Online Review Mining: Health and Environmental Concerns on Beauty Products

Yoon Jin Ma, Illinois State University, USA
Jinseok Kim, University of Illinois at Urbana-Champaign, USA

Keywords: data-mining, beauty, cosmetics, reviews

As a response to the growing trends in healthier lifestyles, consumers are demonstrating an increased interest in and purchase of consumer products that can maximize benefits to their well-being. In particular, the cosmetic/beauty products industry is one of the rapidly growing industries, and is projected to reach revenue of $430 billion globally by 2022 (Allied Market Research, 2016). Due to the changing climate conditions, consumers show strong interest in skin and sun care products. Furthermore, as consumers become more conscious of health and environmental issues, there has been an increased demand for natural beauty products (Matic & Puh, 2016). Despite the sizable beauty market and a growing trend of natural beauty products, there is lack of research on consumer perception and purchasing behavior for such products.

Scholars in advertising, communication, marketing, and public relations have used various text-mining techniques to assess sentiments about brands, social issues, products, and policies (He, Zha, & Li, 2013; Mostafa, 2013). They have analyzed text data from consumer reviews and comments on social media platforms such as Facebook, Twitter, and Yelp. This line of research has been active due to the availability of large-scale social media data combined with increased computational capacities. This computational textual analysis has been supported by social scientists as a complementary method overcoming the data size problem and questionnaire-based approach through analyzing large-scale data unobtrusively generated by media users (Lazer et al., 2009). Introducing large-scale text analysis as its method, this study aimed to investigate the trend of consumers' interests or concerns over beauty products, focusing on health and environmental issues.

A dataset of 249,152 review comments by 177,345 Amazon users on around 75,000 beauty products during the 2004-2013 period was obtained for this study. Before analyzing those consumer reviews, we pre-processed the dataset following the standard procedures in text mining (Jurafsky & James, 2000; Manning & Schütze, 1999). First, all letters were changed to lower case. Second, we removed stop-words, which refers to common words in English such as "a," "the," "and," and "to," using the most widely used text-mining software, Stanford NLP. Next, each review went through tokenization, the process of breaking a text into words. Mechanics such as commas and periods were deleted, and a comment was split by spaces. Then, tokenized words were processed for stemming, the process of reducing inflected (or sometimes derived) words to their base or root form. For example, "wanted," "wants," or "wanting" was changed to "want." Using these pre-processed reviews, we conducted two analyses. First, we generated a list of 14 target words, presumably representing interests or concerns of beauty product consumers on health or environmental issues such as chemicals, natural, organic, toxic, hormone, healthy, safe, damage, dangerous, ethical, harmful, harsh, hazard, and risk. Those words were selected by reviewing product and consumer information about healthy and environmentally friendly living from a cosmetic database (ewg.org). Then, we counted how often each word in the list was found. Especially, we counted the frequencies of each target word with a yearly resolution to find a trend of change over time. Another analysis aimed to detect the main topics of interest reflected in the reviews per year. For this, we filtered the top 100 words frequently appearing in the reviews per year and sorted them into four categories (product, price, delivery, and satisfaction), following the steps described in He et al. (2013) and Mostafa (2013).

During the 2004-2013 period, the most frequently used words were "natural", "healthy", and "chemical". For instance, in 2012, when the most reviews (n=72,265) were posted during the target period, "natural" was observed 3,450 times, followed by "chemical" (n=1,126) and "healthy" (n=982). Overall, the frequencies of the target words increased over this period, which may imply a growing interest of consumers in health-related issues regarding beauty products. However, this increase could be simply due to the increase in the number of comments per year. In that, we normalized the frequencies of the target words by the total number of reviews each year. The normalized trend revealed that most target words decreased slightly over time. An exception was the word "toxic", which increased from 0% in 2004 to 0.18% in 2013. The main topics of the reviews did not center on health or environmental interests; they were mostly dedicated to the descriptions of product, price, delivery, and satisfaction. Specifically, the top three topics appearing in the reviews over time included "product" (n=154,571), "hair" (n=146,359), and "love" (n=77,655). These observations indicate that, contrary to our expectation, beauty product consumers did not demonstrate much interest in health or environmental concerns across their reviews. This study contributed to the body of existing literature by examining consumer trends in beauty products from the vast amount of online review data. These findings can help scholars obtain a better understanding of consumer perception and behavior on beauty products related to health and environment issues.

Although the dataset is sizable in terms of the number of reviewers, its representativeness is unknown because consumers who decided to leave reviews might not fully represent a population. In future research, we plan to analyze two million Amazon user reviews spanning 1996-2014 for in-depth analysis. In addition to the techniques used in this research, the future study will attempt a combination of automatic detection of topics and sentiment analysis.

Allied Market Research. (2016). Cosmetics market by category and by distribution channel. Retrieved from https://www.alliedmarketresearch.com/cosmetics-market

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining. *International Journal of Information Management, 33*(3), 464-472.

Jurafsky, D., & James, H. (2000). Speech and language processing an introduction to natural language processing, computational linguistics, and speech.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., . . . Van Alstyne, M. (2009). Social science: Computational social science. *Science, 323*(5915), 721-723.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: the MIT Press.

Matic, M. & Puh, B. (2016). Consumers' purchase intentions towards natural cosmetics. *EKONOMSKI VJESHNIK/ECONVIEWS, 29*(1), 53-64.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241-4251.