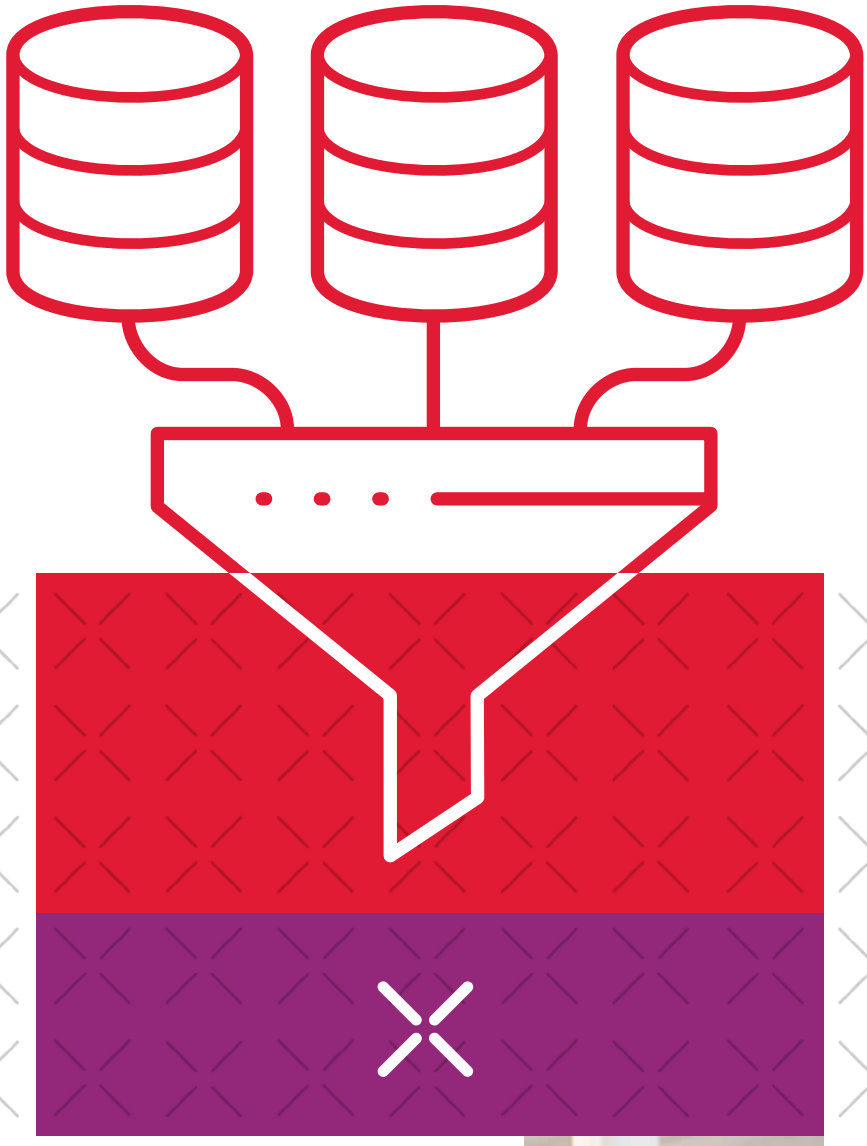
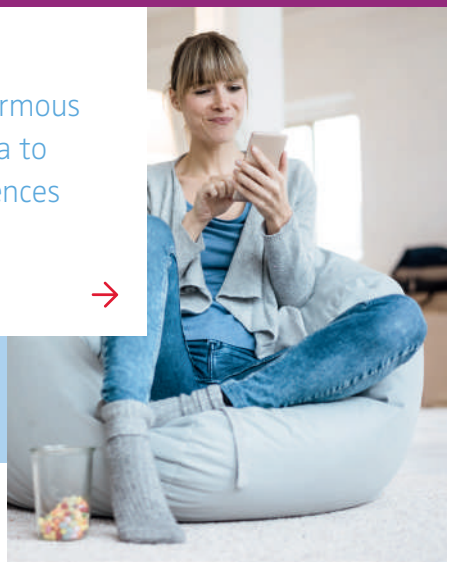


Artificial intelligence can help analyze enormous amounts of textual data from social media to generate insights about consumer preferences and decision-making.



