A Multimodal Symphony: Integrating Taste and Sound through Generative AI

Matteo Spanio Massimiliano Zampini Antonio Rodà
Franco Pierucci

2025-02-01

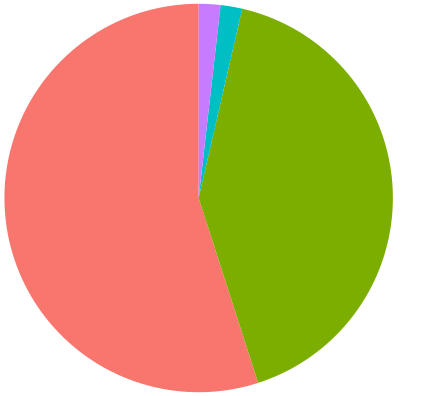
In recent decades, neuroscientific and psychological research by Spence, Wang, Zampini and others has traced direct relationships between taste and auditory perceptions. This article explores multimodal generative models capable of converting taste information into music, building on this foundational research. We provide a brief review of the state of the art in this field, highlighting key findings and methodologies. Additionally, we present an experiment in which we fine-tuned a Large Language Model (LLM) to generate music based on detailed taste descriptions provided for each musical piece. The results are promising: the fine-tuned model produces music that more coherently reflects the input taste descriptions compared to the non-fine-tuned model. This study represents a significant step towards understanding and developing embodied interactions between AI, sound, and taste, opening new possibilities in the field of generative AI.

Demographic Analysis

The data used for this study were collected through an online survey via PsyToolkit's web platform [1]. Inclusion criteria were that participants had to be over eighteen years old and have access to a device capable of playing audio files.

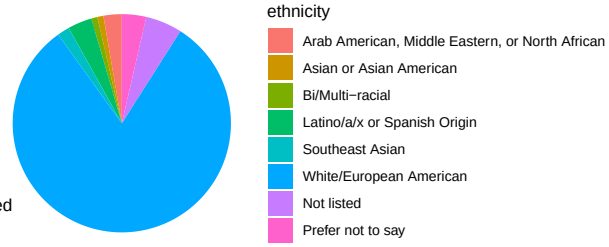
Of 141 people reached, only 111 completed the questionnaire (61 males, 46 females, 2 other, 2 prefer not to say, see Figure 1a), null or partial answers were considered as withdrawal from the questionnaire, therefore only complete answers were taken into consideration for the following analysis.

Overall the reached population has a mean age of 32 years, whit a maximum of 75 and a minimum of 19. The mean time spent on the survey was 9 minutes, with a standard deviation of 3.3. Along with age, gender and execution time also data about musical and eating experience have been collected: Figure 1b displays the ethnicity distribution of the population,



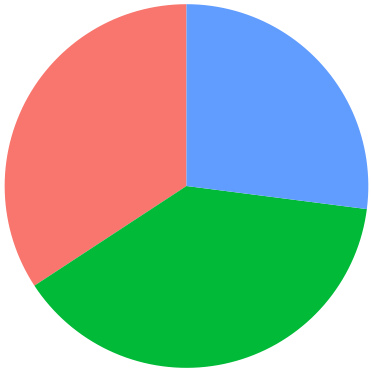
Source: Article Notebook

(a) Gender distribution of participants.



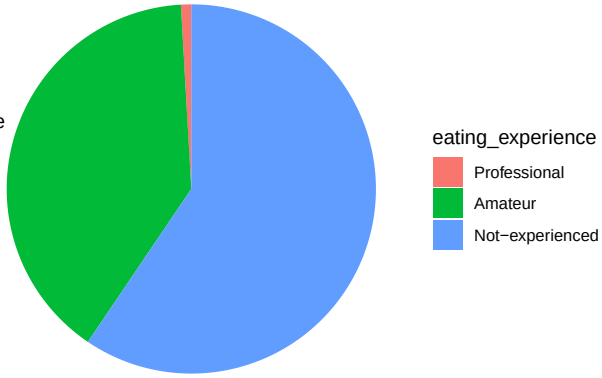
Source: Article Notebook

(b) Distribution of participants' ethnic backgrounds.



Source: Article Notebook

(c) Distribution of participants' hearing experiences.



Source: Article Notebook

(d) Distribution of participants' eating experiences.

Figure 1: Demographic characteristics of the study's participants.

the majority of participants recognize themself as *White/European American*, participants have an almost equally distributed experience in listening to music (see Figure 1c), while just one participant recognized himself as an experienced eater and the major part of the sample population declared to be not-experienced in tasting food (Figure 1d).

Model Preference Analysis

The first task in the survey involved participants listening to two audio clips, each corresponding to a taste description chosen randomly from *sweet*, *sour*, *bitter*, and *salty*. The goal was to determine which audio sample best matched the given taste description. The two clips were generated by different versions of the MusicGEN model [2]: a fine-tuned version and the original [1], base model, released by Meta. Participants were asked to express their preference for the first or second clip by moving a slider ranging from 0 to 10, where 0 indicated a strong preference for the first clip and 10 indicated a strong preference for the second, Figure 2 shows the survey's first question interface.

To ensure randomization and avoid any bias, the taste descriptions and audio clips were presented in a random order. In the analysis the scores are normalized as follows: scores from 0 to 4 are interpreted as a preference for the base model, scores between 6 and 10 indicate a preference for the fine-tuned model, and scores of 5 are treated as neutral.

The underlying research question guiding this task was to assess if the fine tuned model output better matches taste descriptions than the sounds generated by the base model.

Data Visualization

The distribution of scores across all participants is presented in Figure 3a. The histogram and density plot show the spread of scores, allowing us to visually assess the preference for one model over the other. The base model and fine-tuned model preferences are expected to manifest as peaks around the lower and higher end of the score range, respectively.

Figure 3b goes further by breaking down the preferences based on the taste category described in the audio sample. Each taste category is visualized using boxplots, where the median score for each taste can be assessed. This enables us to examine whether the model preference varies depending on the taste label, with the red dashed line at a score of 5 acting as the neutral threshold.

