

# Adaptive Text Extraction from Online Product Reviews for Marketing Intelligence

Thomas Y. Lee

The first step in the design and marketing of new products involves identifying a target market by segmenting the prospective user population in terms of their geographic, psychographic, demographic, and behavioralistic variables [Urban and Hauser 93]. The traditional approach to analyzing users involves combinations of surveys, focus groups, and one-on-one interviews. In this paper, we propose to augment traditional approaches for gathering marketing intelligence with the automated analysis of user-generated online product reviews. Specifically, we present a supervised, machine learning approach for sentential-level adaptive text extraction and mining. Based upon a set of 9700+ digital camera product reviews gathered in January 2008, we evaluate the approach in three ways. First, we document absolute performance using precision and recall on labeled data. We report the sensitivity of the method to various divisions of training-and-test data using n-fold cross-validation. Second, we compare the recall of automated learning with respect to traditional measures for identifying users and their respective needs. Finally, we use multi-dimensional scaling (MDS) to visualize the mapping between brands and segment variables.

Prior research on mining market value from user comments has focused on measuring the economic impact of online reviews. Specifically, how is the presence or absence of reviews and their corresponding textual features (length, valence, sentiment, polarity, etc.) reflected in price or sales [Cabral & Hortascu 06; Chevalier & Mayzlin; Das & Chen 07; Ghose & Ipeirotis 07; Pavlou & Dimoka 06]. Other research attempts to derive the market premium for specific product attributes based upon user comments [Ghose et al. 07]. Rather than learning about products and their prices or sales, this work focuses on learning about the customer population. In particular, we seek to extract segmentation variables including user activities, interests, and general behavior such as how specific customers use the product. For example: do the customers have children, do they travel, and what do they take pictures of? This work builds on early research in adaptive text extraction [Kushmerick & Weld 97; Califf & Mooney 98; Soderland 99; Muslea et al. 99].

In this work, we propose a supervised, population-based algorithm that learns sentence-level extraction patterns. Reviews are first decomposed into sentences. For each sentence, linguistic preprocessing identifies three pieces of meta-information: part-of-speech tags, subject-verb-object (SVO) triples, and the lemmatized forms of each text token. To train the model, we begin with a labeled set of SVO triples. The triples are subdivided into three parts: a prefix, a suffix, and the target where the target represents the text extraction. The three parts, taken together, form an extraction pattern. We initialize a population of patterns by transforming a labeled training set into patterns. Learning takes place in two steps. Two patterns are merged into a child pattern based upon the principle of longest-common-subsequence. The fitness of a child pattern with respect to its parents is measured as a function of the Laplacian estimated error and coverage. A population of patterns evolves by merging randomly selected parents and scoring the resulting children. The evolutionary process seeks to minimize an objective function defined by the sum of the Laplacian estimates over all patterns in the population. The process ends after a maximum number of evolutionary generations or a minimum change in objective function score over a threshold number of generations.

We have gathered 9700+ digital camera reviews from Epinions.com in January 2008. Using MontyLingua (Liu04), we parsed, tagged, chunked and lemmatized nearly 600,000 SVO triples. A training set of randomly selected reviews was independently coded by two research assistants. The training set covers 35 reviews numbering 3041 SVO triples. We have implemented the algorithms and performed 5-fold cross-validation over 60-40, 70-30, and 80-20 train-test splits of the data. For comparison to consumer surveys, we collected survey results from three recent Forrester Research North American and European technographics studies. Finally, following the multi-level hierarchical ordering process used in product design [Ulrich & Eppinger 04; Urban & Hauser 93], we condensed the training data to 524 observations. For the MDS, we used this training data that varies over 16 different products and 5 distinct brands.

All the described work has been completed to date but has not yet been formally written up. In the near future, we intend to focus in three areas: clustering, different forums, and different product domains. Rather than relying a manual process for hierarchically ordering the segmentation variables, we hope to employ automated methods in a post-processing step. Automating the hierarchical clustering will simplify the MDS visualization of changes in market structure and brand positioning over time. By analyzing data from additional forums, we hope to ask and answer questions regarding differences and/or similarities in user populations that post in different sites over the same product domains. Finally, prior research on the economic impact of reviews indicates that impact varies depending upon the product domain. We are interested in evaluating whether this approach is similarly sensitive to different domains. We have already gathered online reviews for three other product domains and are slowly working on teaching undergraduate coders to help label training data.