Master Thesis

École Polytechnique Fédérale de Lausanne
School of Computer and Communication Sciences
Data Science Laboratory

Who likes this beer? Me or the community?

A matched observational study of beer reviews
from two aligned communities

by

Gael Lederrey
Computational Science and Engineering
Section of Mathematics

Under the supervision of Prof. Robert West, EPFL
and the expertise of Prof. Carlos Castillo, Eurecat

June 2017

After many years studying at EPFL, this is the final piece of text I am writing for my studies. And I would like to use these few words to thank everybody who helped me, believed in me and supported me. I would like to thank my professor Bob West for giving me the opportunity to do this amazing project with him. It was an honor to work with him. I would also like to thank all my friends in Physics and Mathematics who made these years of Bachelor and Master much easier than they really were. And finally, I would like to thank my parents. Without them, I would not be here today for many reasons. Thank you!

Abstract

Online rating systems are a standard tool used to share, discuss and sometimes sell products. The study of how the users interact with others inside these communities is fundamental to the better understanding of our society. In most of the literature, only one community is analyzed at a time. This thesis uses the data from two websites on beer reviews: BeerAdvocate and RateBeer. These data are matched to create a subset of products and users present in both communities.

This thesis uses the newly matched datasets to study different social effects happening inside these communities. The results show that users have the tendency to follow the crowd's judgment and the first ratings influence this judgment. The use of Natural Language Processing demonstrates that the communities tend to build their own vocabulary. The users who rate beers on both websites have the tendency to copy-paste the text of their reviews, but they give different ratings in both communities. A Randon Forest model is finally used to predict the reviews between both websites.

This collection of results demonstrates the power of using a subset of products and users present in two communities, and the use of analyses at the individual level. These results are used to give different advice to understand and build better online rating systems.

Contents

Fill with mingled cream and amber,
I will drain that glass again.
Such hilarious visions clamber
Through the chambers of my brain.
Quaintest thoughts — queerest fancies,
Come to life and fade away:
What care I how time advances?
I am drinking ale today.

— Edgar Allan Poe, *Lines on Ale*
July 1848

Introduction

Beer is a worldwide known beverage. Nearly all human being have tasted beer, being either just a sip or the daily beers at Sattelite, the bar at EPFL. It is the drink of the poor and the rich. Every human being who drinks beer on a regular basis has some preferences between the different types of beers such as Indian Pale Ale, Stout, and Lager. Beer is a drink that brings people closer and further away at the same time. Everybody wants to give his opinion on this beverage. Even some famous people have quoted the beer or written poem on it, see the previous page for a poem of Edgar Allan Poe. Since we are in the era of the Web 2.0, we do not only discuss beers at the table of a pub; we do it on the Internet too. In this thesis, we will talk about two websites on beer reviews: RateBeer[1] and BeerAdvocate[2]. RateBeer has been created in May 2000 and is run by John Tucker. BeerAdvocate has been set up a bit earlier, in 1996, by Jason and Todd Alström. In their About page, these two websites are pushing their forum and community forward. The purpose of these two websites is not to sell beers. It is to share, discuss and trade them. Since their community is very active and these websites are doing anything to keep their community with them while growing, it is fascinating to study them.

This thesis studies the social effects that can happen in different kind of communities. One highly studied effect is the *herding effect*. Everybody knows how it works, even if he or she is not part of a scientific community. This effect describes the fact that a single human will have the tendency to follow the crowd even if it is not good for him. In this case, nothing bad can happen to the users, but the quality of a beer perceived by them could suffer from this effect. The main strength of this thesis is the fact that these two datasets are from two communities on the same subject: the beers. These communities have an overlapping subset of users. The sets of beers on both websites are also overlapping. Therefore, it is possible to perform a broad study on the beers as well as on the users. Indeed, it is feasible to match these two datasets to compare them on different aspects.

However, this thesis did not start with the study of the social effects in different communities. The inspiration for this project came from the paper of Wang and Wang [1]. This article develops a Machine Learning model for predicting the future ratings of an Amazon product. This model takes into account two factors: the intrinsic quality of the product and the herding effects. Their model works well to predict the long-term consequences of the herding effects. The idea of this thesis was to develop a similar model applicable to both datasets (BeerAdvocate and RateBeer). However, it was not possible to apply the same kind of model. One of the first reason was that their model is based on the possibility for the users to see the histograms of stars. Therefore, they could use a multinomial distribution to model the intrinsic quality. In the case of BeerAdvocate and RateBeer, it is not possible. Firstly the rating systems are not as simple as in Amazon. Secondly, because the users do not have access to the histograms of the ratings, thus it would not be possible to justify the use of such a model. BeerAdvocate and RateBeer are displaying other information. Therefore, the different technical and visual aspects of the websites have been studied. At around the middle of the thesis, many results could already be found with the lone analysis of the data. Therefore, the idea of building such a Machine Learning model has been aborted, and the aim of the thesis changed to an observational study of the communities and the social effects inside these datasets.

