PREDICTING EMERGING PRODUCT DESIGN TREND BY MINING PUBLICLY AVAILABLE CUSTOMER REVIEW DATA

Conrad Tucker[1] and Harrison M. Kim[1]
(1) University of Illinois at Urbana-Champaign, USA

ABSTRACT
In this work, the authors present a robust framework to enrich new product design process by dynamically capturing customer preference trends. The framework autonomously captures customer preference trends from publicly available product review data which is abundantly available but grossly underutilized. The method overcomes a major challenge that has plagued the product design community -- the lack of large scale, realistic customer data and its meaningful interpretation to guide new product design process. The challenge is from conventional, prevalent use of customer surveys or focus group interviews that are usually costly and time consuming while the size of available data is usually small scale. The framework is composed of three steps – retrieval of customer review texts, mining product feature texts, and predicting future trend of product preference.

Keywords: Text mining, trend mining, online customer reviews, product design

1 INTRODUCTION

The rapid expansion of internet usage, both domestically and globally, is helping to fuel an increased flow of information across many barriers. Technological successes such as Wikipedia, Facebook®, Twitter®, etc., are giving users a sense of empowerment in their ability to create and shape knowledge and information flow. These user-propelled networks have been referred to as *digitized word of mouth* networks that harness the true power of human communication [1].

Online customer reviews are becoming a viable source of large scale product review data. A recent research found that Amazon.com had over 10 million active customer reviews on all product categories [2]. A study by Forrester Research reported that 50% of customers who visited retailer sites with customer review feedback indicated that customer reviews were important or extremely important in their purchasing decisions [2].

Many research domains such as Marketing and Advertising, Medical Research, Quality Assurance, Computer Science [3, 4], etc., continue to investigate the potential of web based networks as a viable approach to data collection. However, research into the acquisition and integration of online information in the product design domain has been limited.

A few of the noticeable benefits of employing such approaches in customer data acquisition include the speed at which large sets of customer review data can be accessed and stored for next generation product design purposes. Another major benefit is that a significant portion of this data is based on customer *revealed preferences*; that is, the feedback given after a customer has purchased and interacted with a product for a considerable amount of time. This differs from *stated preference* data that typically involves a hypothetical purchasing scenario in the form of a survey [5].

The methodology presented in this work: (1) Searches through user specified customer review web sites for a given product; (2) For each unique user review, a text mining and retrieval algorithm is employed to isolate and store information specific to a given product review; (3) The stored text data for the entire review population is time stamped and mined for frequent feature patterns; (4) A time series predictive model is generated based on a specific time horizon; (5) Design engineers can utilize the generated model to understand evolving customer preference trends for next generation product design ideas.

This paper is organized as follows. This section provides a brief motivation and background; Section 2 describes previous works closely related to the current research; Section 3 describes the methodology; Section 4 introduces the cell phone case study and Section 5 concludes the paper.

2 RELATED WORK

2.1 Traditional Customer Preference Acquisition Techniques

There are several well established methodologies in the product design community that have been employed in the product design and development process. The Quality Function Deployment (QFD) is a methodology that attempts to translate customer requirements (CR) (also known as the Voice of the Customer (VOC)) into functional engineering targets in an effort to generate new design ideas and enhance quality [6].

Another popular approach to customer preference quantification is the Discrete Choice Analysis (DCA) technique, which includes the Probit Model and Logit Models (multinomial, mixed, nested, etc.) to name but a few. By quantifying product attribute levels, design engineers can estimate the demand of next generation products and also investigate creative approaches to help address customer preferences [7].

A major challenge resulting in the aforementioned customer preference modeling techniques is that the quality of the customer feedback is highly dependent on the framing of the survey questionnaires presented to current/potential customers. Recent research investigating the validity of customer surveys have revealed that a phenomenon called *self-generated validity effects* may adversely affect the quality of the customer response data [8]. Consequently, the creativity of the engineering design solution may be adversely affected if customer needs are not fully understood. For example, design engineers may be focused on developing creative approaches to enhancing the speed of a product while the emerging needs of the customer are more aligned with environmental safety. In addition to these complexities, the size of the survey data set is often quite limited due to the time and financial costs of generating surveys.

