pp. 3-10.

[8] Suresh, K. P. and S. Chandrashekara, "Sample size estimation and power analysis for clinical research studies," *Journal of Human Reproductive Sciences*, Jan-Apr 2012, 5(1), pp. 7-14.

[9] J.S. Tanaka, "How big is big enough?": sample size and goodness of fit in structural equation models with latent variables, *Child Development*, 1987, 58, pp. 134-146.

[10] Ahuvia, " For the love of money: materialism and product love," in *SV - Meaning, Measure, and Morality of Materialism*, eds. Floyd W. Rudmin and Marsha Richins, Provo, UT: Association for Consumer Research, 1992, pp. 188-198.