[1] Link to RateBeer's website: http://ratebeer.com

[2] Link to BeerAdvocate's website: http://beeradvocate.com

This thesis concentrates on the two main strengths of these datasets: the overlapping sets of beers and users in both datasets. The analysis starts with the matched beers and tries to understand how the ratings can be influenced using one of the datasets to control for the other one. Then, the analysis continues with the matched users. There is also the possibility to control for the users who rated the same beer on both websites. The analyses on this subset of beers and users are fascinating because the only thing that differs is the community, the website. Indeed, since these users are rating the same beer on both websites, their opinion on this beer is the same. Thus, the only effects that could appear come from the differences between the two websites and the two communities.

Since this thesis is an observational study on two datasets, the reader will not find much theory. The Machine Learning models used in this thesis, as well as the Natural Language Processing algorithms, are not given in details. If the reader does not understand what a specific function or algorithm does, we strongly advise him to have a look at the documentation of the package mentioned or the theory behind the different algorithms used.

This thesis is separated into three main parts. Beforehand, a small literature review is presented in Section Literature Review. The first part is about some general and technical information. The datasets and the websites are presented in Section Framework. Then, some basic studies on the data are performed in Section Understanding the Data. Finally, the algorithm for matching the beers and the users is explained in Section Matching. The second part of this thesis is using the matched beers. First, a study is performed on the influence of the score[3] on the ratings in Section Influence of the Scores on the Ratings. Then, the predictability of the ratings on both websites is investigated in Section Predictability of the Ratings. Finally, a study on the herding effects is done in Section Herding Effects. The third and final part is mostly using the matched users and the beers they rated on both websites. It starts with a little bit of Natural Language Processing for classifying and extracting features from the text of the reviews in Section Text Classification and Features Extraction. Then, the communities and how the users belong to them is studied in Section Communities. The last part is the study of the users who rated the same beer on both websites in Section Same User, same Beer, two Websites. Finally, The results are recalled, and a conclusion is given in Section Conclusion. The reader can find supplementary figures in the Appendix that are always discussed or mentioned in the previous sections.

[3] see Subsection Websites in Section Framework.

Literature Review

In this section, a literature review based on the different aspects covered in this thesis is given. First, the Online Rating Systems are presented, then the Herding effects and finally some literature about the two datasets used in this thesis is given.

Online Rating Systems

Online rating systems can take many different forms. For example, Amazon[4] has a five-stars system, Reddit[5] has a up-down arrow system, Facebook[6] has the "like" system, and BeerAdvocate and RateBeer have a system of ratings per aspects, see subsection Websites. Many experiments have been conducted to understand the behavior of the users when rating a product or a comment. For example, the work of Glenski & Weninger [2] tries to understand the effects of upvotes and downvotes on the behavior of the users on Reddit. The paper of Zhang *et al.* [3] is using data from TripAdvisor[7] to understand the dynamics of the ratings. In their paper, using a model taking into account the reviews and the popularity of the restaurants, they show that the ratings converge towards a positive perspective. The paper of Hu *and al* [4] is using Amazon's ratings to investigate the effects of the online word-of-mouth communication. This effect is very well known in real life. However, this paper, among many others, shows the effect of this exchange in an online environment. They developed a "*Brag-and-Moan*" model to predict the Amazon's rating. This model is named that way due to the U-shape of the distribution of ratings on Amazon, see this link Amazon Echo Dot for Alexa (click on the stars to see the histogram of ratings). Finally, the paper of Mudambi & Schuff [5] uses the Amazon reviews and looks at the helpfulness of the customers' reviews. Their work treats the negative and moderate reviews, and they found that the helpfulness can differ depending on the product type, *i.e.* sometimes bad reviews are not helpful at all, at times they are.

As the reader can see, many different online rating systems have been studied in these papers. The difference between these systems makes it difficult to link them together. The users' behavior is different between all of these websites, simply because they do not visit them for the same reasons. However, with the field of data science and computational social science growing quickly, the knowledge about these rating systems is becoming more and more precise.

Herding Effects

The herding effects are one of the most studied social effects in online communities. For example, the paper of Banerjee [6] is about the theory behind the herding effects. There are also some papers on the empirical viewpoints of the herding effects, such as the paper of Leskovec *et al.* [7] and the paper of Anderson [8]. There are also papers on the social influences in a network, which are directly linked to herding effects, as in the paper of Meyers & Leskovec [9] working on Twitter[8] data. However, this thesis concentrates on online rating systems. Therefore, more example of literature on the herding effects in rating systems are given: the paper of Chevalier & Beinecke [10] on sales book published in the Journal of Marketing Research, the paper of Luca [11] on restaurants' reviews, and the paper of Zhu [12] on video games published in the

[4] Link to the Amazon's website: http://amazon.com/

[5] Link to the Reddit's website: http://reddit.com

[6] Link to the Facebook's website: http://facebook.com

[7] Link to the TripAdvisor's website: http://tripadvisor.com

[8] Link to the Twitter's website: http://twitter.com

Journal of Marketing as well. With the journals given as examples, the reader can see that the herding effects interests scientists not only on a social science point-of-view but also for marketing and business purpose. Many other examples of papers about herding effects and social influences can be found in the section "Related Work" of the paper of Glenski & Weninger [2].