[1] A full description of the model and its finetuning process is available at the publication related to this analysis.

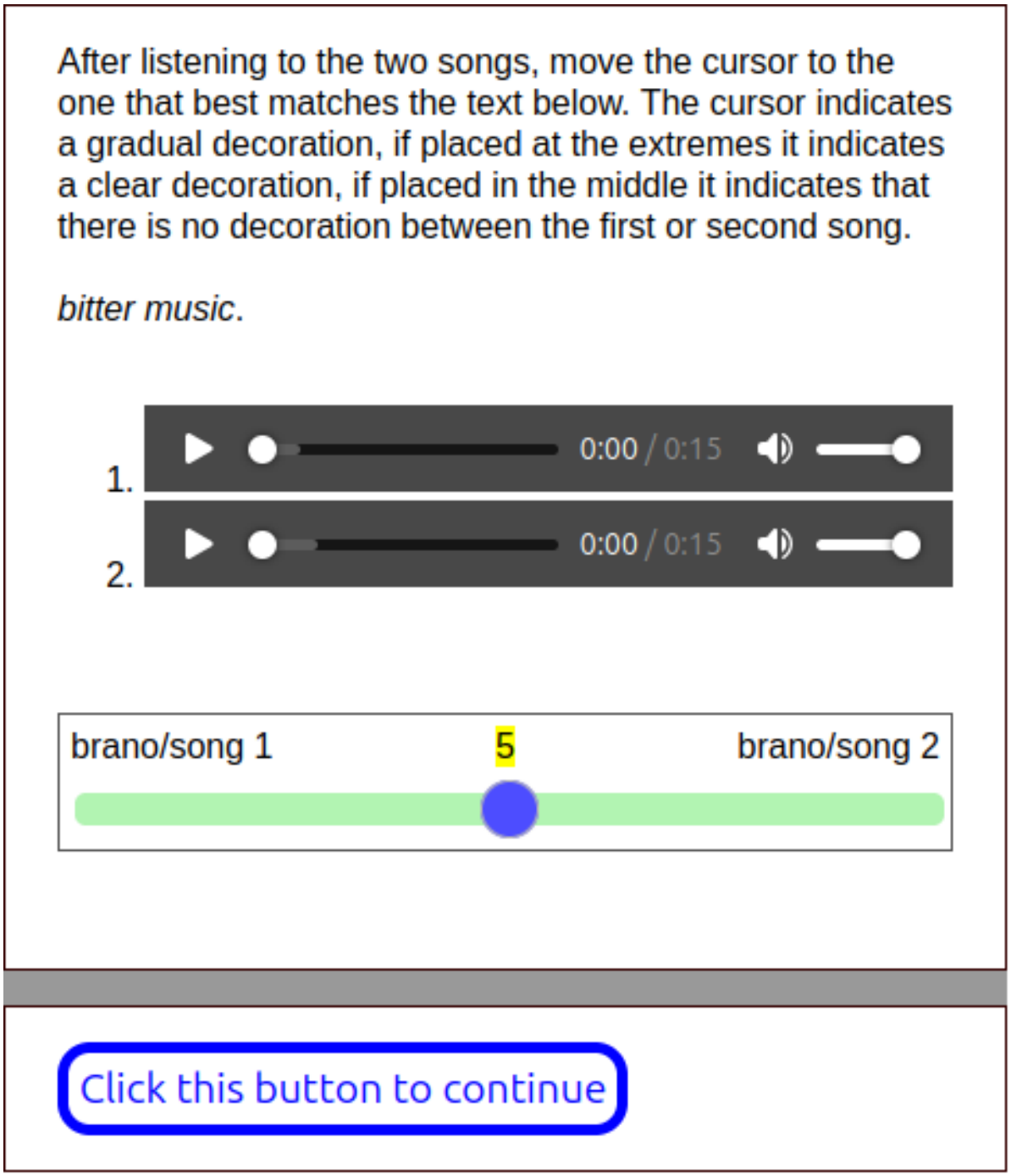
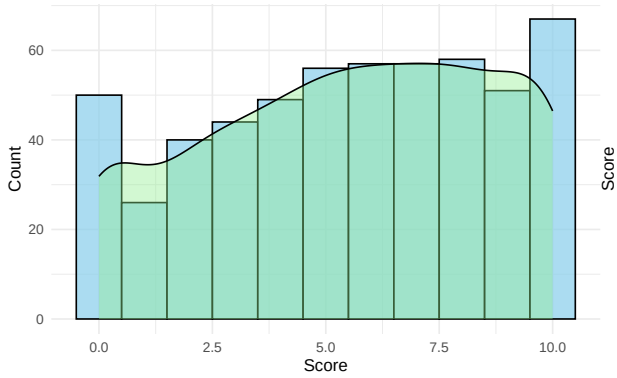


Figure 2: Screenshot of the survey's first task interface.

