Food Quality and Preference 82 (2020) 103888

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Automated sentiment analysis of Free-Comment: An indirect liking measurement?

M. Visalli[a,*], B. Mahieu[a], A. Thomas[b], P. Schlich[a]

[a] Centre des Sciences du Goût et de l'Alimentation, CNRS, INRAE, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France
[b] SensoStat, Dijon, France

1. Introduction

Predicting consumers liking thanks to emotion based measurements is a hot topic in sensory science. Several implicit methods for measuring emotions exist (Lagast, Gellynck, Schouteten, De Herdt, & De Steur, 2017), but the review did not mention Sentiment Analysis (SA), as it is relatively new. SA automatically determines the emotions of the author of a text from that text (Mohammad, 2016). When the underlying emotion is mentioned explicitly in the text, SA is consider explicit, otherwise it is considered implicit (Hu & Liu, 2004). Balahur, Hermida, and Montoyo (2012) nuanced this assertion, considering that expressions of sentiment are directly related to expressions of emotions in text and thus sentiments can be expressed directly (e.g. “I like Nokia phones.”), indirectly (e.g. “This phone is light as a feather.”) or implicitly, by describing a situation which points the reader towards a specific sentiment (e.g. “I paid 200€ for this phone and it broke in two days.”). More on implicit aspect extraction in sentiment analysis can be read in the review from (Tubishat, Idris, & Abushariah, 2018). Most applications of sentiment analysis rely on the exploitation of a large corpus of text coming from, for example, social networks. In sensory science, sentiment analysis has already been used to extract Twitter data for food-related consumer research (Vidal, Ares, Machín, & Jaeger, 2015). The authors highlighted the potential for sensory and consumer research but also identified several limitations, such as the cost of data analyses and the lack of representativeness of Twitter data. Another original application overcoming the second limitation consisted of applying sentiment analysis to Free JAR and Free-Comment data (Luc, Lê, & Philippe, 2020). Luc et al. used different lexicons and R libraries to compute sentiment scores from textual data and thus quantified the feelings of consumers about the tasted products. They disliked the subjective classification of the words and proposed considering machine learning as an alternative. Machine learning algorithms build mathematical models based on training data to make predictions that rely on patterns and inference without being explicitly programmed to perform the task (Ravi & Ravi, 2015). When applied to sentiment analysis, machine learning consists of using models with large datasets of text to detect known models or new ones that are similar to those that the machine learning already knows in the corpus of interest. Sentiment analysis using machine learning would avoid the time-consuming manual curation of the lexicon, reduce technical issues due to the inability to determine the exact meaning of words with multiple uses or misspellings and reduce subjectivity. In this study, an already trained machine learning algorithm of sentiment analysis was applied to sensory descriptions generated by consumers having performed a Free-Comment task to test if derived sentiment scores allowed the discrimination of products. Then, these emotional valence measurements were compared to hedonic scores given during the same studies to assess whether sentiment scores could be considered indirect liking measurements. The results are presented and discussed on the basis of 4 datasets from different product universes: chocolates, perfumes and wines.

2. Materials and methods

2.1. Sentiment analysis algorithm trained with machine learning

While the simplest possible approach relies on a manually curated lexicon of words or phrases that impart negative or positive sentiment to a sentence, the machine learning approach consists of training models that detect sentiments in a corpus of text. The training process requires a large dataset of text records already labeled with the sentiments for each record. The input text is tokenized into individual words, and stemming is applied. Then, depending on the algorithm, several features are used to train a classifier using concepts such as N-grams, part-of-speech tagging, word embedding and other natural language processing tools. Once the training process is completed, the classifier can be used to predict the sentiment of any new piece of text.