To alleviate some of these challenges, we propose a customer preference acquisition model based on large scale customer review data. The *unstructured* nature of online customer review data relieves respondents from the traditional predefined structure of a survey type approach and enables respondents to provide an unbounded assessment of their product preferences. The next section provides some background literature in the text mining research domain.

2.2 Online Customer Preference Acquisition Techniques

Considerable research has been done in both document summarization and text classification relating to individual words or group of words and their descriptive relations [9]. While many of the early text extraction and mining algorithms focused on document summarization, there have been several works more closely related to feature extraction as it relates to customer review data [10, 11]. Jindal and Liu investigate comparative sentence mining and employ sequential rules to compare customer sentiments towards similar products [12]. Dave et al. propose a feature selection classification algorithm that analyzes customer review data and automatically partitions words into positive and negative domains with relatively high accuracy [9]. A more recent contribution by Hu and Liu builds upon the work by Dave et al. by proposing a semantic classifier of product review sentences that does not need a set of training texts to build the classifier [13]. It has been reported in the literature that attempting to mine customer review data in order to separate positively and negatively associated words can be very challenging [14]. Although many methodologies have been proposed to try and address this issue in text mining, we propose overcoming these inherent challenges of text approximation by focusing on customer review sites that have partitioned the reviews into predefined categories of *pros* and *cons*. Customer review websites have begun adopting predefined *pros* and *cons* formats, ranging from customer product review sites such as Cnet to hotel and travel related sites such as Priceline.

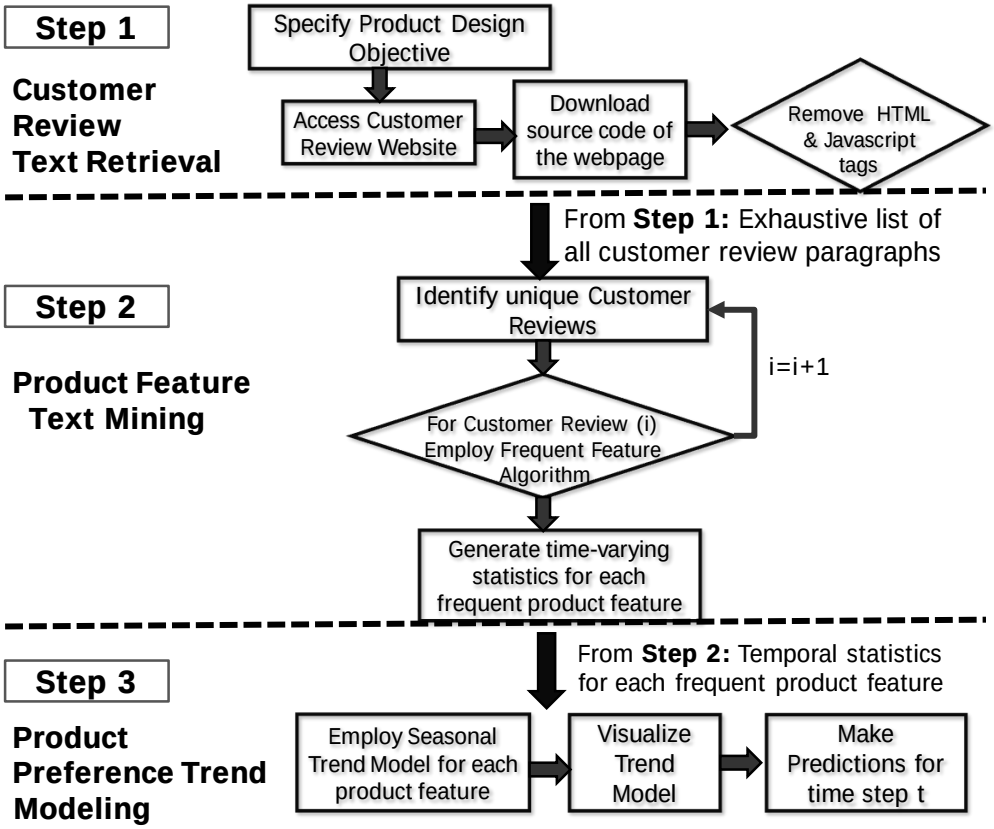

Fig. 1. Overall flow of from Customer Review Text Retrieval to Product Preference Trend Modeling

Another related field of text mining is search query analysis. In a recent research finding, online search queries were used to predict seasonal flu patterns within the same geographical region [3]. That is, there was a direct correlation between the temporal frequency of certain key query words that describe a flu (for example, fever, aches, sneezing, etc.) and the number of hospital visits for flu like symptoms [3].