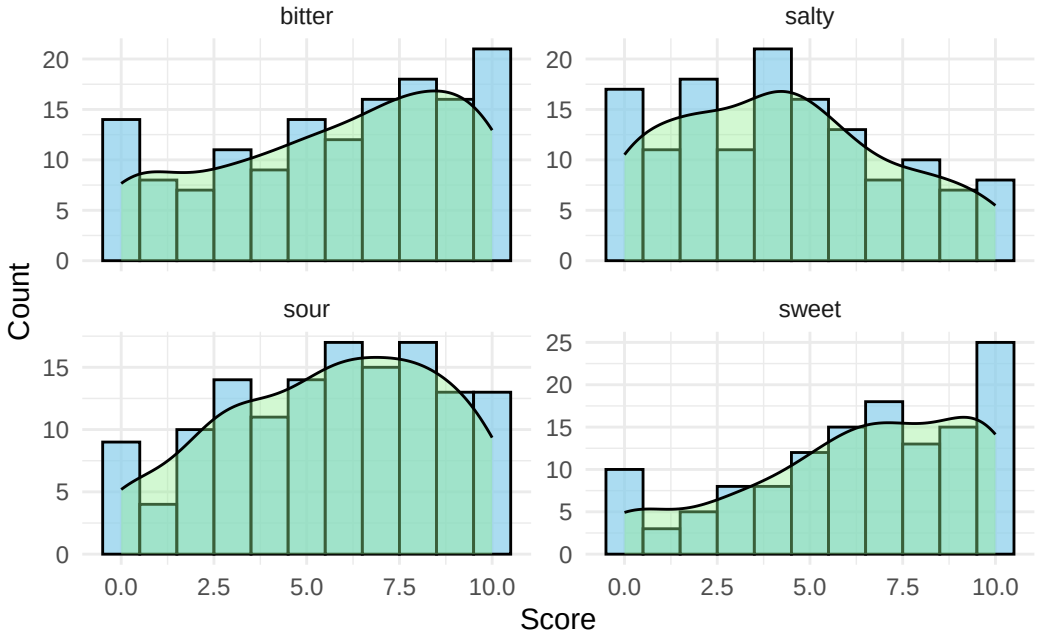


Source: Article Notebook

(a) Overall evalutation of the models.

Source: Article Notebook

(b) Score boxplot by taste category.



Source: Article Notebook

(c) Score distribution by taste category.

Figure 3: Score distribution between the two models.

Hypothesis Testing

Next, we assess whether the average score for the audio samples significantly differs from a neutral score of 5, under the assumption that the preference for the fine-tuned model should be greater than this threshold. To do this, we need to verify whether the data follows a normal distribution. In order to assess normality of data we applied both visual and computational methods, then firstly a Q-Q plot was generated to visually inspect the normality of the score distribution, see Figure 4. The resulting plot shows deviations from the expected straight line, indicating that the scores do not follow a normal distribution.

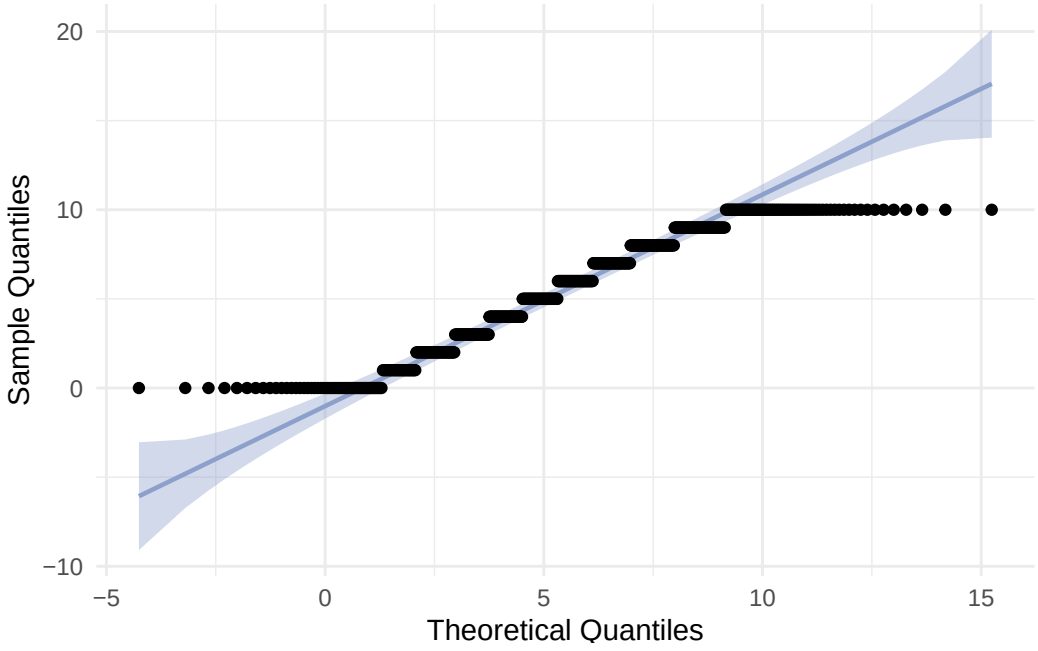


Figure 4: Q-Q plot to assess the normality distribution of the collected data.

Source: Article Notebook

Source: Article Notebook

In addition the Shapiro-Wilk test confirmed that the data significantly deviate from a normal distribution (with a resulting p-value equals to 2.516×10^{-14}). Therefore, we decide to apply the non-parametric Wilcoxon signed-rank test to see if the models preference score expressed by participants has a mean major than the null preference (score equals to five), more formally we are testing the hypothesis:

$$H_0 : \mu = 5 \quad H_1 : \mu > 5$$

where H_0 means that there is no preference between the two models while H_1 means that the fine-tuned model is preferred over the other one with a 95 confidence interval.

Source: Article Notebook

The result of the Wilcoxon test shows a p-value of 1.498×10^{-4}, which is less than 0.05, indicating that we can reject the null hypothesis and conclude that the median score is indeed significantly greater than 5. This supports the hypothesis that the participants prefer the fine-tuned model overall.

Post-Hoc Analysis by Taste

While the Wilcoxon test confirms that the overall preference goes to the fine-tuned model the boxplots reveal a variation of the score by taste, to confirm the variation we perform separate Wilcoxon tests for each taste group (*sweet*, *sour*, *bitter*, *salty*). We use a Bonferroni correction to adjust for the multiple comparisons and control the family-wise error rate. The results of the post-hoc tests are shown below, in Table 1.

Table 1: Results of the Wilcoxon test performed on data filtered by taste.

	p.value	adjusted.p.value
bitter	0.0034381	0.0137523
salty	0.9969070	1.0000000
sour	0.0076040	0.0304159
sweet	0.0000043	0.0000171

Source: Article Notebook

The analysis reveals that the fine-tuned model was significantly preferred for the sweet taste category, with an adjusted p-value of 1.7060421×10^{-5}, well below the conventional threshold of 0.05. This suggests a strong alignment between the musical outputs and participants' expectations of sweetness. Conversely, the bitter and sour categories also exhibited significant preferences, with adjusted p-values of 0.0137523 and 0.0304159, respectively. However, these results, while statistically significant, indicate a less robust preference compared to the sweet category. Notably, the salty taste group did not demonstrate a significant preference for the fine-tuned model, as indicated by an adjusted p-value near to 1. This lack of significance suggests that the model's performance may not align with participants' expectations for salty flavors, warranting further investigation into the underlying factors influencing this outcome.