BeerAdvocate and RateBeer

The datasets of BeerAdvocate and RateBeer have been scraped in 2012 by Julian McAuley and are available on SNAP[9] [13]. Many works have been done using one or both of these datasets. A non-exhaustive list of these works is given. The paper of McAuley, Leskovec, and Jurafsky [14] is using both datasets as well as others. This work developed a model to understand the texts associated with multi-aspects reviews. It can understand the text, link the sentences to the different aspects rated, and summarize a review correctly. The paper of McAuley & Leskovec [15] is also using both datasets as well as others. In this paper, they investigate the experience levels of the users and how a recommender system that takes into account this experience would work. The work of Guardia-Sebaoun, Guige, and Gallinari [16] uses both datasets and others. This paper is about adding the time in Recommender Systems using a latent method. Finally, the work of Danescu *et al.* [17] is only using the datasets of BeerAdvocate and RateBeer. This work is about users' lifecycle and linguistic changes over the years. They even use the linguistic changes of the users to predict their lifespan in these communities.

The list of papers given in the previous paragraph is not exhaustive. However, no paper has been found using both BeerAdvocate and RateBeer (or any other datasets) to extract the similar items of these websites and compare them. All of the papers in this subsection have used these datasets, and they compared their results at an aggregated level. This project takes an interest in comparing the datasets on an individual level on both the products and the users.

[9] Link to the SNAP's website: http://snap.stanford.edu

Framework

In this section, the framework for this thesis is presented. Indeed, the data are coming from two different websites: BeerAdvocate and RateBeer. There are fundamental differences in the scrapped data and the websites. The main differences are shown in this section.

Datasets

BeerAdvocate and RateBeer allow users to rate beers on five aspects: appearance, taste, palate, smell, and overall. The first four are all sensory aspects. The last one is the overall appreciation of the beer. The data were collected by Julian McAuley and are available on SNAP [13, 14, 15]. Table 1 show the size of the two datasets used in this thesis.

	#Users	**#Beers**	**#Reviews**
BeerAdvocate	33'388	66'055	1'586'614
RateBeer	29'265	112'171	2'924'163

Table 1: Summary of the datasets.
The beers and users without ratings are not taken into account.

Table 2 gives some statistics on the number of reviews per beer and user.

	Beers		**Users**	
	BeerAdvocate	RateBeer	BeerAdvocate	RateBeer
Mean	24.02	26.07	47.52	99.92
Max	3290	3696	5817	16'364
75th percentile	7	13	16	16
Median	2	4	3	3
25th percentile	1	2	1	1
At least 10 reviews	14'174	35'328	10'190	8'924

Table 2: Statistics on the number of reviews per beer and per user.

The data span from the 2nd of August 1996 to the 11th of January 2012 for BeerAdvocate and from the 12th of April 2000 to the 13th of January 2012 for RateBeer. It is interesting to see the development of these two websites over the years. The number of reviews over the years is given in Figure 1. The number of users joining every year is shown in Figure 2. It is growing every year. If a simple linear regression is used to extrapolate the number of reviews in 2017, an addition of around 2.5 million reviews for BeerAdvocate and around 3.8 million reviews for RateBeer is expected. Therefore, scraping the data in 2017 would at least double the total number of reviews. The number of users joining every year is also increasing. Even if this growth starts to flatten, around 3'000 users every year could be added at the minimum. Nowadays, this would give an increase of around 15'000 users on both websites. However, the data were not scraped during this project to get more for the analysis of the data. All analyses presented in this thesis can be applied to the new data when they will be scraped.