Understanding Consumer Preferences from Social Media Data

Bradley Taylor

KEYWORDS
Text Analysis, AI, Preference Measurement

THE AUTHOR
Bradley Taylor
Black Swan Data,
Executive Data Science Director
(formerly GfK SE)

Reading "between the lines" of product reviews on a large scale When people consume in the digital space, they not only click and buy, but often comment on products, brands and services on social media, on platforms or the sites of online stores. The enormous amount of textual data consumers produce online is in fact a treasure box that hasn't yet been fully opened. But as data volumes grow, so do new algorithms to process and analyze unstructured data. Artificial Intelligence (AI) is one of the domains that can help open this treasure box further to better understand consumer decision-making.

In a GfK research project, we tested how we can learn consumer preferences and predict choices from publicly available social media and review data which are related to sales data. The common AI tool "Word Embeddings" has shown to be a powerful way to analyze the words that people use. It enabled us to reveal consumers' preferred brands, favorite features and main benefits. Language biases uncovered by the analysis can indicate preferences, and they fit reasonably well to actual brand sales within various categories. Especially when data volumes were large, the method produced very accurate results and it is completely passive (see Box 1). We have been using free, widespread online data without affecting respondents or leading them into ranking or answering questions they would otherwise not have even thought of. The analysis is fast to run, and no fancy processing power is needed.

Predicting the most preferred brands in one category To test if brand preferences could be derived from online reviews, we first ran the AI-based text analysis for one category (TVs) and different amounts of data and compared the outcome of the analysis to actual sales data. Specifically, we ran 3 experiments: using data from one online retailer only, encompassing a total of 3,000 reviews; using data from multiple retailers totalling 4,500 reviews (a random subsample of the whole data); and using the entire data set of 53,000 reviews.

The results are displayed in Figure 1. The first column shows the sales ranks of 5 brands in the category. It is important to note that the sales difference between brands C, D and E was quite small, and therefore we had expected some confusion. The 2nd column shows the results of 3,000 reviews scraped from a single online retailer. With this limited amount of data, the rank is clearly wrong, ranking the most-sold brands A and B on ranks 3 and 4 instead of 1 and 2. The 3rd column introduces multiple retailers with a random subsample of 4,500 reviews. In this experiment, Brand A is now in the correct position (1), but we see confusion on Brand B and the others. The 4th column, using the complete dataset of 53,000 reviews, shows the correct ranking for Brand A and B – the major volume drivers in the category – and confusion of Brands C, D, E.

BOX 1

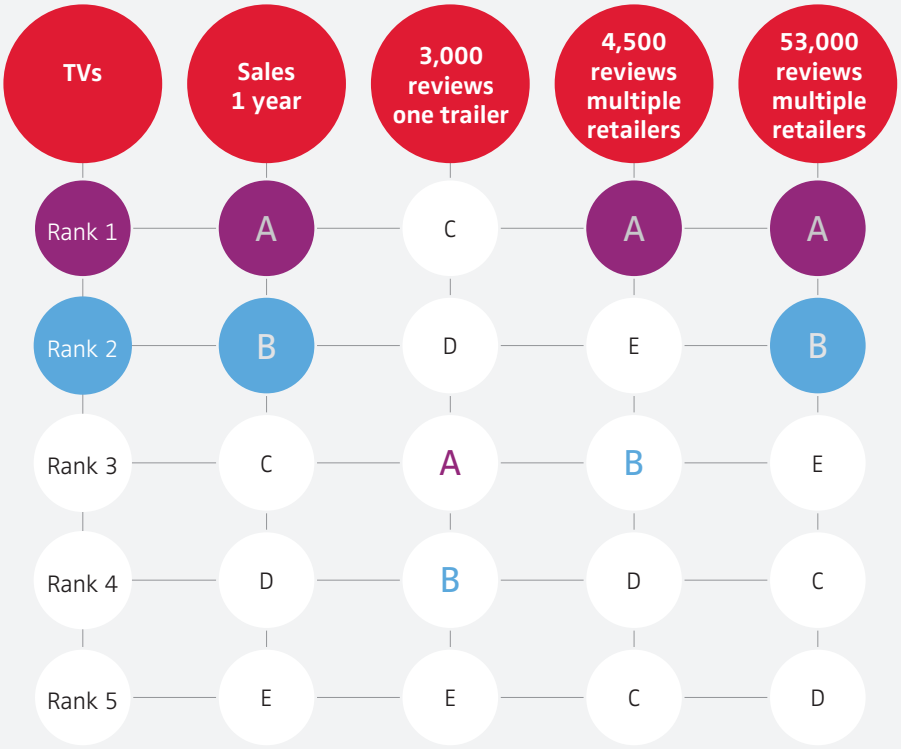
AI based natural language processing

We processed product review data via an AI method in the sub-domains of Natural Language Processing called Word Embeddings. The review data was scraped from online stores which have customer reviews of products sold. Additionally, we used a Twitter data set from tweets and comments which contained keywords around the category.