The methodology presented in this work differs from the aforementioned product review based algorithms by assessing individual customer reviews (in its entirety) and employing a text classification algorithm to determine the most relevant product features being expressed by customers. Since each user review is time-stamped, the temporal nature of certain positive and negative product features can be quantified, hereby enabling us to model product feature preference trends.

3. METHODOLOGY

The proposed methodology enables design engineers to model product trends and identify emerging features that may be popular in future product design models or identify obsolete features that should be excluded from future product designs. By integrating free, publicly available customer review data, the proposed design methodology can have wide applicability to many areas of product design. The flow diagram in Figure 1 presents the overall proposed framework from customer review text retrieval to product preference trend modeling. A detailed explanation of each step of the flow diagram in Figure 1 will be presented in the following section.

3.1 Step 1: Customer Review Text Retrieval

The process of acquiring text based data begins by specifying the source of the customer review data. Figure 2 presents a snapshot of a typical cell phone customer review. To overcome some of the text analysis challenges discussed in section 2.2, data is acquired through customer review web sites that have a predefined partition of the positive (pros) and negative (cons) customer reviews

(ex:www.cnet.com). The pre-partitioned format of the customer feedback platform greatly reduces the complexities that would have resulted from trying to mine the raw data for negative customer opinions. A product design research web site was developed as a part of research, which acquires all of the raw customer review data (in html source code format) from multiple online sources.

"Good for games, but not for business"

by reality34 on March 17, 2010

Pros: Easy user interface
Bright clear screen
Many useful applications

Cons: No tactile response
Applications costly
Signal quality

Fig. 2. Text Summarized Customer Review Data (Partitioned into Pros and Cons)

3.1.1 Link between preference review text and new product design

The unconstrained nature of online customer review data is a departure from traditional survey type approaches and has the potential to enhance the overall product design process by enabling design engineers to understand the entire spectrum of customer needs. The formulation of a survey questionnaire, ironically assumes that the customer cares about the specific items presented within the survey. In addition to this bias, it has been reported in the literature that an individual's reported purchase intentions are biased towards a social norm whenever they are asked to make predictions about their future behavior (ex: An individual may express preference towards a green product due to social norms, despite whether or not this product will satisfy their needs) [8]. By extracting customer preferences though an unguided web-based format, design engineers can acquire large scale unpredictable, honest customer feedback in a timely and efficient manner.

Assessing customer's needs and preferences is critical in designing next generation products with improved, creative features. Collecting feedback from consumers often plays an important role as it was observed in many design firms' cases such as IDEO [15]. Instead of relying on focus group exercise or surveys based on physical prototypes due to the limitations described above, we can utilize a wealth of product review data. Review data is readily available on the web based on current or past generations of products. As a result, before the idea generation begins, design engineers can now be presented with the customer needs which have in the past been acquired through customer surveys and focus groups. By providing design engineers with real time assessment of customer preference trends, our proposed methodology will serve as a guide during the idea generation and concept feasibility phases. There are however some fundamental challenges of acquiring unstructured customer review which we aim to address in Step 2 of the product design process.