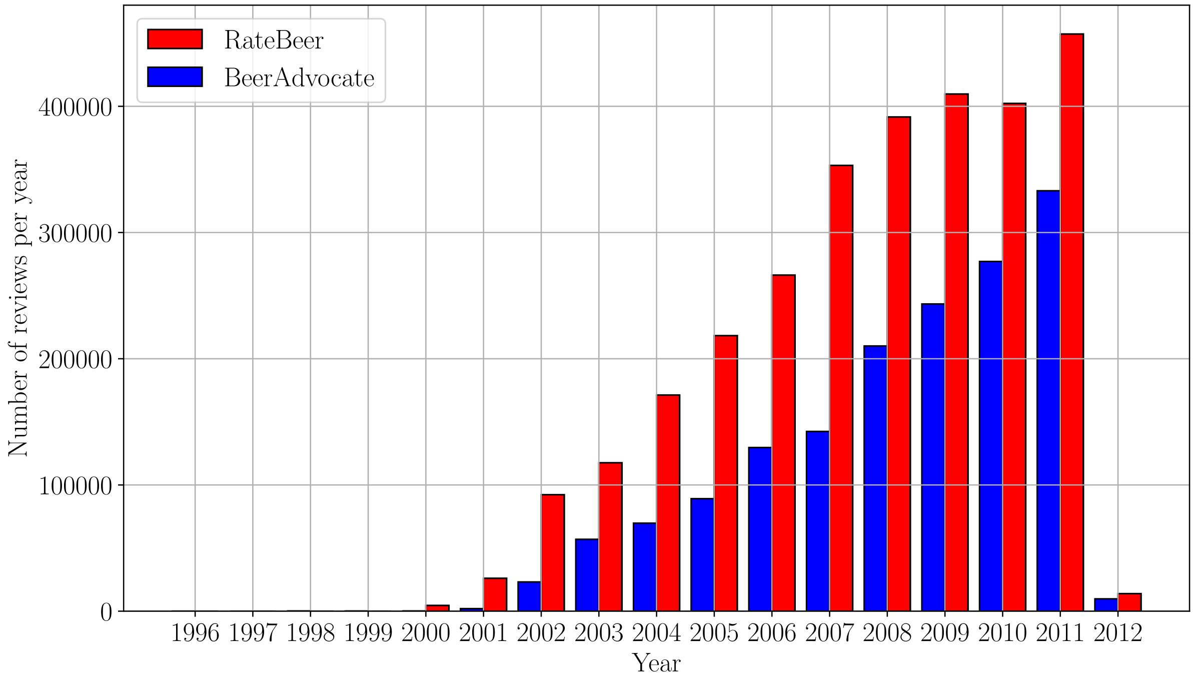


Figure 1: Number of reviews for BeerAdvocate and RateBeer per year.

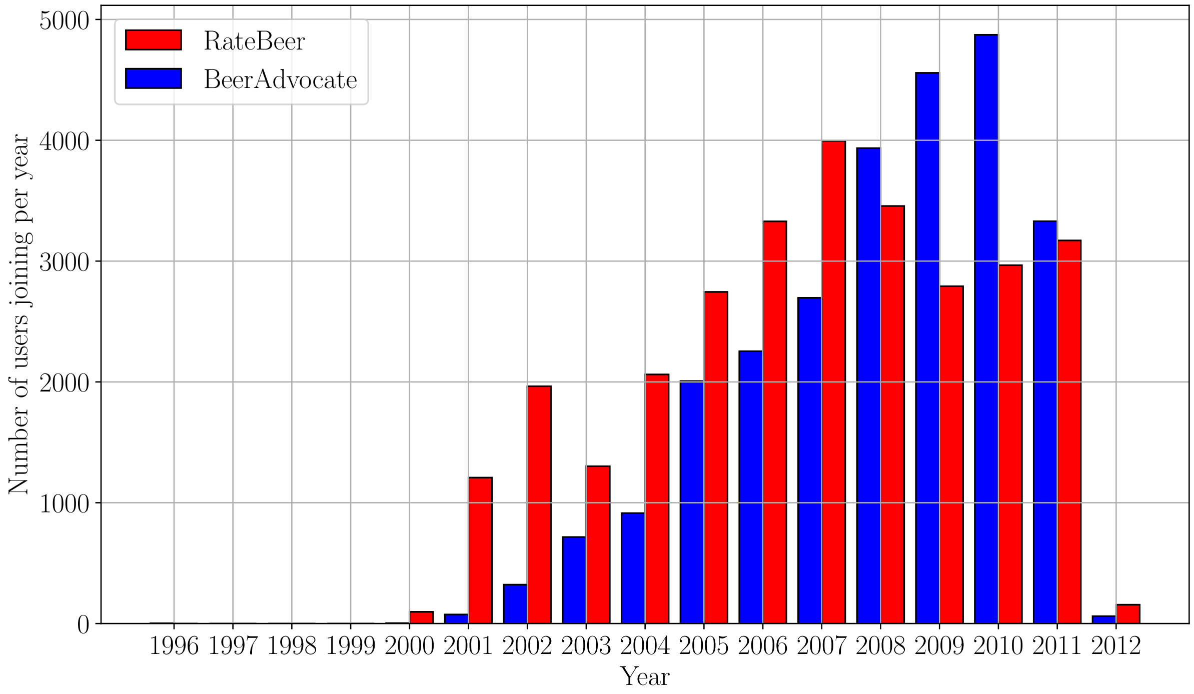


Figure 2: Number of users joining every year for BeerAdvocate and RateBeer.