Word Embeddings enabled us to analyse the relationships of words to each other. This is accomplished by transforming words of a specific context into numerical vectors. The library used is called WordVectors. In our case the name of the category served as the context and we searched for words which were semantically most similar. For example, "TV" or "fridge" were used as the context (transformed into a vector) to look for similarities. We then used a closest-to function which generated a list of words with scores indicating how similar words were to this context. In the analysis, all special characters, numbers and white spaces were removed. All words were transformed to lower case and all stop words were removed. Sentence order of words is not important for this exercise.

In the next step, we used the degree of similarity of words to the context to get a rank order of brands and compare this order to actual unit sales rankings of the last 12 months. The point-of-sale data came from our GfK retail panels, which consists of multiple retailers' data aggregated to a total category picture by brand. All data reviews and point-of-sale data were from the U.K. market.

FIGURE 1 > Prediction accuracy of brand preferences in the TV category for different amounts of available data

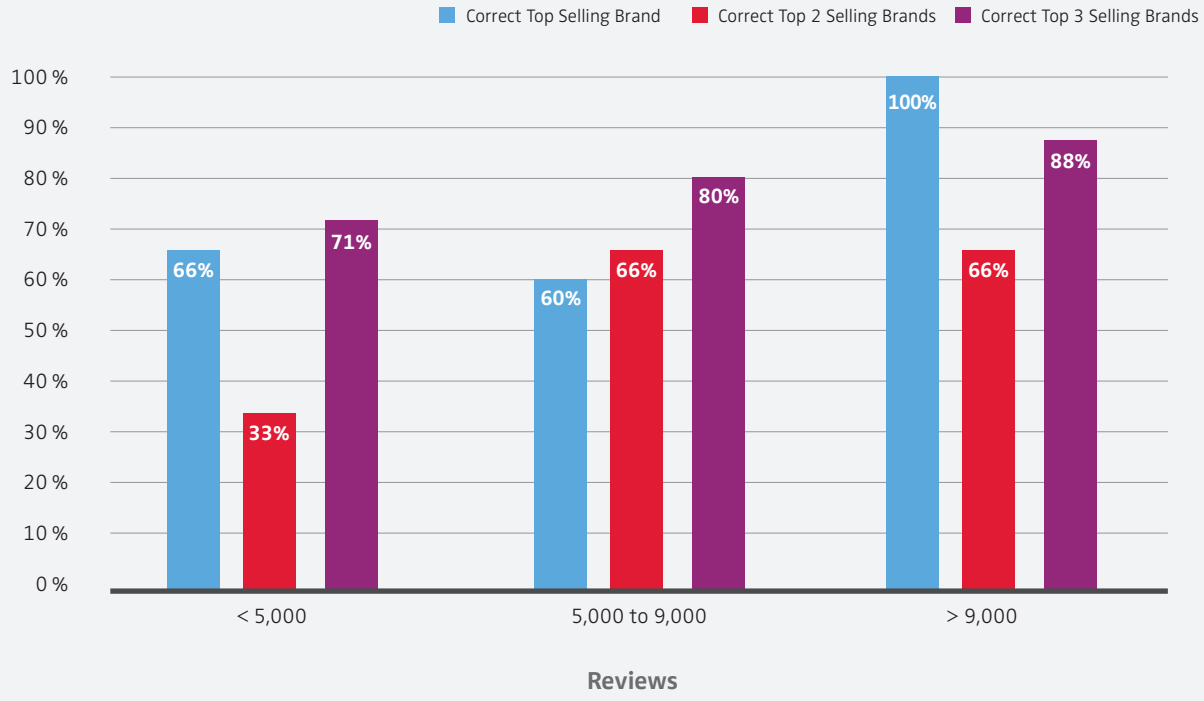


Predicting brand preferences in multiple categories ✕ Next, we ran experiments across 10 categories, again using point-of-sale data as the reference for the ranking. As in the first experiment, more data lead to more accurate results. The categories we tested were Tablets, Hand-held Vacuums, Full-size Vacuums, Smartphones, Hair Dryers, Fridges, Laptops, TVs, Shavers, Dishwashers and Washing Machines. We see that, to get correct brands in the top positions in at least 80% of all cases, a sample of 5,000+ is required, and that accuracy can be enhanced significantly with larger amounts of data (Figure 2). In the Smartphone category (using Twitter data), we made the interesting observation that the correct order prevails bar Alcatel. This brand was characterized by a disproportionate amount of spam, which could be the reason for the incorrect ordering of Alcatel.