Source: Article Notebook

Since the finetuned model did not show to perform well on the salty group, we performed a Wilcoxon test to test if its mean is significally lower than the tie situation ($H_0 : \mu_{\text{salty}} =$

$5, H_1 : \mu_{\text{salty}} < 5$) the result is that the first model is statistically preferred over the finetuned variant according to the Wilcoxon test with a p-value equal to 0.0031167 without Bonferroni correction[2].

Recognisability of Tastes

In the second task of the survey, participants were asked to listen to musical pieces generated exclusively by the fine-tuned model to better investigate the intrinsic qualities carried by the generated music. Following each listening session, participants were required to quantify the flavors they perceived in the music using a graduated scale, ranging from 1 to 5 (where 1 means *not at all* and 5 means *very much*), for each of the four primary taste categories: salty, sweet, bitter, and sour. Unlike the first task, there were no imposed labels for specific flavors, allowing participants the freedom to associate values with each taste based on their personal interpretations of the musical experience. Additionally, to enrich the assessment, participants had to evaluate their emotional responses to the music by rating various non-gustatory parameters, including happiness, sadness, anger, disgust, fear, surprise, hot, and cold. Figure 5 displays the web interface used to collect participants' responses.

The underlying research questions guiding this task were:

1. Can the music generated by the model induce sensory-gustatory responses?
2. What correlations exist between music and taste?
3. Which emotions mediate these sensory responses?

ANOVA test

To address the first research question, we performed an Analysis of Variance (ANOVA) to evaluate whether the participants' ratings of stimuli varied according to distinct stimulus characteristics. The dependent variable was the value assigned by participants to each stimulus, while the independent variables included stimulus-related factors. The dataset was filtered to include only participants identifying as *Male* or *Female*, excluding participants classified as *Professional Eaters* due to insufficient representation of this category. This preprocessing step ensured robust and meaningful comparisons between groups.

Table 2: Results of the ANOVA test.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
prompt	3	29.402	9.801	8.739	0.000
adjective	11	188.478	17.134	15.279	0.000

[2]The Bonferroni correction has not been applied due to the non indepenent nature of the test, in fact the test has been performed after the results of the previous Wilcoxon test, which, instead, was testing independently 4 groups.

Table 2: Results of the ANOVA test.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
hearing_experience	2	37.299	18.650	16.630	0.000
eating_experience	1	0.711	0.711	0.634	0.426
sex	1	0.069	0.069	0.061	0.804
prompt:adjective	33	214.757	6.508	5.803	0.000
Residuals	3764	4221.057	1.121	NA	NA

Source: Article Notebook

The results of the ANOVA are summarized in Table 2, which presents the degrees of freedom (Df), sum of squares (Sum Sq), mean squares (Mean Sq), F-statistics, and p-values for each factor and interaction.

Prior to interpreting the results, the homoskedasticity assumption was assessed by examining the residuals. A Shapiro-Wilk test indicated evidence of heteroskedasticity ($p < 0.05$). Despite this violation, the ANOVA analysis proceeded, following recommendations from prior research [3], [4], [5] suggesting that ANOVA is robust to deviations from normality under moderate violations, particularly with large sample sizes such as the one in this study. The results show that different prompts and adjectives lead to significantly different adjectives quantified by participants, similarly the adjectives used influence the participants' feelings. Furthermore the significant interaction effect implies that the effect of one variable depends on the level of the other. In other words, the way a prompt influences feelings may vary depending on the adjective used. This result highlights that the participants to the survey deliberately operated consistent choices while evaluating the stimuli.

Post-Hoc Analysis

To further explore the results of the ANOVA, we conducted Tukey's Honest Significant Difference (HSD) test. This post-hoc analysis is particularly useful for identifying which specific group means are significantly different from each other after finding a significant overall effect in the ANOVA. Given that our analysis revealed significant main effects for prompt, adjective, their interaction and the hearing experience, it is essential to determine the nature of these differences.

The Tukey test compares all possible pairs of group means while controlling for the family-wise error rate, thus providing a robust method for multiple comparisons. This is crucial in our context, as we aim to understand how different prompts, adjectives and hearing experience levels influence participants' evaluations of the stimuli. Upon executing the Tukey test, we examined the adjusted p-values for each comparison. The results indicated several significant