Websites

These two websites have the same purpose, which is allowing users to give their opinion on a lot of different beers. However, the framework is fundamentally different. Figures 53 and 54 in the Appendix, see Subsection Websites, show the page of the same beer on both websites. The visual is quite different. However, this does not have a significant impact on the analyses of the human behavior. Indeed, both of these websites provide the basic statistics such as average rating, and number of ratings. They also show a big number named as a *score*. A description of the different visual aspects seen by a simple user is given in the next paragraphs.

BeerAdvocate shows two different scores. The first one is the **BeerAdvocate Overall Score** (BA Score). It is a "*proprietary weighted point (not percentile) system that represents the final overall score for a beer.*"[10]. This score ranges from 100 for the best beers to an unknown value under 60 for the worst beers. The smallest value found in the data is 55. The second score is the **Bros Score**. It is the score given by the founders of this website. The scale for this score is the same as the BA Score.
RateBeer also shows two different scores. The **Overall Score** is a "*score based on its [a beer's] percentile ranking among all beers*"[11]. This score has a scale from 0 for the worst beers to 100 for the best beers. The second score is based on the same principle. However, it only takes into account the beers of the same style hence the name **Style Score**.

The second difference between these two websites is the number of reviews shown. Indeed, BeerAdvocate shows twenty-five reviews per page and RateBeer shows only ten reviews. The Internet Archive Wayback Machine [18] can be used to "go back in time" and search for the number of reviews displayed to the users. Figure 3 shows the results for BeerAdvocate and RateBeer.

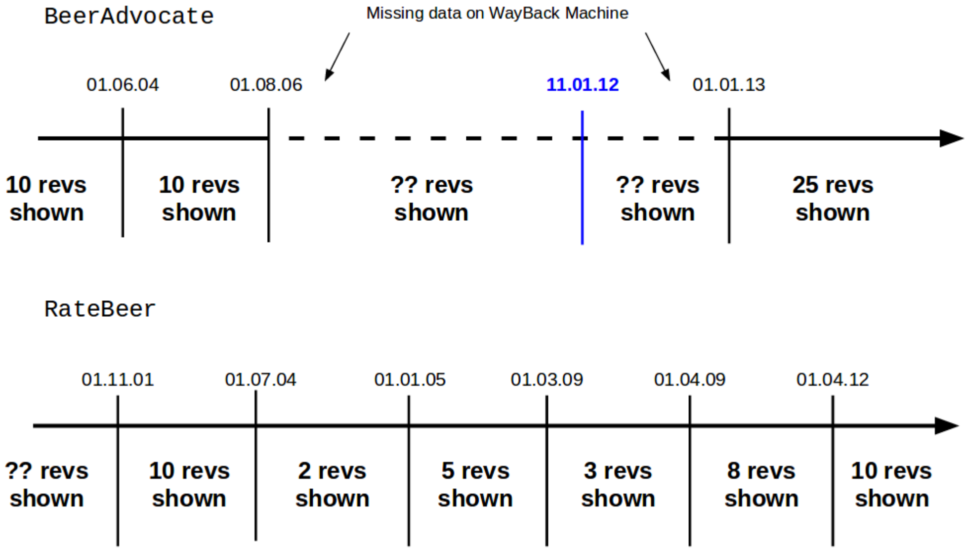


Figure 3: Time line of the number of reviews shown to the users.

Much data is missing for BeerAdvocate on the Wayback Machine. However, it still shows that the number of reviews shown to the users is not consistent over time.

[10] cited from a post on the forum of Beeradvocate: BeerAdvocate Ratings, Explained
[11] cited from a page on RateBeer : RateBeer Ratings Quality

Regarding the reviews themselves, they are shown similarly on both websites. The users can see the final rating and the ratings given to each of the aspects by the previous users. The date of the review, as well as the text, is given. On BeerAdvocate, the text is not mandatory. They decide not to show the reviews without text. The user can still click in a checkbox to display these reviews as well. For the thesis, it is not a problem since the data collected only contain the reviews with text and they are shown to the users. There is also a small difference in the vocabulary used for the reviews. Indeed, BeerAdvocate is using the words *smell* and *feel* while RateBeer is using *aroma* and *palate*. These two pairs define the same aspect. For the rest of the thesis, the denominations from RateBeer are used.

The main difference between BeerAdvocate and RateBeer is the rating system. As shown in Table 3, BeerAdvocate and RateBeer have a different policy on the rating scales.

	BeerAdvocate		RateBeer	
	(min, max, Δ)	# values	(min, max, Δ)	# values
Aroma	(1, 5, 0.5)	9	(1, 10, 1)	10
Appearance	(1, 5, 0.5)	9	(1, 5, 1)	5
Taste	(1, 5, 0.5)	9	(1, 10, 1)	10
Palate	(1, 5, 0.5)	9	(1, 5, 1)	5
Overall	(1, 5, 0.5)	9	(1, 20, 1)	20

Table 3: Ratings of the aspects on both websites.