3.2 Step 2: Product Feature Text Mining

Customer review data in html format is stored on an SQL database on the web site which the authors developed. The next step is to determine the product features being expressed by customers. This is a non-trivial problem as customer reviews are presented in an unstructured, unpredictable style. We first begin by employing a PHP: Hypertext Preprocessor (PHP) based version of the Brill Tagger for each customer review sentence [16]. This enables each customer review sentence to be decomposed into their respective Part of Speech (POS), otherwise known as Part of Speech Tagging. For example, the customer feedback "Easy user interface" presented in Figure 2 would be transformed into:

- [easy-**JJ** user-**NN** interface-**NN]**

where **JJ** is the Tag code for an *adjective* and **NN** is the Tag code for a *noun*, with similar syntax used for other parts of speech. The results of the Tagger algorithm will therefore identify the product feature space within customer reviews that are represented by nouns. For each time step, a frequent

words algorithm searches through customer reviews to identify the most frequently expressed customer product preferences. The sensitivity of the frequent words search algorithm is dependent on the user specified minimum threshold (MinSup) for what is characterized as *frequent*.

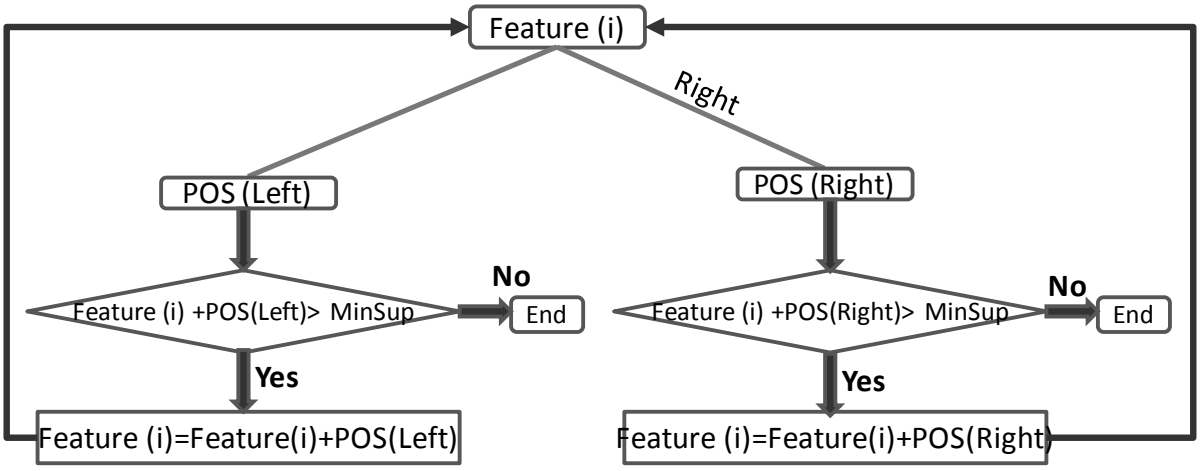

Fig. 3. Algorithm flow of product feature search aggregation.

One major challenge in refining the accuracy of the frequent product feature algorithm is to efficiently analyze and understand customer textual input. For example, the Tagger results above reveal that both the words *user* and *interface* are nouns, representing candidate product features. However from observation we realize that the customer is referring to a single product attribute *user interface*, rather than two separate attributes *user* and *interface*. We employ an Apriori-like algorithm that satisfies the anti-monotone Apriori property: *if any length k pattern is not frequent in the database, its length* (k+1) *super-pattern can never be frequent* [17, 18]. Figure 3 presents a visual flow of the frequent product feature extraction algorithm that enables words like *user* and *interface* to be grouped as one product feature *user interface*.

Once the frequent features have been identified for each time step, design engineers may want to know the *subjective* and *objective* terms that customers use to describe them. An example of a subjective descriptive term would be the word *easy*, used to describe the product feature *user interface*. An example of an objective description would be *16 Gigabytes* used to describe the product feature *hard drive*. We employ a Bayesian Classification learner to determine the associated terms (objective or subjective) relating to the most frequent features. This is mathematically represented as:

$$p\left(W_j|F_i\right) = \frac{p\left(F_i|W_j\right)\cdot p\left(W_j\right)}{p\left(F_i\right)} \tag{0}$$

where,

- F_i represents a particular product feature
- W_j represents the particular word (adjective, adverb, etc.) describing feature F_i

This allows us to classify a particular descriptive term (W_j) based on the probability estimates from equation (1) by assigning the objective/subjective term to the feature with the highest probability.