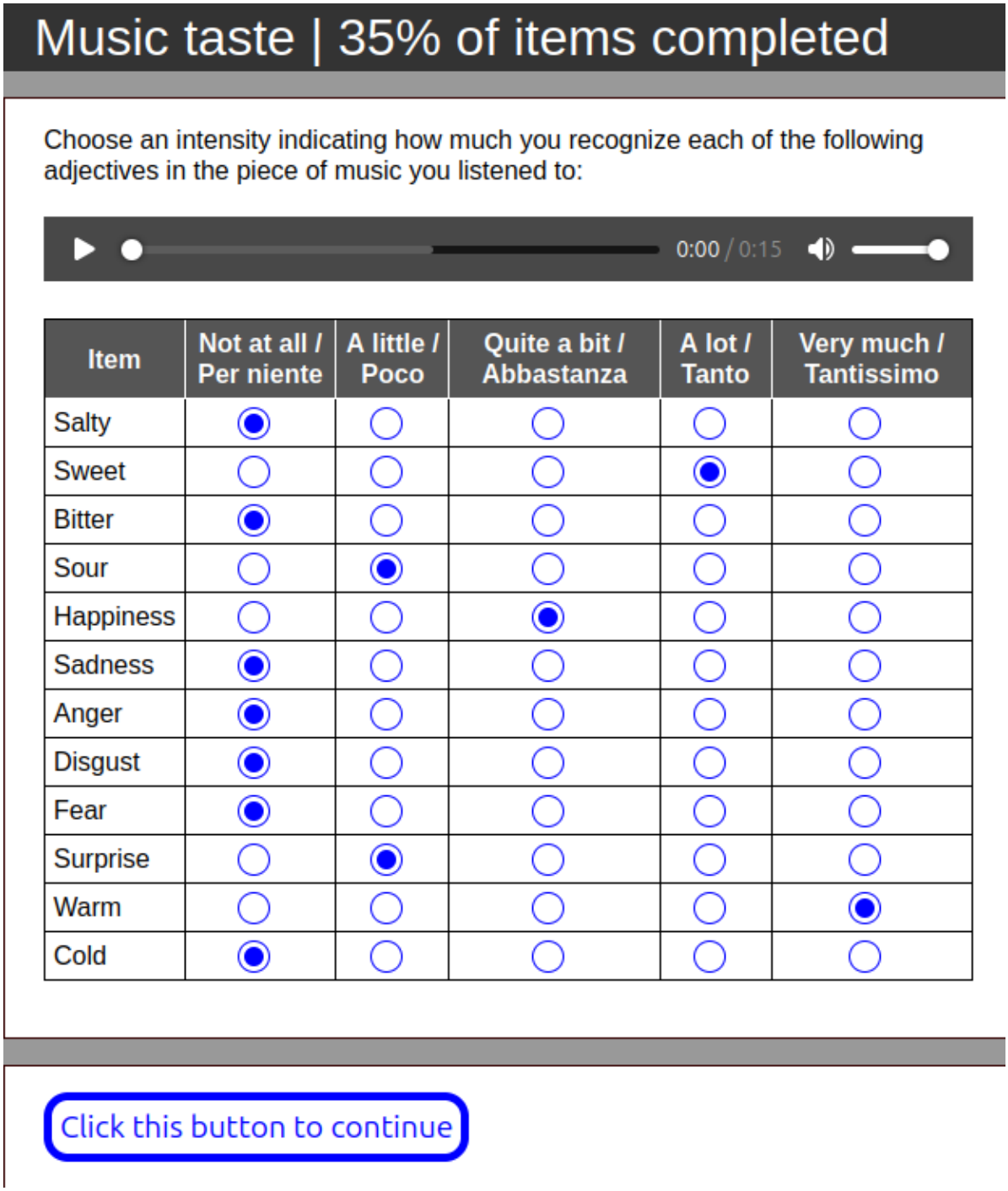


Figure 5: Screenshot of the survey's second task interface.

differences between specific combinations of prompts and adjectives, as summarized in the table below:

Source: Article Notebook

Table 3: Tukey test between different prompts with a p-value lower 0.05.

	diff	lwr	upr	p adj
sour-bitter	0.1539030	0.0274663	0.2803397	0.0095792
sour-salty	0.1942878	0.0661296	0.3224460	0.0005747
sweet-sour	-0.1932957	-0.3214540	-0.0651375	0.0006229

Source: Article Notebook

Table 4: Tukey test between different adjectives with a p-value lower 0.05.

	diff	lwr	upr	p adj
bitter-anger	0.4204204	0.1439129	0.6969280	0.0000443
cold-anger	0.4774775	0.2009699	0.7539850	0.0000012
hot-anger	0.5315315	0.2550240	0.8080391	0.0000000
sad-anger	0.5765766	0.3000690	0.8530841	0.0000000
sweet-anger	0.3723724	0.0958648	0.6488799	0.0006699
disgust-bitter	-0.6336336	-0.9101412	-0.3571261	0.0000000
happy-bitter	-0.3093093	-0.5858169	-0.0328018	0.0136899
surprise-bitter	-0.2852853	-0.5617928	-0.0087777	0.0360645
disgust-cold	-0.6906907	-0.9671982	-0.4141831	0.0000000
happy-cold	-0.3663664	-0.6428739	-0.0898588	0.0009180
sour-cold	-0.2852853	-0.5617928	-0.0087777	0.0360645
surprise-cold	-0.3423423	-0.6188499	-0.0658348	0.0030580
fear-disgust	0.4324324	0.1559249	0.7089400	0.0000213
happy-disgust	0.3243243	0.0478168	0.6008319	0.0070884
hot-disgust	0.7447447	0.4682372	1.0212523	0.0000000
sad-disgust	0.7897898	0.5132822	1.0662973	0.0000000
salty-disgust	0.4684685	0.1919609	0.7449760	0.0000021
sour-disgust	0.4054054	0.1288979	0.6819130	0.0001074
surprise-disgust	0.3483483	0.0718408	0.6248559	0.0022833
sweet-disgust	0.5855856	0.3090780	0.8620931	0.0000000
hot-fear	0.3123123	0.0358048	0.5888199	0.0120394
sad-fear	0.3573574	0.0808498	0.6338649	0.0014571
hot-happy	0.4204204	0.1439129	0.6969280	0.0000443
sad-happy	0.4654655	0.1889579	0.7419730	0.0000026
sour-hot	-0.3393393	-0.6158469	-0.0628318	0.0035312

	diff	lwr	upr	p adj
surprise-hot	-0.3963964	-0.6729040	-0.1198888	0.0001798
salty-sad	-0.3213213	-0.5978289	-0.0448138	0.0081111
sour-sad	-0.3843844	-0.6608919	-0.1078768	0.0003509
surprise-sad	-0.4414414	-0.7179490	-0.1649339	0.0000122

Source: Article Notebook

Table 5: Tukey test between different hearing experience groups with a p-value lower 0.05.