Using different scales for the aspects, RateBeer directly demonstrates the impact of each of the ratings. That is a good thing because the users directly know which aspect is the most important for the final rating. However, at the same time, it makes the rating harder for the user. They cannot directly compare the rating given to an aspect to another one. BeerAdvocate, on the other hand, uses the same scale for all the aspects. They allow for half points.[12] The weights for the final ratings are available on a page in the BeerAdvocate forum.[13]. Table 4 show the weights of each aspect for the final rating. We see that they are quite similar except for Taste and Overall. Their weights are swapped between the two websites.

	BeerAdvocate	RateBeer
Aroma	24%	20%
Appearance	6%	10%
Taste	**40%**	**20%**
Palate	10%	10%
Overall	**20%**	**40%**

Table 4: Weights of the aspects for the final rating.

This difference in the weights is used later in the thesis, see Table 9 in Subsection The Aspects.

[12] In 2017 (July 4, 2017), the value for Δ is 0.25. We searched in the data to make sure that the ratings never had a quarter-point. Therefore, we can say that this change occurred after the data were scrapped.

[13] Link to the forum page with the weights of BeerAdvocate: How to Review a Beer

Understanding the Data

In this section, a first analysis is done on the data. The previous section explained the technical and visual differences between both websites. It is then mandatory to check if these differences are reflected in the data, *i.e.* the users' reviews.

The Ratings

The very first thing to study is the distributions of the ratings for the two websites. The term "ratings" means here the final rating computed with the weights in Table 4. Indeed, even if the number of beers, the number of users, and the number of reviews are different between the two websites, the same distribution of ratings is expected. Figure 4 shows the histogram of ratings' distribution for BeerAdvocate and RateBeer. The distributions are not the same.

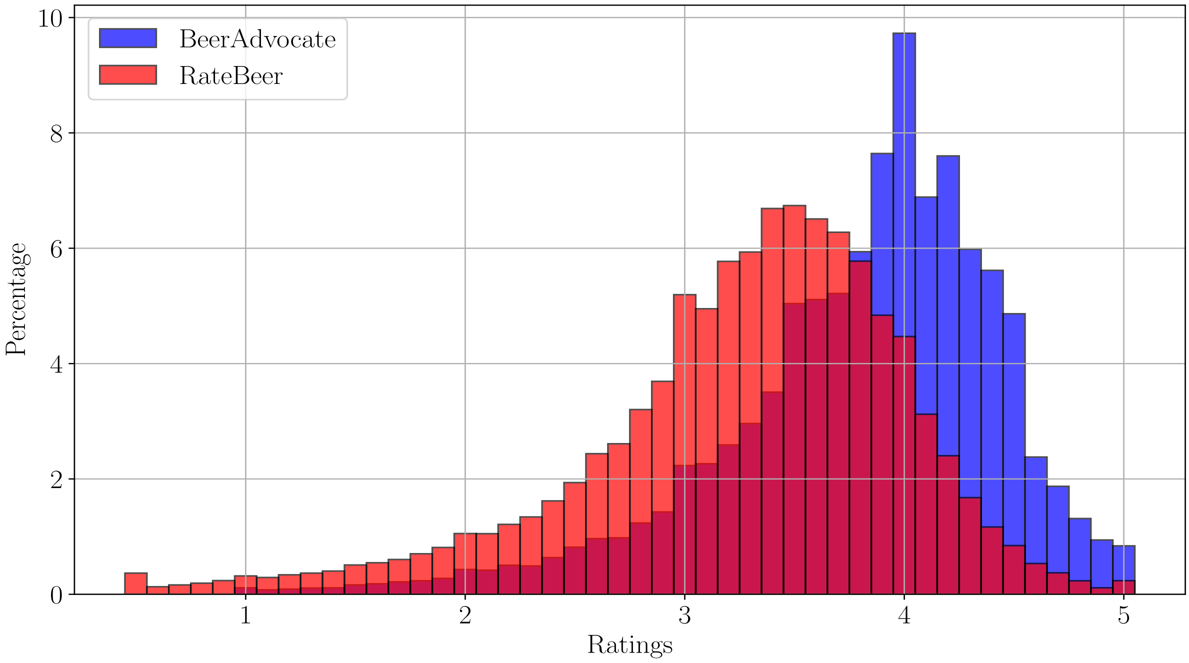


Figure 4: Histogram of the distribution of ratings for BeerAdvocate and RateBeer. The y axis is in percentage due to the different number of reviews.

The reason why these distributions are so different cannot be explained yet. The first hypothesis that can be stated is the fact that the rating systems from both websites are different. As the very first test for this hypothesis, the effects of the aspects with their weights on the final rating is studied. More advanced studies are done later in the thesis. In the Appendix, Subsection Effects of the Aspects on the Final Rating, the matrices of the effects of each of the aspects on the final rating are given. It is then possible to compute the average value for the effects of each aspect and sum them to calculate the average effect on the final rating. The results are given in Table 5.