The Microsoft Text Analytics API (“What is the Text Analytics API?”, 2019) was used as the algorithm for sentiment analysis. The algorithm uses a combination of techniques in text analysis, including word processing, morphosyntactic analysis, word positioning, and word associations. The service has been benchmarked (Analytics, 2015) and proven effective (Harfoushi, Hasan, & Obiedat, 2018). Other algorithms exist (Comparison of the Most Useful Text Processing APIs, 2018), but

* Corresponding author.
E-mail address: michel.visalli@inrae.fr (M. Visalli).

https://doi.org/10.1016/j.foodqual.2020.103888
Received 10 December 2019; Received in revised form 7 January 2020; Accepted 16 January 2020
Available online 18 January 2020

Table 1
Investigated datasets.

Dataset	Panelists	Products	Location	FC modalities
D1	96	4 milk chocolates	Home	Flavor/Texture/Emotions
D2	88	4 ambiance perfumes	Lab	–
D3	63	5 dark chocolates	Home	–
D4	60	4 wines	Home	Mouth/Odor/Sight

Microsoft API has the advantage of being free and easily available using Web services or the R library.

2.2. Datasets

Table 1 details the 4 FC datasets used as examples. These studies were sponsored by the partners mentioned in acknowledgements, and they were designed for getting sensory descriptions of their products. Thus, they were not specially designed for the purpose of this article, but selected *a posteriori* because they presented a gradient of hedonic differences (large in D1, small in D4).

All Free-Comment tasks involved French consumers. Products were presented according to Williams Latin square designs and coded with 3-digit random numbers. For at-home studies, a minimum delay of 24 h between two product evaluations was enforced. Separate free comments were given by sensory modality for D1 and D4, whereas an overall description was given for both D2 and D3. A single overall liking score between 0 and 10 was given for each product after the Free-Comment task for D1, D3, D4 and before the Free-Comment task for D2.

2.3. Sentiment scores

No labeled or training data were needed to use the service, and it is important to note that the authors used the service as a black box and did not train the classifier. Thus, each Free-Comment dataset was submitted only to the latest (3.0) Microsoft Text Analytics online REST API, enabling each individual description to be converted to a sentiment score between 0 (negative feeling) and 1 (positive feeling), with 0.5 being neutrality or indetermination. Raw data were used, without the correction of grammar or spelling errors.

2.4. Data analysis

For each dataset, means and standard deviations of liking and sentiment scores (by modality and averaged over modalities, if applicable) were computed, and additive two-way ANOVA models with subjects and products as the source of variations, followed by product pairwise multiple comparisons (Tukey HSD tests) were used. F-products, p-values, and post hoc groups were reported. Pearson correlations were computed between liking and sentiment scores by considering the scores of each product by subject (centered by subject) as observations. For datasets D1 and D4, sentiment scores were computed by sensory modality and on their averages over the modalities. To facilitate comparison with sentiment scores, hedonic notes were transformed between 0 and 1.

3. Results

Table 2 shows that the sentiment scores were significantly correlated with liking and had rather good Pearson coefficients, especially for the average sentiment scores by sensory modalities. ANOVA was more discriminative with liking, but the post hoc groups were the same regardless of the score.

Table 3 also shows that the global sentiment score was significantly correlated with liking and had a good Pearson coefficient. ANOVA was still more discriminative with liking, and the post hoc groups were still the same.

Table 4 shows that the global sentiment score was significantly correlated with liking and had a lower Pearson coefficient. With the liking score, 1NV was discriminated from MAD, EQU and SAO. With sentiment scores, BRA was discriminated from EQU. In this study, the rankings of the products were clearly different depending on the liking or sentiment scores.

Table 5 shows that the mouth, odor and mean sentiment scores were significantly correlated with liking but not with the variable height Pearson coefficients. While the mean liking scores discriminated BOR from VAL and LAN, the sentiment scores did not. The rankings of the products differed depending on the liking or sentiment scores but were not so different, with the exception of sight—the only sentiment score not significantly correlated to liking.