Understanding preferences of product features ✕ As an extension to the method described before, we further tested if semantic networks from Word Embeddings can help us understand the bias toward not just brands but product feature preferences. Such an attempt to understand what people desire goes beyond buzz or sentiment analysis and toward mapping consumers' minds. The outcome of such a brand-plus-feature analysis could be similar to classic surveys on attribute preferences.

Our sample consisted of data from 36,000 reviews of TV sets. Picture quality emerged as the number one feature and sound quality came second. Sound quality's being so high up should not be surprising since the TV experience actually combines vision and sound, despite its categorization as a

FIGURE 2 > Ability to correctly identify top-selling brands from 10 different categories for different sample sizes



visual device. The next most-desired feature was ease of use, and then came the ability to connect with services such as YouTube or Netflix. We learned that unless the process for connecting the TV to the Wi-Fi router and internet is easy, people will not use the TV services for streaming. Not only does getting connected need to be self-explanatory, but the way in which third-party app providers link to the TV platform must be very user-friendly. There seems to be a double-barrelled problem here for adoption!

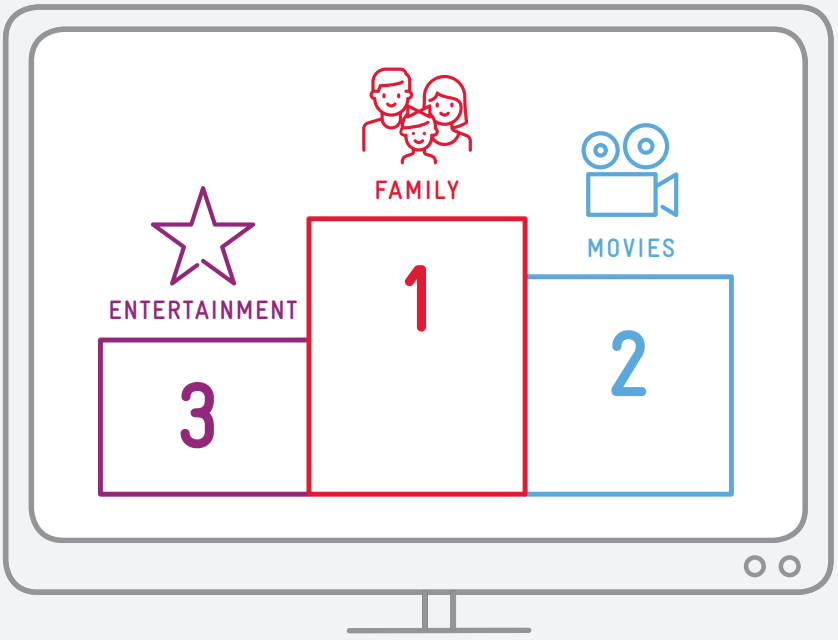
Last among the top five features was the look and style of the TV. Given the current focus on other elements of advertising, we can see how using AI to understand the consumer can help inform the marketing and even the innovation strategy of many manufacturers.