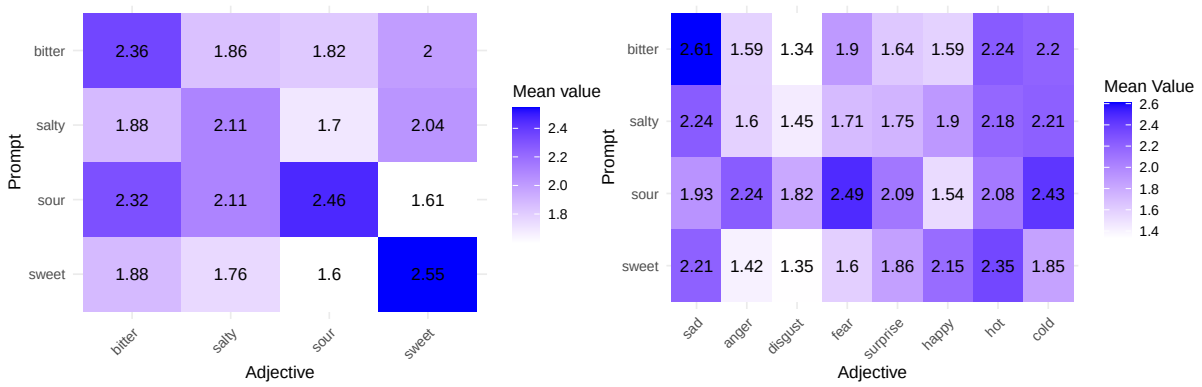
	diff	lwr	upr	p adj
Amateur-Professional	0.2268160	0.1318865	0.3217454	0.0000001
Not-experienced-Amateur	-0.1684003	-0.2698267	-0.0669739	0.0002967

Source: Article Notebook

These significant comparisons illuminate the subtleties in participants' responses to different stimuli. Certain prompt-adjective combinations elicited stronger emotional reactions than others, indicating that the interaction between prompts and adjectives significantly shapes participants' perceptions. Notably, some combinations yielded adjusted p-values below the conventional threshold of 0.05, signifying statistically significant differences. This finding reinforces the ANOVA results, confirming that the presentation of prompts and adjectives can meaningfully impact emotional responses.

Interaction between *prompt* and *adjective*

The ANOVA analysis evidenced also a significant interaction between prompt and adjectives used to evaluate the sounds, as we know, the design of the experiment fixed the prompt before generating the audio files, therefore adjectives has to be intended as dependent variables while the prompts are independent; in other words, participants assigned different values to the adjectives to the sounds on the basis of their generation prompt. This interaction can be seen in Figure 6. In particular Figure 6a shows the mean value assigned to each taste adjective by their prompt, we can clearly seen the major diagonal emerge by the 4×4 matrix, this means that, the mean value assigned to the adjective that matches the prompt of each sound is the highest. The rest of the interaction between adjectives and prompt can be seen in Figure 6b , a deeper analysis of emotional aspect assigned to the sounds is presented in the next section.



Source: Article Notebook

(a) Heatmap of perceived taste in correspondence of the intended one.

Source: Article Notebook

(b) Heatmap of perceived emotional response in correspondence of the suggested taste.

Figure 6: Heatmaps

Factorial analysis

Source: Article Notebook

To explore the underlying relationships between sensory qualities and emotional states, we conducted a factor analysis on the standardized data, excluding the 'taste' column. The initial step involved calculating the correlation matrix, which revealed notable relationships among the adjectives. Specifically, we observed that negative emotions were positively correlated, while the pair happy-sad exhibited a negative correlation. Furthermore, sweetness demonstrated a strong correlation with happiness and warmth, whereas bitterness was associated with anger and fear. Sourness, on the other hand, was evidently correlated with disgust and fear. Other variables did not show strong correlations at first glance, prompting us to proceed with the factor analysis.

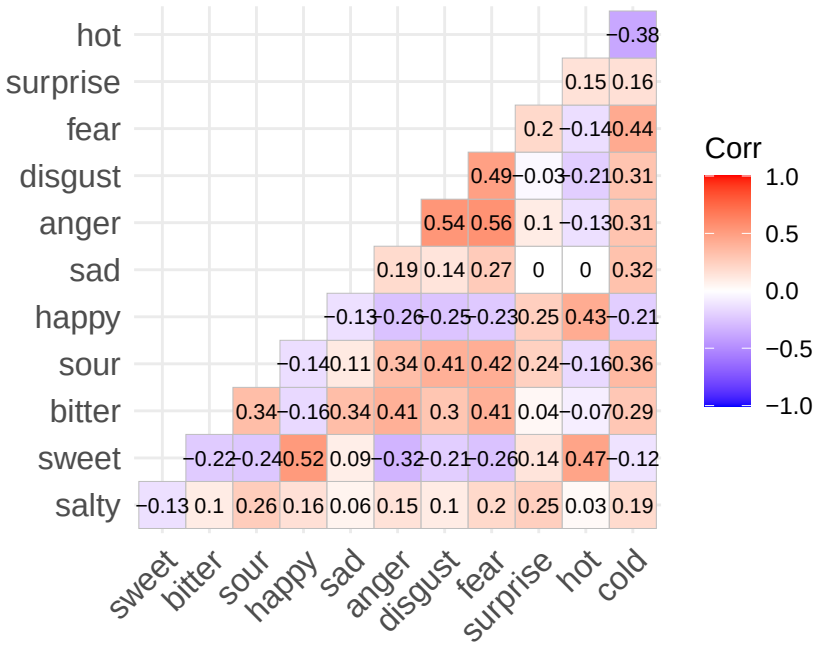


Figure 7: Correlation matrix

Source: Article Notebook

The correlation matrix is illustrated in Figure 7, showcasing these relationships clearly.

```
Parallel analysis suggests that the number of factors =  4  and the number of components =  1
```

