**Deriving Novel Parameters for the Assortment Decision Using Textual Analytics**

Rudraksh Syal, Srinija Vobugari, Ananth Nath, Kiran Samayam

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

[rsyal@purdue.edu](mailto:rsyal@purdue.edu), [svobugar@purdue.edu](mailto:svobugar@purdue.edu), [natha@purdue.edu](mailto:natha@purdue.edu), [ksamayam@purdue.edu](mailto:ksamayam@purdue.edu)

GitHub Link: <https://github.com/rudrakshsyal/RShiny-Application>

RShiny Application Link: <https://rsyal.shinyapps.io/RShinyApplication/>

**Abstract**

We employ a text content analysis framework that has been successful in other studies to identify important features of automobile service parts. We then show how such knowledge could be integrated into assortment planning model to help provide a category manager additional insight into which products to stock and which to remove. The motivation for this research is the fact that assortment planning decision is considered one of the most important decisions a firm will make because of the direct impact that the products offered (or not offered) will have on the firms key performance measures.

**Keywords:** Assortment Planning, Product Reviews, Text Analytics, Content Analysis, Sentiment

Analysis

**Business Problem**

Many important characteristics of the assortment decision that can be mined from consumer reviews can help improve an assortment that traditional predictive analytics techniques do not provide. One of the fundamental requirements needed to support this decision are accurate measures of demand, which may be unit forecasts, propensity to sell measures for very sparse demand situations, and consumer choice substitution behavior, which has been investigated heavily as of late. A lot is invested in this area by retailers to help them estimate demand as accurately as possible, with the idea that as demand accuracy is increased, a firm’s assortment performance will achieve or exceed expected results, and thus translate into satisfied customers.

**Analytics Problem**

Leveraging the information of the consumer reviews, we have found out that sentiment analysis techniques complemented with domain-specific smoke term dictionaries that are highly correlated with the assortment domain context could help in building a better performing text analytics framework. We construct a new “smoke” words dictionaries for service parts that hit on four major factors that matter to category managers in this retail domain: (1)fit, (2)quality, (3)price, and (4)vendor. We extend a text analytics framework that has proven successful in other automotive areas, in addition to non-related areas.

**Data**

There were two distinct datasets used in this study. The first data set used are product reviews from Amazon.com recorded between May 1996 and July 2014. This data set was obtained from Julian McAuley at UCSD (McAuley, Pandey, & Leskovec, 2015). Out of the 24 categories provided, we narrowed the focus their “Automotive-Replacement Parts” category.

**Methodology Selection**

According to a previous research study on around 250,000 automotive service parts reviews different types of data dictionaries were generated for words, bigrams and trigrams for each of the four product attributes – fit, quality, price, and vendor. By continuously feeding these dictionaries in to the holdout /test set, the quality of the dictionaries were improved through an iterative process.

**Model Building**

Based on this study, our team has built a custom function that feeds this dictionary with scores for unigram, bigrams and trigrams and finally categorises the reviews into – bad fit, poor quality, high priced, or unsatisfied with vendor. The model uses a method where a higher(more negative) score indicates more negative sentiment towards the corresponding category. We use an ensemble scoring methodology where the score would be taken from trigram dictionary initially and in case of absence of trigram scores, the scores would be obtained from bigram and unigram sequentially. The model classifies reviews to any of the four categories after scoring the review in each of them. The output of this model also acts as the backend data for the app.

**Functionality**

The Assortment planning(ShinyApp tool) dashboard helps the decision makers helps in content analysis of the reviews through different sections of the tool.

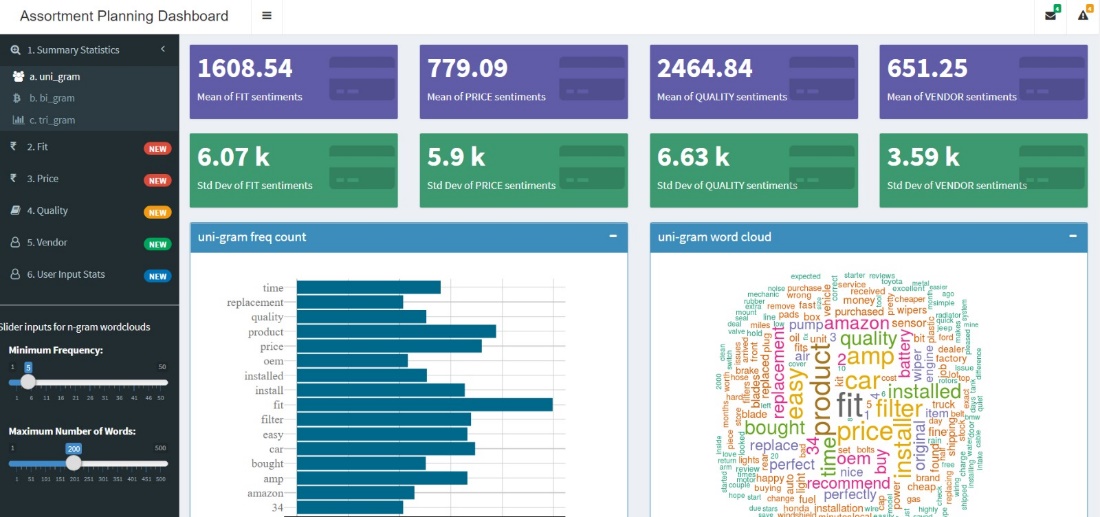
*Summary Statistics:* Smoke Dictionary Statistics for unigram , bigram and trigram scores and their distribution.

*Category Wise Tabs*: The Fit, Price , Quality and Vendor tabs show the sentiment scores and the word clouds after scoring the reviews and classifying them into these categories(using smoke dictionary).

*User Input Stats*: By inputting different reviews here, the tool shows the sentiment scores in each category and classifies them into one of the categories.

**GUI Design and Quality**

Below is the GUI Design which allows all the functionalities mentioned in the previous sections along with the options like hovering, filtering word frequencies, notification tool and so on.



Access to different sections of the Application

Notification Bar

Word Cloud  
(Filter Controls to the left side )

Word Freq & Max Words Filter

**Conclusions**

Using content analysis approach, smoke words dictionary was built with scores assigned to the Ngrams and an analytical model was created which feeds this dictionary into the model and classifies the reviews to four categories. With dynamic flow of data to this application, managers can understand the consumer opinions for the products at a glance and get notified with alerts and summaries.  
  
**References**

Ding, X., Liu, B., & Yu, P. S. (2008). *A holistic lexicon-based approach to opinion mining.* Paper presented at the Proceedings of the 2008 International Conference on Web Search and Data Mining.

Dellarocas, C., Gao, G., & Narayan, R. (2010). Are consumers more likely to contribute online reviews for hit or niche products? *Journal of Management Information Systems, 27*(2), 127-158

Gamon, M. (2004). *Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis.* Paper presented at the Proceedings of the 20th international conference on Computational Linguistics.

*Moe, W. W., & Schweidel, D. A. (2012). Online product opinions: Incidence, evaluation, and evolution. Marketing Science, 31(3), 372-386.*

*Riloff, E. (1996). Automatically generating extraction patterns from untagged text. Paper presented at the Proceedings of the national conference on artificial intelligence.*

*Riloff, E., Patwardhan, S., & Wiebe, J. (2006). Feature subsumption for opinion analysis. Paper presented at the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.*