4. Discussion and conclusion

Concerning sentiment analysis, using a machine learning algorithm opened up interesting perspectives. The Microsoft API was able to extract sentiment scores from description data regardless of whether the data were oriented by sensory modality or not. Unlike previous studies mentioned in the introduction, the API did not require any manual intervention, rendering the analysis of results almost immediate. The API was able to distinguish subtle nuances in descriptions, for example, the differences between different levels of sweet (NB: in this article, descriptions have been translated from French, where the word “sweet” is only associated with sugar, not with persons). The sentences “It is melty and sweet”, “It is melty, pleasant to eat, but slightly too sweet” and “Hazelnut; too sweet; I do not smell cocoa” had respective sentiment scores of 0.75, 0.63 and 0.25.

The products from datasets D1 and D2 were discriminated in the same way regarding liking or sentiment scores, but the liking differences were particularly marked between the 2 groups of products (more than half of the scale for D2 and between two-tenths and three-tenths for D3). When the differences were more subtle than those in D3 and D4, the sentiment scores were not discriminant (D4) or discriminated in a different way (D3). The visual descriptions had, in most cases, no emotional valence for the algorithm, though they could have such valence for the wine tasters; thus, sensory modalities such as sight were not appropriate for computing sentiment scores. In addition, as the sentiment scores were based on product descriptions, the algorithm would be unable to give different scores to products having been sensory characterized in the same way.

The sentiment scores were significantly correlated with liking scores, except for sight in D4. The hedonic scores could have been biased because they were given in the same session as the descriptions, after for datasets D1, D3, D4 and before for D2. Such a cognitive correlation between descriptions and liking should have played in favor of a systematic increase in correlations between liking and sentiment scores, which is not supported by the heterogeneity of those correlations among other studied products. However, significance is easily reached with a large number of observations (approximately 300 by dataset here), and the coefficients of correlation varied from fairly good (0.703 in D1) to poor (0.261 in D4). In any case, the liking scores were more discriminant than the sentiment scores.

As sentiment score was hypothesized to be an indirect measurement of liking, this less discriminatory power is not so surprising, but several issues contradict the preceding positive outcomes and have to be discussed.

First, the scores can be difficult to interpret. 0.5 could be either neutrality or indetermination. In addition, to truly consider sentiment scores a proxy of liking, it would have been better if the “negative” liking scores beyond 0.5 were reflected by negative sentiment scores beyond 0.5 and the same for the inverse. This was true for D1, but false for D2, where the less-liked perfumes still had positive sentiment scores; it was also false for D3, where some chocolates had negative sentiment

Table 2
Dataset D1 (milk chocolates). Means and standard deviations of liking and sentiment scores by product (post hoc test group), F-product of the ANOVA model (p-value) and Pearson coefficient of correlation (p-value) between liking and sentiment scores subject centered.

Dataset D1	Liking	Sentiment			
		Flavor	Texture	Emotion	Mean
P1	0.72 ± 0.21 (b)	0.69 ± 0.21 (b)	0.56 ± 0.25 (b)	0.76 ± 0.22 (b)	0.67 ± 0.16 (b)
P2	0.72 ± 0.20 (b)	0.66 ± 0.20 (b)	0.52 ± 0.20 (b)	0.76 ± 0.21 (b)	0.65 ± 0.14 (b)
P3	0.17 ± 0.21 (a)	0.46 ± 0.21 (a)	0.35 ± 0.20 (a)	0.44 ± 0.26 (a)	0.40 ± 0.14 (a)
P4	0.17 ± 0.20 (a)	0.44 ± 0.21 (a)	0.30 ± 0.18 (a)	0.44 ± 0.26 (a)	0.41 ± 0.16 (a)
F Product (p)	199.257 (< 0.001)	37.048 (< 0.001)	30.889 (< 0.001)	51.626 (< 0.001)	86.894 (< 0.001)
Correlation (p)		0.580 (< 0.001)	0.538 (< 0.001)	0.624 (< 0.001)	0.703 (< 0.001)

Table 3
Dataset D2 (perfumes). Means and standard deviations of liking and sentiment scores by product (post hoc test group), F-product of the ANOVA model (p-value) and Pearson coefficient of correlation (p-value) between liking and sentiment scores subject centered.