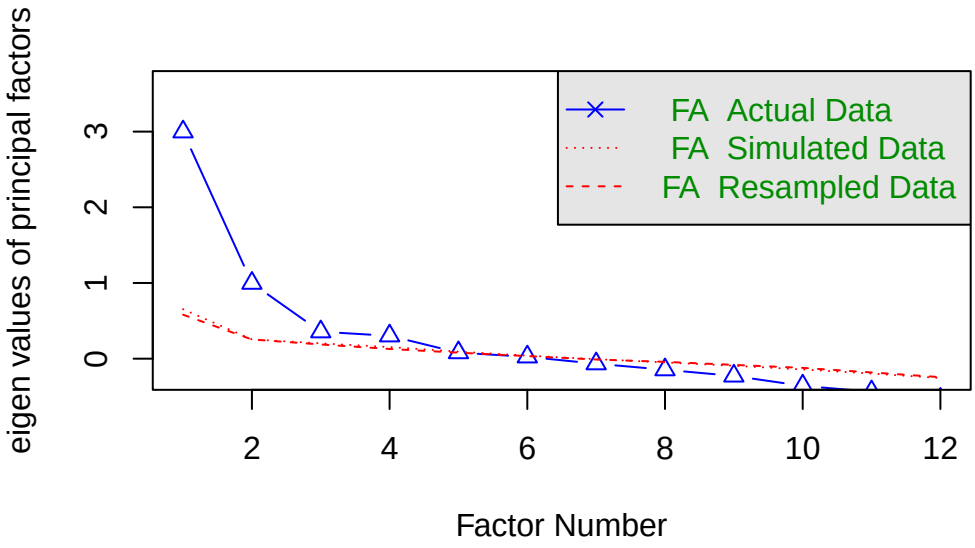


Figure 8: Parallel Analysis Scree Plots

Source: Article Notebook

To determine the optimal number of factors for our analysis, we employed parallel analysis, which indicated an optimal number of 4 factors. This estimation serves as a foundation for our subsequent factor analysis. Following this, we executed the factor analysis using the identified number of factors, applying an oblique rotation (oblimin) to allow for potential correlations among the factors. The results of the factor analysis, including the factor loadings, are presented in the output. The factor loadings indicate how strongly each variable contributes to the identified factors, providing insights into the underlying structure of the data.

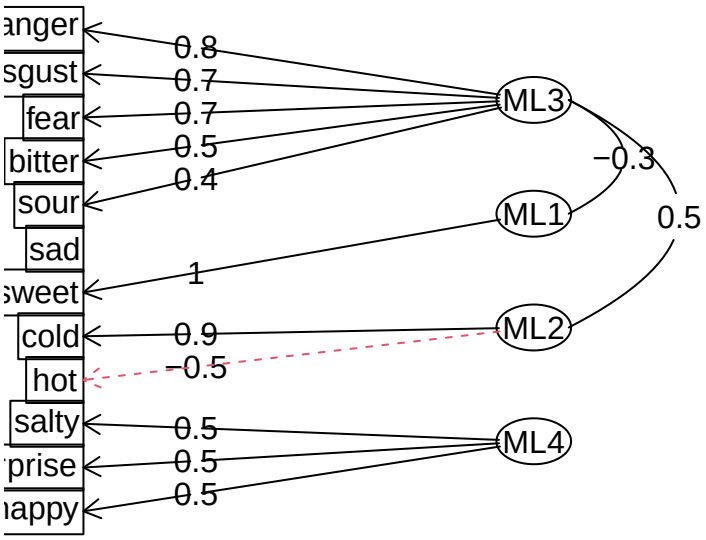


Figure 9: Factor analysis graph.

Source: Article Notebook

```
Loadings:
         ML3    ML1    ML2    ML4
salty          -0.231  0.111  0.535
sweet           0.992
bitter    0.502
sour      0.385 -0.128  0.178  0.226
happy    -0.197  0.302 -0.132  0.492
sad       0.292  0.259  0.236
anger     0.779
disgust   0.694                -0.133
fear      0.662         0.133  0.113
surprise                0.120  0.526
hot       0.140  0.361 -0.458  0.267
cold                    0.882

                 ML3   ML1   ML2   ML4
SS loadings    2.083 1.360 1.146 0.971
Proportion Var 0.174 0.113 0.096 0.081
Cumulative Var 0.174 0.287 0.382 0.463
```

Source: Article Notebook

To visualize the relationships among the factors, we generated biplots for various factor combinations. The biplots, shown in the subsequent figures, illustrate the distribution of variables across the identified factors, highlighting the clustering of adjectives associated with similar emotional states.

Prompt	Fattore 1	Fattore 2	Fattore 3	Fattore 4
bitter	$\mu = -0.01$, $\sigma = 0.82$	$\mu = -0.06$, $\sigma = 0.99$	$\mu = 0.04$, $\sigma = 0.91$	$\mu = -0.16$, $\sigma = 0.69$
salty	$\mu = -0.13$, $\sigma = 0.87$	$\mu = -0.03$, $\sigma = 0.98$	$\mu = 0.01$, $\sigma = 0.94$	$\mu = 0.02$, $\sigma = 0.94$
sour	$\mu = 0.50$, $\sigma = 1.02$	$\mu = -0.39$, $\sigma = 0.79$	$\mu = 0.23$, $\sigma = 0.93$	$\mu = 0.11$, $\sigma = 0.74$
sweet	$\mu = -0.30$, $\sigma = 0.74$	$\mu = 0.43$, $\sigma = 1.04$	$\mu = -0.26$, $\sigma = 0.82$	$\mu = 0.05$, $\sigma = 0.80$

Source: Article Notebook

Lastly, we performed a multi-factor analysis to further explore the dimensions of negativity and positivity within the data. The results of this analysis are depicted in the multi-factor diagram, which categorizes the factors into two overarching themes: Negativity and Positivity.

References

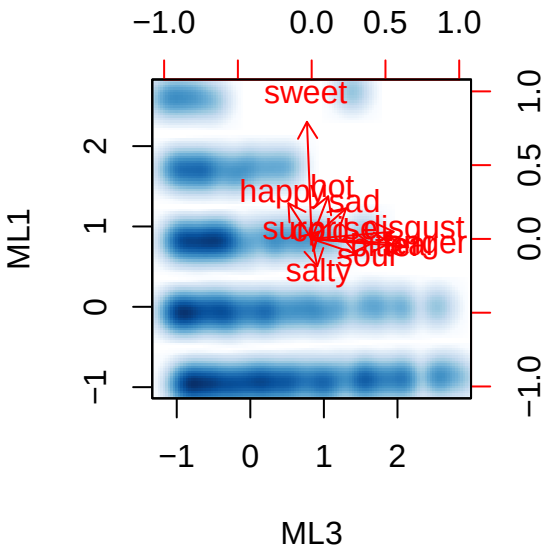
[1] G. Stoet, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017, doi: 10.1177/0098628316677643.