The association of customer sentiment with a given product feature allows us to transform the once unstructured customer review data into a high dimensional structured representation that can be used for traditional statistical analysis and data mining in the subsequent product design process [5, 19] .
This structured format can be downloaded directly from the research web site where engineers can search through and access their product of interest. It is important to note in Figure 4 that not all feature words (W_j) are present for each customer, which is a more realistic representation of market conditions, as different customers would express varying preferences of product features. For example in Figure 4, Customer 1 expresses a preference for Feature 1 using word W_1 while Customer M

expresses preference for Feature N using word W_8. While each customer is referring to the same product, they are interested in different features and use different words to represent their preferences.

Customers	Feature 1	…	Feature N
Customer 1	W_1	-	-
.	-	-	W_7
.	-	W_9	W_{11}
.	W_3	-	-
Customer M	-	-	W_8

Fig. 4. Transformation of unstructured customer review text data to structured customer feature preference data

3.2.1 Link between newly discovered preference and new product features

The ability to identify the positive and negative product features (over time) most frequently expressed by customers will serve as a valuable feedback tool in the product design process. Design engineers will be able to enhance the technological features that customers respond positively towards. They can also either substitute or eliminate product features receiving significant negative reviews. Equally as important as the feedback is the magnitude and speed (algorithm speed as well as speed of customer review input) at which this feedback can be acquired using the proposed online customer review process. As of August 2010, there exist more than 17,000 customer reviews on the research web site that spans more than 400 product models. Therefore enterprise decision makers will not only be able to acquire reviews about their own products, but also benchmark their electronic feedback to that of competitors also being discussed within the customer review platform. All of this insight can be acquired in a few seconds to generate a set of time series frequent product feature results.

3.3 Step 3: Product Preference Trend Modeling

The Holt-Winters exponential smoothing technique is employed to model the time series data due to its relatively high predictive accuracy compared to other models such as the STL, Box-Jenkins, AutoRegressive Integrated Moving Average (ARIMA), to name but a few [20, 21]. The model uses a weighted averaging technique that takes into account the local level, the trend, and the seasonal components of the time series. The (k) step-ahead forecasting model can therefore be represented as:

$$\hat{y}_t(k) = L_t + kT_t + I_{t-a+k} \tag{0}$$

where

Level L_t (the level component):

$$L_t = \alpha\left(y_t - I_{t-s}\right) + \left(1-\alpha\right)\left(L_{t-1} + T_{t-1}\right) \tag{0}$$

Trend T_t (the slope component):

$$T_t = \gamma\left(L_t - L_{t-1}\right) + (1-\gamma)T_{t-1} \tag{0}$$

Season I_t (the seasonal component):

$$I_t = \delta\left(y_t - L_t\right) + (1-\delta) I_{t-s} \quad (0)$$

The smoothing parameters α, γ, δ α, γ, δ are in the range [22] and are typically chosen to be anywhere from 0.02 to 0.2 as a default, although they can also be estimated by minimizing the sum of squared errors for one time step ahead [20, 21]. The starting values L_0, T_0 and I_0 also have to be initiated with conventional estimates [20].

3.4 Predictive Model Validation

There are several well established statistical techniques proposed to validate the predictive accuracy of time series models but there seems to be little consensus on a single time series model evaluation measure [23]. The more frequently used statistical validation techniques are the Mean Absolute Percent Error (MAPE) which is mathematically represented as:[23, 24]

$$MAPE = \frac{1}{F}\sum_{i=t+1}^{t+F}\frac{|x_i - y_i|}{x_i} \quad (0)$$

The Maximum Absolute Percent Error (MaxAPE), mathematically represented as:

$$MaxAPE = \max(i = t+1,...,t+F)(\frac{|x_i - y_i|}{x_i} \quad (0)$$

And the normalized Mean Sum of Squared Errors (NMSEE)

$$NMSSE = \frac{1}{F}\sum_{i=t+1}^{t+F}\frac{(x_i - y_i)^2}{Var(X)} \quad (0)$$

Where
F: represents the number of future time steps being modeled
x_i: represents the actual data point at time step t+1
y_i: represents the predicted data value at time step t+1
Var(X): represents the variance of the actual data points (X)

The level of predictive error allowed will depend on the tolerances specified by the design engineers and enterprise decision makers. Once an appropriate predictive model has been generated, the error tolerances from the model validation statistics can guide design engineers in terms of the potential accuracy of future product preference trend predictions. The cell phone design example presented in the next section will be used to address these questions as they relate to product design.

