A Comparison of Sentiment Analysis Tools

Emergent Research Forum (ERF)

Amirkiarash Kiani
Ryerson University
amirkiarash.kiani@ryerson.ca

Sameh Al Natour
Ryerson University
salnatour@ryerson.ca

Ozgur Turetken
Ryerson University
turetken@ryerson.ca

Abstract

Sentiment analysis (SA), an analytics technique that assesses the "tone" of text, has emerged as a viable alternative to help users decide what to read without analyzing the whole text. In this study, we compare two main SA techniques (lexicon-based and machine-learning) for analyzing sentiments in the context of consumer product reviews. Given that a noted gap in prior research has been the almost sole focus on short textual information that concerns specific contexts (e.g., tweets about the Winter Olympics), we examine the role of contextual factors. Specifically, we examine reviews that address a multitude of product/service contexts, and which vary significantly in length.

To test the research model, we collected 625 consumer reviews that ranged in length from 10 to 551 words, and which concerned six goods belonging to the three product/service categories commonly cited in the literature: a) search goods: laptop computers and paper notebooks; b) experience goods: hotels and restaurants; and c) credence goods: car repair and multi-vitamins. To analyze the reviews, we used two tools: 1) VADER (Valence Aware Dictionary for sEntiment Reasoning); a parsimonious rule-based sentiment analysis tool that has been shown to be especially effective in analyzing social media posts, and 2) Google Cloud Natural Language API (henceforth called Google), which uses a machine-learning approach. To assess the effectiveness of the tools, we compared the sentiment scores generated by each tool to the star ratings that were provided by the original authors of the consumer reviews. Hence, the (absolute) difference between the SA scores and the star ratings served as our dependent measure.

The results of an ANOVA indicated that the lexicon-based approach (VADER) to sentiment analysis outperforms machine-learning in almost all product contexts regardless of review length (average error $M = 16.6\%$ vs. $M = 19.8\%$; $F = 13.81$, $p < 0.01$). The main effect of review length was also statistically significant ($F = 14.04$, $p < 0.01$), where the SA tools' accuracy was better for short and medium length reviews. Overall, both tools' average accuracy was highest for credence goods, followed by experience and then search goods ($F = 5.82$, $p < 0.05$). Overall, VADER demonstrated higher accuracy over Google for credence and search goods. Machine learning (Google) had a better accuracy only in the case of long reviews about experience goods ($F = 5.64$, $p < 0.05$).

The results of this study provide evidence that there are differences in the accuracy of various SA tools, and that it is important to consider contextual factors when choosing a SA tool. Building on these results, we will examine whether SA can be a substitute or a complement to star ratings in consumer decision making.

Keywords

Sentiment analysis, product reviews, text mining.

Acknowledgements

This research was partially supported by a joint research grant from Ryerson University and Hong Kong Polytechnic University.