Dataset D2	Liking	Sentiment
P1	0.65 ± 0.24 (b)	0.70 ± 0.18 (b)
P2	0.64 ± 0.25 (b)	0.69 ± 0.21 (b)
P4	0.46 ± 0.29 (a)	0.59 ± 0.20 (a)
P3	0.39 ± 0.33 (a)	0.52 ± 0.22 (a)
F Product (p)	21.770 (< 0.001)	15.712 (< 0.001)
Correlation (p)		0.549 (< 0.001)

Table 4
Dataset D3 (dark chocolates). Means and standard deviations of liking and sentiment scores by product (post hoc test group), F-product of the ANOVA model (p-value) and Pearson coefficient of correlation (p-value) between liking and sentiment scores subject centered.

Dataset D3	Liking	Sentiment
MAD	0.68 ± 0.20 (b)	0.49 ± 0.12 (ab)
EQU	0.67 ± 0.19 (b)	0.53 ± 0.13 (b)
SAO	0.66 ± 0.21 (b)	0.48 ± 0.12 (ab)
BRA	0.60 ± 0.24 (ab)	0.47 ± 0.13 (a)
1NV	0.51 ± 0.31 (a)	0.51 ± 0.15 (ab)
F Product (p)	6.303 (< 0.001)	2.600 (0.037)
Correlation (p)		0.447 (< 0.001)

scores and positive liking scores. However, differences were not large enough and interpreting one as negative and the other one as positive may be exaggerated.

Second, the algorithm was not specifically trained on sensory datasets and could also deal with political speeches or opinions of hotel consumers. For example, in D3, 1NV (perceived as sweeter) was the less-liked dark chocolate at the panel level, while BRA (perceived as more bitter) had the worst average sentiment. Thus, except when associated with quantificators ("too much", "not enough", etc.), whatever the context, sweet alone was considered by the algorithm to be positive, with a sentiment score of 0.98, and bitter was considered negative, with a score of 0.05. This is highly questionable in a product space such as dark chocolates. Training the classifier with sensory datasets and even exclusively with dark chocolate datasets would certainly increase quality of sentiment scores obtained from food description.

Third, in some cases, the sentiment score was slightly dubious. For example, the descriptions of "tender, not too hard, does not stick", "no taste", and "at the beginning, a good taste of chocolate, then an acidic flavor appears in the mouth, making the taste unpleasant" were associated with sentiments scores of 0.21, 0.76 and 0.78 and liking scores of 0.8, 0.4 and 0.2, respectively. The Microsoft API is a black box that is subject to changes and improvements, and it should be verified whether other algorithms give "better" results, even if some correlations have been established between different sentiment analysis algorithms (Yoon et al., 2017). It should be kept in mind that these methods were originally developed for use on long texts and were not necessarily expected to work at this level of specificity. This has been confirmed in a recent study that suggested that automated classified sentiments were not as valuable as those made by humans (Jussila, Vuori, Okkonen, & Helander, 2017).

Considering the positives and negatives, from this study, it seems that the scores from sentiment analysis are not reliable enough to be considered indirect measures of liking scores, at least when the sensory differences between products are small. However, without taking the measure so far, sentiment analysis using a classifier trained with machine learning remains an interesting way to extract a sentiment score from text with sensory descriptions. The benefits of applying such methods to sensory analysis remain to be investigated.

CRediT authorship contribution statement

M. Visalli: Conceptualization, Formal analysis, Methodology, Software, Writing - original draft, Writing - review & editing, Visualization. **B. Mahieu:** Methodology, Resources, Data curation, Validation, Writing - review & editing. **A. Thomas:** Methodology, Writing - review & editing, Funding acquisition. **P. Schlich:** Methodology, Writing - review & editing, Supervision, Funding acquisition.