[2] J. Copet *et al.*, "Simple and controllable music generation," in *Proceedings of the 37th international conference on neural information processing systems*, in NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[3] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance," *Review of Educational Research*, vol. 42, no. 3, pp. 237–288, 1972, doi: 10.3102/00346543042003237.

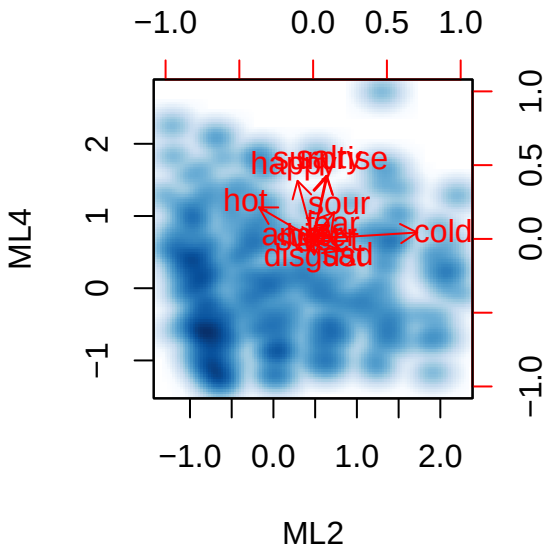
[4] M. R. Harwell, E. N. Rubinstein, W. S. Hayes, and C. C. Olds, "Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases," *Journal of Educational Statistics*, vol. 17, no. 4, pp. 315–339, 1992, Accessed: Jan. 21, 2025. [Online]. Available: http://www.jstor.org/stable/1165127

[5] L. M. Lix, J. C. Keselman, and H. J. Keselman, "Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance "f" test," *Review of Educational Research*, vol. 66, no. 4, pp. 579–619, 1996, Accessed: Jan. 21, 2025. [Online]. Available: http://www.jstor.org/stable/1170654



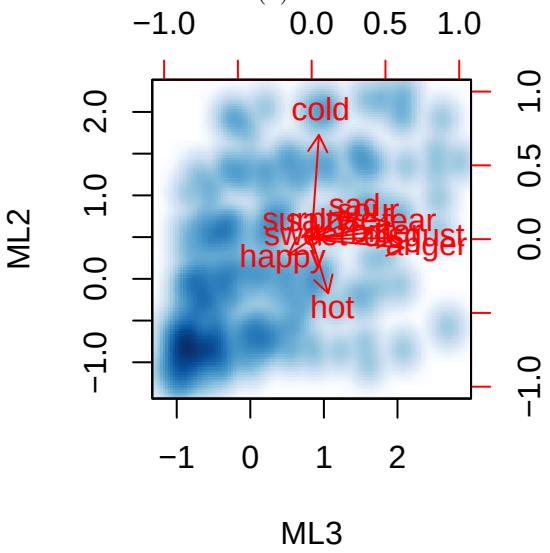
Source: Article Notebook

(a)



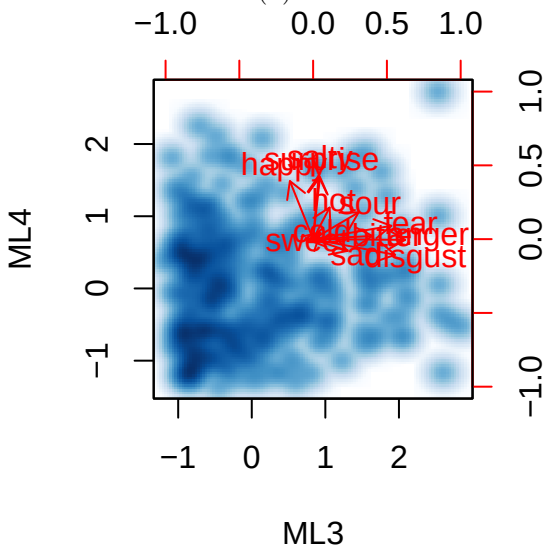
Source: Article Notebook

(b)



Source: Article Notebook

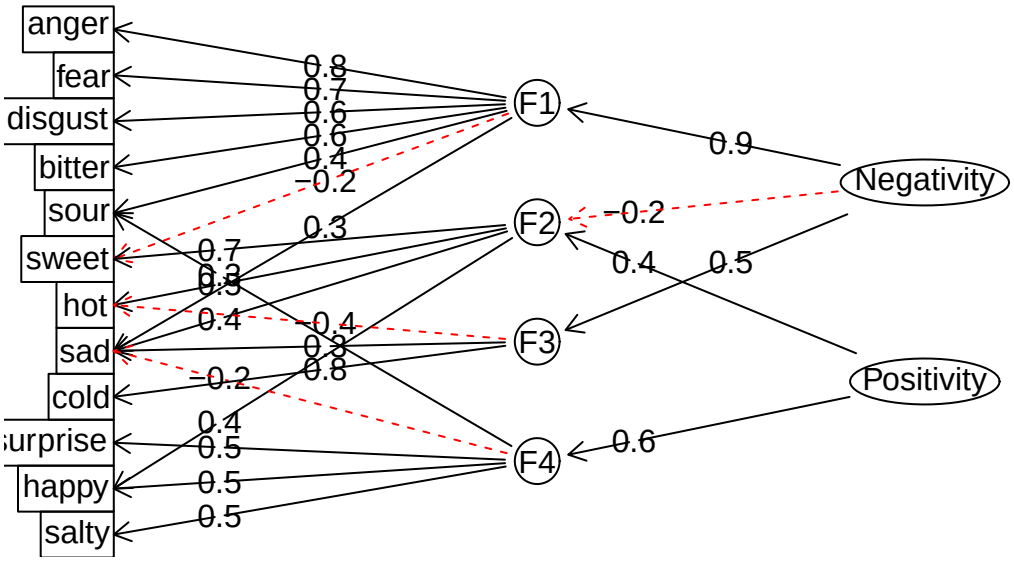
(c)



Source: Article Notebook

(d)

Figure 10: Loadings biplot over the four factors.



Source: Article Notebook

Figure 11: Hierarchical (multilevel) factors' structure.