4 APPLICATION: PREDICTIVE MODELING IN FUTURE CELL PHONE DESIGN

By generating predictive models to forecast product feature trends, design engineer can observe which product design features are becoming obsolete or popular over time and incorporate these findings in future product design decisions. A visual representation of the trend mining model is presented in Figure 5 with the solid lines representing the time series historical data acquired from customer reviews and the dashed lines representing the Holt-Winters forecasts beyond the actual data. The predictive accuracy of the Holt-Winters forecast can be validated by shifting the *Model Validation Bar* seen in Figure 5 leftwords. The *Model Validation Bar* will calculate the disparity between the actual customer review data and the Holt Winters forecast using the statistical techniques defined in section 3.4. This validation mechanism allows design engineers to quantify the accuracy of customer reviews based on historial data and forecast future customer preferences towards certain product features.

The graphical predictive model enables engineers to analyze the customer preference trends of the entire product industry (ex: entire smart phone market in Figure 5) or isolate a particular product type

for analysis. If we take a closer look at Figure 5 (Smart Phone>Apple,all models), we can observe that in June, 2007, the product feature *Keyboard* has a relatively low preference but as we move through time, we observe that the *Keyboard* feature becomes the 3rd most popular feature expressed by customers.

With large scale customer preference trend data such as that proposed in this work, designers can better assess the emerging problems of a current product or favorable features of past products and integrate these findings in the next generation product concept phase. Design engineers also have the