Table 5
Dataset D4 (wines). Means and standard deviations of liking and sentiment scores by product (post hoc test group), F-product of the ANOVA model (p-value) and Pearson coefficient of correlation (p-value) between liking and sentiment scores subject centered.

Dataset D4	Liking	Sentiment			
		Mouth	Odor	Sight	Mean
BOR	0.66 ± 0.18 (b)	0.65 ± 0.26 (a)	0.62 ± 0.21 (a)	0.67 ± 0.20 (a)	0.62 ± 0.15 (a)
GAM	0.60 ± 0.21 (ab)	0.68 ± 0.24 (a)	0.60 ± 0.18 (a)	0.64 ± 0.23 (a)	0.64 ± 0.13 (a)
VAL	0.54 ± 0.21 (a)	0.60 ± 0.22 (a)	0.59 ± 0.19 (a)	0.63 ± 0.19 (a)	0.59 ± 0.11 (a)
LAN	0.53 ± 0.19 (a)	0.61 ± 0.29 (a)	0.60 ± 0.18 (a)	0.68 ± 0.20 (a)	0.59 ± 0.14 (a)
F Product (p)	5.875 (< 0.001)	1.139 (0.335)	0.442 (0.722)	1.115 (0.33)	1.167 (0.323)
Correlation (p)		0.463 (< 0.001)	0.261 (< 0.001)	0.067 (0.301)	0.517 (< 0.001)

Acknowledgments

The authors would like to thank Robert et Marcel®, Sicarex, Barry Callebaut® and l'Occitane en Provence® for providing their products.

The Région Bourgogne Franche-Comté and SensoStat are thanked for their funding for Benjamin Mahieu's Ph.D. program.

References

Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems, 53*(4), 742–753.

Comparison of the Most Useful Text Processing APIs. (2018). Retrieved November 7, 2019, from https://activewizards.com/blog/comparison-of-the-most-useful-text-processing-apis/.

Harfoushi, O., Hasan, D., & Obiedat, R. (2018). Sentiment analysis algorithms through azure machine learning: Analysis and comparison. *Modern Applied Science, 12*(7), 49.

Hu, M., & Liu, B. (2004). Mining Opinion Features in Customer Reviews. Retrieved from www.aaai.org.

Introducing Text Analytics in the Azure ML Marketplace. (2015). Retrieved December 2, 2019, from https://blogs.technet.microsoft.com/machinelearning/2015/04/08/introducing-text-analytics-in-the-azure-ml-marketplace/.

Jussila, J., Vuori, V., Okkonen, J., & Helander, N. (2017). Reliability and Perceived Value of Sentiment Analysis for Twitter Data.

Lagast, S., Gellynck, X., Schouteten, J. J., De Herdt, V., & De Steur, H. (2017). Consumers' emotions elicited by food: A systematic review of explicit and implicit methods. *Trends in Food Science & Technology, 69*, 172–189.

Luc, A., Lê, S., & Philippe, M. (2020). Nudging consumers for relevant data using Free JAR profiling: An application to product development. *Food Quality and Preference, 79*, 103751.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion Measurement,* 201–237.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14–46.

Tubishat, M., Idris, N., & Abushariah, M. A. M. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, opportunities, and open challenges. *Information Processing and Management, 54*(4), 545–563.

Vidal, L., Ares, G., Machín, L., & Jaeger, S. R. (2015). Using Twitter data for food-related consumer research: A case study on "what people say when tweeting about different eating situations". *Food Quality and Preference, 45*, 58–69.

What is the Text Analytics API? (2019). Retrieved November 6, 2019, from https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/overview.

Yoon, S., Parsons, F., Sundquist, K., Julian, J., Schwartz, J. E., Burg, M. M., ... Diaz, K. M. (2017). Comparison of different algorithms for sentiment analysis: Psychological stress notes. *Studies in Health Technology and Informatics, 245*, 1292.