



**HAL**  
open science

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

Michel Visalli, Benjamin Mahieu, Arnaud Thomas, Pascal Schlich

► **To cite this version:**

Michel Visalli, Benjamin Mahieu, Arnaud Thomas, Pascal Schlich. Automated sentiment analysis of Free-Comment: An indirect liking measurement?. Food Quality and Preference, 2020, 82, pp.103888. 10.1016/j.foodqual.2020.103888 . hal-02866671

**HAL Id: hal-02866671**

**<https://hal.inrae.fr/hal-02866671>**

Submitted on 21 Jul 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial | 4.0 International License

## 1 Title

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

## 3 Authors

4 Visalli<sup>1</sup>, M., Mahieu<sup>1</sup>, B., Thomas<sup>2</sup>, A., Schlich<sup>1</sup>, P.

5 <sup>1</sup>Centre des Sciences du Goût et de l'Alimentation, CNRS, INRA, Univ. Bourgogne Franche-Comté, F-21000  
6 Dijon, France.

7 <sup>2</sup>SensoStat, Dijon, France

## 8 1. Introduction

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

40 during the same studies to assess whether sentiment scores could be considered indirect liking  
41 measurements. The results are presented and discussed on the basis of 4 datasets from different  
42 product universes: chocolates, perfumes and wines.

## 43 2. Materials and methods

### 44 2.1. Sentiment analysis algorithm trained with machine learning

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

53 The Microsoft Text Analytics API (“What is the Text Analytics API?”, 2019) was used as the algorithm  
54 for sentiment analysis. The algorithm uses a combination of techniques in text analysis, including  
55 word processing, morphosyntactic analysis, word positioning, and word associations. The service has  
56 been benchmarked (“Introducing Text Analytics in the Azure ML Marketplace”, 2015) and proven  
57 effective (Harfoushi, Hasan, & Obiedat, 2018). Other algorithms exist (“Comparison of the Most  
58 Useful Text Processing APIs,” 2018), but Microsoft API has the advantage of being free and easily  
59 available using Web services or the R library.

### 60 2.2. 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

61 Table 1: investigated datasets

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

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

### 72 2.3. Sentiment scores

73 No labeled or training data were needed to use the service, and it is important to note that the  
74 authors used the service as a black box and did not train the classifier. Thus, each Free-Comment

75 dataset was submitted only to the latest (3.0) Microsoft Text Analytics online REST API, enabling each  
 76 individual description to be converted to a sentiment score between 0 (negative feeling) and 1  
 77 (positive feeling), with 0.5 being neutrality or indetermination. Raw data were used, without the  
 78 correction of grammar or spelling errors.

## 79 2.4. Data analysis

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

## 88 3. Results

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)

89 Table 2: Dataset D1 (milk chocolates). Means and standard deviations of liking and sentiment scores  
 90 by product (post hoc test group), F-product of the ANOVA model (p-value) and Pearson coefficient of  
 91 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)

92 Table 3: Dataset D2 (perfumes). Means and standard deviations of liking and sentiment scores by  
 93 product (post hoc test group), F-product of the ANOVA model (p-value) and Pearson coefficient of  
 94 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)

95 Table 4: Dataset D3 (dark chocolates). Means and standard deviations of liking and sentiment scores  
 96 by product (post hoc test group), F-product of the ANOVA model (p-value) and Pearson coefficient of  
 97 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)

98

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

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

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

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

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

#### 117 **4. Discussion and conclusion**

118 Concerning sentiment analysis, using a machine learning algorithm opened up interesting  
119 perspectives. The Microsoft API was able to extract sentiment scores from description data  
120 regardless of whether the data were oriented by sensory modality or not. Unlike previous studies  
121 mentioned in the introduction, the API did not require any manual intervention, rendering the  
122 analysis of results almost immediate. The API was able to distinguish subtle nuances in descriptions,  
123 for example, the differences between different levels of sweet (NB: in this article, descriptions have  
124 been translated from French, where the word “sweet” is only associated with sugar, not with  
125 persons). The sentences "It is melty and sweet", "It is melty, pleasant to eat, but slightly too sweet"  
126 and "Hazelnut; too sweet; I do not smell cocoa" had respective sentiment scores of 0.75, 0.63 and  
127 0.25.

128 The products from datasets D1 and D2 were discriminated in the same way regarding liking or  
129 sentiment scores, but the liking differences were particularly marked between the 2 groups of  
130 products (more than half of the scale for D2 and between two-tenths and three-tenths for D3). When

131 the differences were more subtle than those in D3 and D4, the sentiment scores were not  
132 discriminant (D4) or discriminated in a different way (D3). The visual descriptions had, in most cases,  
133 no emotional valence for the algorithm, though they could [have](#) such valence for the wine tasters;  
134 thus, sensory modalities such as sight were not appropriate for computing sentiment scores. In  
135 addition, as the sentiment scores were based on product descriptions, the algorithm would be  
136 unable to give different scores to products having been sensory characterized in the same way.