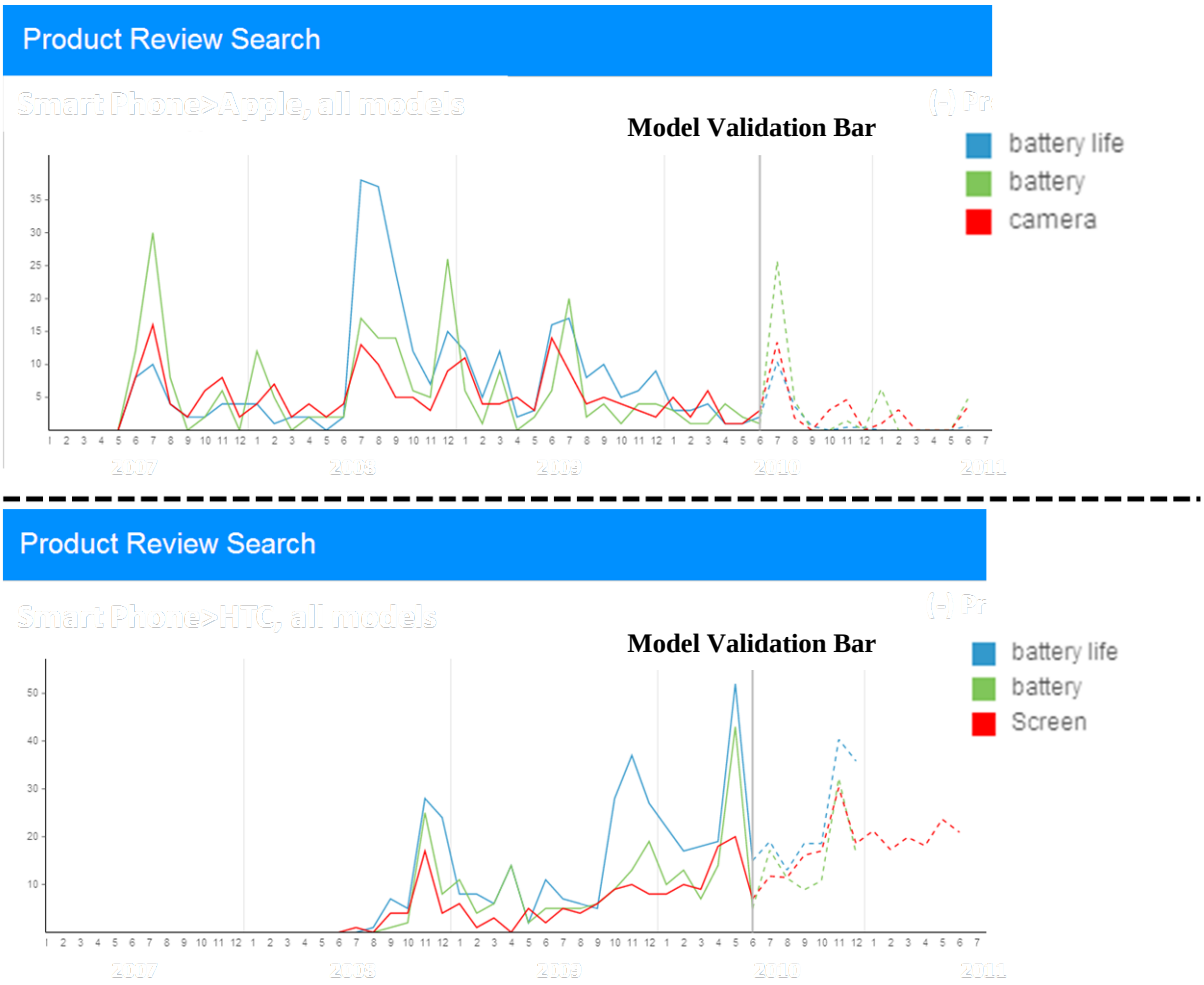

Fig. 5. Trend Comparison of negative product features between Apple and HTC

ability to benchmark their product to competitors by comparing the time series product features to those of competitors. Figure 5 presents a scenario where the *negative* product preferences are compared between two product brands, Apple and HTC. Figure 5 reveals that the negative product feature opinions of the two brands begin to differ at the 3rd highest negative product feature camera (for Apple) and screen (for HTC). Design engineers can incorporate this negative customer feedback into the design of next generation products by developing creative solutions to help address product deficiencies and minimize the threat of competitors. One of the major benefits of modeling customer preference trends using online, publicly available data is that models can be frequently updated as new customer reviews are accessed and mined. This data is readily available for download in a structured .CSV format.

5 CONCLUSION

In this work, we have proposed resolving the product design challenge of large scale customer data acquisition by proposing an online, publicly available customer product review methodology presented on the research web site. We propose a methodology that transforms unstructured customer

preference data into a time series representation of product feature preferences. Although the primary focus is on consumer electronics, the proposed methodology can be extended to other engineering fields such as automotive design, aviation logistics, etc. We aim to expand on the proposed online trend mining research by developing demand models for next generation products based on the time series customer review data.

ACKNOWLEDGEMENTS

The work presented in this paper is supported by Sandia National Labs, SURGE and the National Science Foundation under Award No. CMMI-0726934. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of Sandia National Labs, SURGE or the National Science Foundation. The authors would like to acknowledge Aukrit Unahalekhaka and Sukolsak Sakshuwong for data collection and programming aspects of this work.