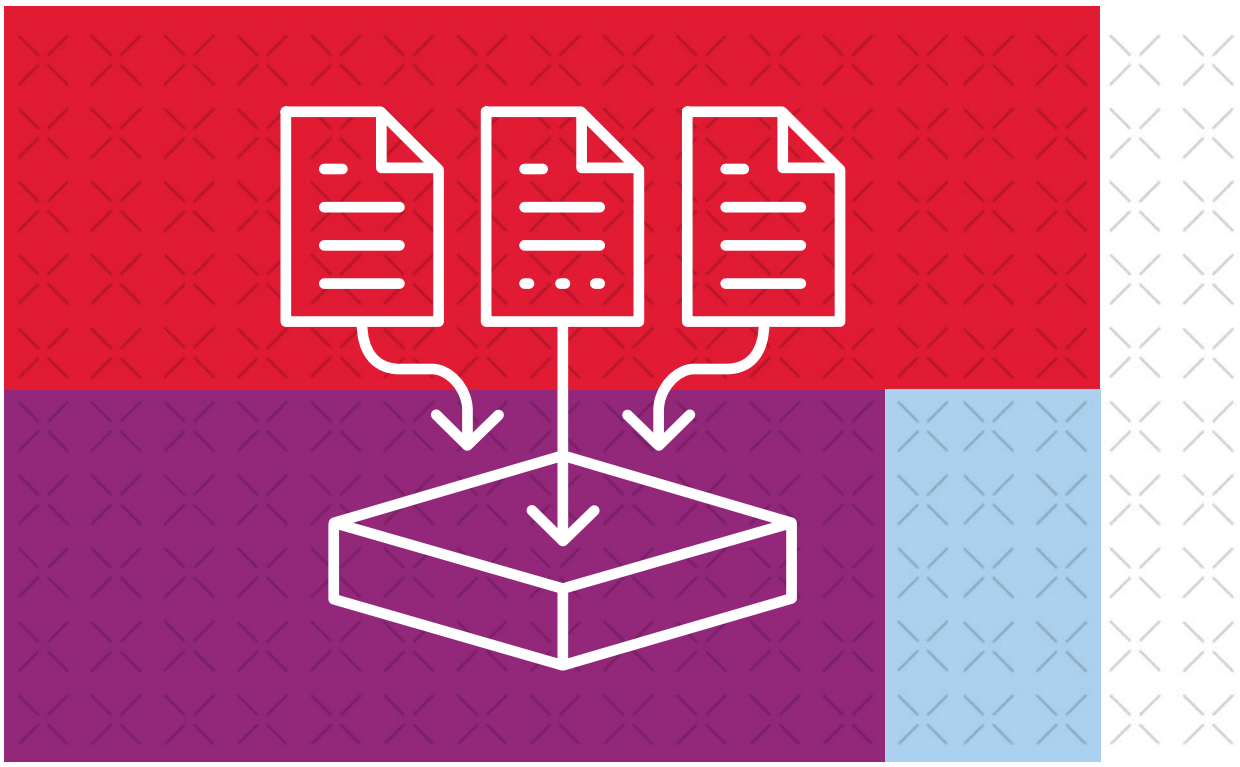
Understanding product benefits ⨯ Our extended analysis helped us understand not only top features but also purchase motivations and the benefits that consumers seek. Figure 3 lists the main topics that emerged and the roles they play. Family was ranked first, giving us the insight that TVs are not all about entertainment, or simply watching movies, but rather are devices which can be used and enjoyed by the entire family. Although the top features are picture and sound quality, the purchase motivation for TV sets goes beyond features and into their beneficial effects on home life. The communication implication is that TVs have a connection to social constructs such as families, and should be spoken about in a far more socially-connected manner. Simply stating their ability to provide the best picture and the most engaging movies and gaming experiences does not sufficiently address purchase motives.

Critical issues for future applications of AI-based text analysis ⨯ We have demonstrated that this type of analysis can help generate very useful insights. However, there is still a lot to learn, and several critical issues need to be considered.

› **The amount of available data is key for prediction quality** ⨯ Our results show that sample size is important in achieving reliable results. According to the results we presented in Figure 1 and Figure 2, we currently recom-

FIGURE 3 › Most-preferred benefits of TVs





mend a sample size of over 10,000 reviews or other text. In the range of +/- 10,000 samples, researchers should only look at the words closest to the context. When moving further away, the orders become less certain, and one can be less sure of the population's preferences.

› **Include reviews from different sources** In addition to the sheer amount of data, diversity of sources has shown to improve results. Figure 1 shows that using only one retailer introduced biases and, together with a small sample of reviews, led to the poor performance of the method. Due to some retailers and brands having exclusive offers, other more representative answers can be worked out by including as many retailers as possible. This is particularly important for brand rankings, and less so for feature or benefit rankings, so long as the brands in the dataset represent the features and benefits of the whole category.

› **Consider ambiguous meanings of words** Results must be interpreted with care when words become confused in their meaning. One example we encountered was "Hoover," which is both the trademark of a vacuum cleaner brand and is used generically as a verb meaning "vacuum." The word is also being used to describe the entire category. Unless the data is cleaned by removing the category and activity uses of the word, the results will inevitably show the brand in an overinflated position.

We have clearly begun to be able to represent the human in the machine. AI is helping us to understand consumers, to create mathematical models, and to represent complex preferences and consumer choices. It can accelerate our understanding of what is important to people, and hence their decision-making processes. AI augments the intelligence and skill of our experts and workforce, filling in knowledge gaps and reducing the cognitive load that the volumes of available diverse and continuous data sources represent. Eventually, AI and big data might enable us to learn insights from existing data that have previously been delivered by costly surveys and active questioning.

↓

FURTHER READING

http://arxiv.org/pdf/1301.3781.pdf
https://implicit.harvard.edu/implicit/