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

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

149 First, the scores can be difficult to interpret. 0.5 could be either neutrality or indetermination. In  
150 addition, to truly consider sentiment scores a proxy of liking, it would have been better if the  
151 “negative” liking scores beyond 0.5 were reflected by negative sentiment scores beyond 0.5 and the  
152 same for the inverse. This was true for D1, but false for D2, where the less-liked perfumes still had  
153 positive sentiment scores; it was also false for D3, where some chocolates had negative sentiment  
154 scores and positive liking scores. [However, differences were not large enough and interpreting one  
155 as negative and the other one as positive may be exaggerated.](#)

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

165 Third, in some cases, the sentiment score was slightly dubious. For example, the descriptions of  
166 “tender, not too hard, does not stick”, “no taste”, and “at the beginning, a good taste of chocolate,  
167 then an acidic flavor appears in the mouth, making the taste unpleasant” were associated with  
168 sentiments scores of 0.21, 0.76 and 0.78 and liking scores of 0.8, 0.4 and 0.2, respectively. The  
169 Microsoft API is a black box that is subject to changes and improvements, and it should be verified  
170 whether other algorithms give “better” results, even if some correlations have been established  
171 between different sentiment analysis algorithms (Yoon et al., 2017). It should be kept in mind that  
172 these methods were originally developed for use on long texts and were not necessarily expected to

173 work at this level of specificity. This has been confirmed in a recent study that suggested that  
174 automated classified sentiments were not as valuable as those made by humans (Jussila, Vuori,  
175 Okkonen, & Helander, 2017).

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

## 182 **Acknowledgments**

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

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

## 187 **References**

- 188 Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Detecting implicit expressions of emotion in text:  
189 A comparative analysis. *Decision Support Systems*, 53(4), 742–753.
- 190 Comparison of the Most Useful Text Processing APIs. (2018). Retrieved November 7, 2019, from  
191 <https://activewizards.com/blog/comparison-of-the-most-useful-text-processing-apis/>
- 192 Harfoushi, O., Hasan, D., & Obiedat, R. (2018). Sentiment Analysis Algorithms through Azure Machine  
193 Learning: Analysis and Comparison. *Modern Applied Science*, 12(7), 49.
- 194 Hu, M., & Liu, B. (2004). *Mining Opinion Features in Customer Reviews*. Retrieved from [www.aaai.org](http://www.aaai.org)
- 195 Introducing Text Analytics in the Azure ML Marketplace. (2015). Retrieved December 2, 2019, from  
196 [https://blogs.technet.microsoft.com/machinelearning/2015/04/08/introducing-text-analytics-](https://blogs.technet.microsoft.com/machinelearning/2015/04/08/introducing-text-analytics-in-the-azure-ml-marketplace/)  
197 [in-the-azure-ml-marketplace/](https://blogs.technet.microsoft.com/machinelearning/2015/04/08/introducing-text-analytics-in-the-azure-ml-marketplace/)
- 198 Jussila, J., Vuori, V., Okkonen, J., & Helander, N. (2017). *Reliability and Perceived Value of Sentiment*  
199 *Analysis for Twitter Data*.
- 200 Lagast, S., Gellynck, X., Schouteten, J. J., De Herdt, V., & De Steur, H. (2017). Consumers' emotions  
201 elicited by food: A systematic review of explicit and implicit methods. *Trends in Food Science &*  
202 *Technology*, 69, 172–189.
- 203 Luc, A., Lê, S., & Philippe, M. (2020). Nudging consumers for relevant data using Free JAR profiling: An  
204 application to product development. *Food Quality and Preference*, 79, 103751.
- 205 Mohammad, S. M. (2016). Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual  
206 States from Text. *Emotion Measurement*, 201–237.
- 207 Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and  
208 applications. *Knowledge-Based Systems*, 89, 14–46.

- 209 Tubishat, M., Idris, N., & Abushariah, M. A. M. (2018). Implicit aspect extraction in sentiment  
210 analysis: Review, taxonomy, oppportunities, and open challenges. *Information Processing and*  
211 *Management, 54*(4), 545–563.
- 212 Vidal, L., Ares, G., Machín, L., & Jaeger, S. R. (2015). Using Twitter data for food-related consumer  
213 research: A case study on “what people say when tweeting about different eating situations.”  
214 *Food Quality and Preference, 45*, 58–69.
- 215 What is the Text Analytics API? (2019). Retrieved November 6, 2019, from  
216 <https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/overview>
- 217 Yoon, S., Parsons, F., Sundquist, K., Julian, J., Schwartz, J. E., Burg, M. M., ... Diaz, K. M. (2017).  
218 Comparison of Different Algorithms for Sentiment Analysis: Psychological Stress Notes. *Studies*  
219 *in Health Technology and Informatics, 245*, 1292.
- 220