It is clear that some of the aspects have the tendency to increase the ratings on one of the two websites. In average, only palate has a zero effect due to the same weight and almost equivalent scale. We see that the effect of overall is lessened compared to the effect of taste in the other direction. Since they have the same inverted weights, it means that this difference in average effect is due to the difference in scale. In the end, if all of these effects are summed, the final effect has a positive ratio, meaning that the BeerAdvocate ratings will be 0.25 points higher than the RateBeer ratings only due to the different weighting and scales.

aspect	average effect
Appearance	-0.12
Aroma	0.17
Overall	-0.45
Palate	0.00
Taste	0.65
Total	0.25

Table 5: Average effect of each aspects. The average is computed on the matrices in the Subsection Effects of the Aspects on the Final Rating. A negative value means higher value for RateBeer than BeerAdvocate. A positive value means the inverse.

The previous table does not give the whole explanation of the differences between the ratings. However, it already gives a starting point. However, the ratings in both datasets have to be compared. Therefore, using the z-score (Standard score) is a good option to get rid of the difference in average. It is defined by:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where:

x is the raw value.
μ is the mean value of the population.
σ is the standard deviation of the population.

The significant impact of standardizing the ratings this way is that the mean is equal to 0 and the standard deviation is equal to 1. Table 6 gives some statistics on the raw ratings and the standardized ratings.

	Raw ratings		**Z-score**	
	BeerAdvocate	RateBeer	BeerAdvocate	RateBeer
Max	5.000	5.000	1.947	2.345
75th percentile	4.200	3.800	0.669	0.719
Median	**3.900**	**3.400**	0.189	0.177
25th percentile	3.480	2.900	-0.482	-0.501
Min	1.000	0.500	-4.446	-3.752
Mean	**3.782**	**3.269**	0	0
STD	0.626	0.738	1	1

Table 6: Some statistics about the raw ratings and the standardized ratings.

Table 6 shows a big difference in the ratings between BeerAdvocate and RateBeer. BeerAdvocate ratings are skewed towards the high ratings. It is impossible to draw a conclusion on the origin of this difference yet, as stated before. On the other hand, with the use of the standardized score, it is possible to compare them. Indeed, the percentiles are closer than with the raw ratings. Figure 5 shows the two distributions of standardized ratings. There are some differences in the two distributions, but it is better for comparing the ratings between the two websites.

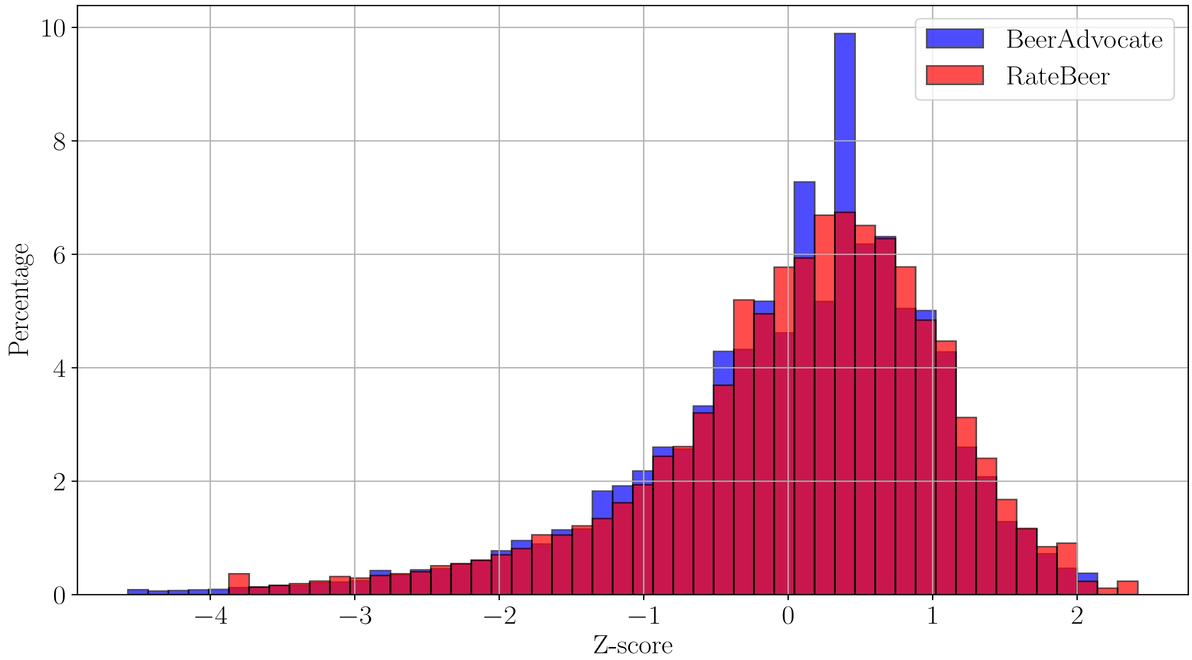


Figure 5: Histogram of the distribution of standardized ratings for BeerAdvocate and RateBeer. The y axis is in percentage due to the different number of reviews.