REFERENCES

1. Dellarocas, C., *The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms.* Manage. Sci., 2003. **49**(10): p. 1407-1424.
2. Chen, Y. and J. Xie, *Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix.* Manage. Sci., 2008. **54**(3): p. 477-491.
3. Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data.* Nature, 2008. **457**(7232): p. 1012-1014.
4. Ferguson, T., *Online patient-helpers and physicians working together: a new partnership for high quality health care.* BMJ Journal, 2000. **321**(7269): p. 1129-1132.
5. Tucker, C.S. and H.M. Kim, *Data-Driven Decision Tree Classification for Product Portfolio Design Optimization.* Journal of Computing and Information Science in Engineering, 2009. **9**(4): p. 041004.
6. Pullman, M.E., W.L. Moore, and D.G. Wardell, *A comparison of quality function deployment and conjoint analysis in new product design.* The Journal of Product Innovation Management, 2002 **19**: p. 354-364.
7. Wassenaar, H.J. and W. Chen, *An Approach to Decision Based Design with Discrete Choice Analysis for Demand Modeling.* Transactions of ASME: Journal of Mechanical Design, 2003. **125**(3): p. 490-497.
8. Chandon, P., V.G. Morwitz, and W.J. Reinartz, *Do intentions really predict behavior? Self-generated validity effects in survey research.* Journal of Marketing, 2005. **69**(2): p. 1-14.
9. Dave, K., S. Lawrence, and D.M. Pennock, *Mining the peanut gallery: opinion extraction and semantic classification of product reviews*, in *Proceedings of the 12th international conference on World Wide Web*. 2003, ACM: Budapest, Hungary.
10. Hatzivassiloglou, V. and K.R. McKeown, *Predicting the semantic orientation of adjectives*, in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. 1997, Association for Computational Linguistics: Madrid, Spain.
11. Wiebe, J., *Learning Subjective Adjectives from Corpora*, in *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. 2000, AAAI Press.
12. Jindal, N. and B. Liu, *Mining comparative sentences and relations*, in *proceedings of the 21st national conference on Artificial intelligence - Volume 2*. 2006, AAAI Press: Boston, Massachusetts.
13. Hu, M. and B. Liu, *Mining and summarizing customer reviews*, in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, ACM: Seattle, WA, USA.
14. Hu, X. and B. Wu, *Classification and Summarization of Pros and Cons for Customer Reviews*, in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*. 2009, IEEE Computer Society.
15. Kelley, T. and T. Peters, *The Art of Innovation : Lessons in Creativity from IDEO, America's Leading Design Firm*. 2001: Currency.

16. Brill, E., *A simple rule-based part of speech tagger*, in *Proceedings of the workshop on Speech and Natural Language*. 1992, Association for Computational Linguistics: Harriman, New York.
17. Agrawal, R. and R. Srikant, *Fast Algorithms for Mining Association Rules in Large Databases*, in *Proceedings of the 20th International Conference on Very Large Data Bases*. 1994, Morgan Kaufmann Publishers Inc.
18. Pei, J. and J. Han, *Constrained frequent pattern mining: a pattern-growth view*. SIGKDD Explor. Newsl., 2002. **4**(1): p. 31-39.
19. Tucker, C.S. and H.M. Kim, *Optimal Product Portfolio Formulation by Merging Predictive Data Mining with Multilevel Optimization.* Transactions of ASME: ASME Journal of Mechanical Design, 2008. **130**(4): p. 041103-1-15.
20. Chatfield, C., *The holt-winters forecasting procedure.* Applied Statistics, 1978. **27**(3): p. 264-279.
21. Smith, B.L., B.M. Williams, and R. Keith Oswald, *Comparison of parametric and nonparametric models for traffic flow forecasting.* Transportation Research Part C: Emerging Technologies, 2002. **10**(4): p. 303-321.
22. *Towards a better understanding of modeling feasibility robustness in engineering design.* Transactions of ASME: Journal of Mechanical Design, 2000. **122**(4): p. 385-394.
23. Hyndman, R. and A. Koehler, *Another look at measures of forecast accuracy.* International Journal of Forecasting, 2006. **22**(4): p. 679-688.
24. Shimshoni, Y., N. Efron, and Y. Matias, *On the Predictability of Search Trends*. 2009.

Contact: Harrison M. Kim
University of Illinois at Urbana-Champaign
Department of Industrial & Enterprise Systems Engineering (IESE)
104 S Mathews Ave., Urbana, IL 61801, USA
Phone: (217) 265-9437
Fax: (217) 244-5705
E-mail: hmkim@uiuc.edu
URL: http://esol.ise.illinois.edu

Harrison Kim is an Assistant Professor in the Department of Industrial and Enterprise Systems Engineering at the University of Illinois at Urbana-Champaign. He joined the University of Illinois in 2005 and has been leading the Enterprise Systems Optimization Lab. His research interests include a variety of areas of system design and optimization utilizing mathematical programming, optimization, data mining, and informatics.

Conrad Tucker is a PhD student in the Department of Industrial and Systems Engineering at the University of Illinois at Urbana-Champaign. His research interests are in developing product portfolio design methodology utilizing data mining and knowledge discovery algorithms.