The Aspects

It is interesting to study the influence of the aspects on the final rating. The difference between BeerAdvocate and RateBeer is not studied here. So far, it is known that the weights for the final rating are different between BeerAdvocate and RateBeer. It is therefore expected that the aspects with a high weight will have a better correlation with the final rating. The Pearson correlation coefficient is used to inspect the relationship between the aspects and the final rating. It is defined by:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

where:

cov	is the covariance
σ_X	is the standard deviation of X.
σ_Y	is the standard deviation of Y.

The function `pearsonr` from the package `scipy.stats` [19] is used to avoid to code it again. The results are given in Table 7 for BeerAdvocate and in Table 8 for RateBeer. Last line and last column show that the most correlated aspects with the final rating are overall and taste. It corresponds to the weighting of the final ratings. This result was expected. Regarding the correlation of the aspects themselves, we see that the most correlated aspects are overall and taste. This result is a bit more interesting than the previous one. It is possible to conclude from this correlation that the users have the tendency to give a good/bad overall rating if the beer had a good/bad taste. For the other aspects, the correlation is never 0 (smallest value is 0.502 for overall and appearance in BeerAdvocate). Therefore all the aspects are always a bit correlated. This conclusion is coherent since an excellent beer should fulfill all the aspects, for example.

	Aroma	Appearance	Taste	Palate	Overall	Rating
Aroma		0.561	0.717	0.617	0.616	0.837
Appearance	0.561		0.547	0.567	0.502	0.642
Taste	0.717	0.547		0.734	0.790	0.948
Palate	0.617	0.567	0.734		0.702	0.808
Overall	0.616	0.502	0.790	0.702		0.867
Rating	0.837	0.642	0.948	0.808	0.867	

Table 7: Pearson correlation coefficient between the different aspects of the BeerAdvocate ratings and the final ratings.

	Aroma	Appearance	Taste	Palate	Overall	Rating
Aroma		0.538	0.794	0.601	0.789	0.878
Appearance	0.538		0.547	0.565	0.566	0.667
Taste	0.794	0.547		0.696	0.876	0.929
Palate	0.601	0.565	0.696		0.701	0.775
Overall	0.789	0.566	0.876	0.701		0.959
Rating	0.878	0.667	0.929	0.775	0.959	

Table 8: Pearson correlation coefficient between the different aspects of the RateBeer ratings and the final ratings.

Now that some knowledge about the correlation between the different aspects of the reviews has been acquired, it is interesting to find which one is easier to predict using the others. It is known from the two previous tables that overall and taste are the most correlated aspects. Therefore, it would be logical that they are easier to predict. In order to verify this statement, a simple linear regression from the package `sklearn.linear_model` called `LinearRegression` [20] is used. One aspect at a time is predicted using the other four aspects for each of the reviews. A 10-fold cross-validation is done to get a better approximation of the RMSE. The results are given in Table 9. The two aspects with the smallest RMSE values are overall and taste for both websites. It confirms the fact that these two values are the most useful for the users. The high correlation, as well as the high weight for these two aspects, also plays a role in these successes.

	BeerAdvocate			RateBeer		
	Mean	STD	**RMSE**	Mean	STD	**RMSE**
Aroma	3.736	0.698	**0.462**	3.177	0.820	**0.467**
Appearance	3.842	0.616	**0.478**	3.432	0.813	**0.634**
Taste	3.793	0.732	**0.375**	3.226	0.811	**0.358**
Palate	3.744	0.682	**0.425**	3.257	0.830	**0.554**
Overall	3.816	0.721	**0.421**	3.300	0.838	**0.369**

Table 9: Mean value, standard deviation, and RMSE of the predictions of each aspect using the four other aspects.

The Score

In the data scrapped by Julian McAuley, all the information shown in the Figure 53 and in Figure 54 in the Appendix can be extracted. For the moment, only the reviews, the scores, and the average value are parsed. Once all of this information is gathered together, it is possible to plot the scores in function of the mean ratings.

The secondary scores in function of the weighted average given on both websites are shown in Figure 60 for BeerAdvocate and in Figure 61 for RateBeer, see Subsection Influence of the Scores on the Ratings - Supplementary Figures in the Appendix. They are not shown here because these scores are not that interesting. There are 19'203 beers for BeerAdvocate and 35'389 for RateBeer with a score and a weighted average. The results with the primary scores are shown in Figure 6 for BeerAdvocate and in Figure 7 for RateBeer.

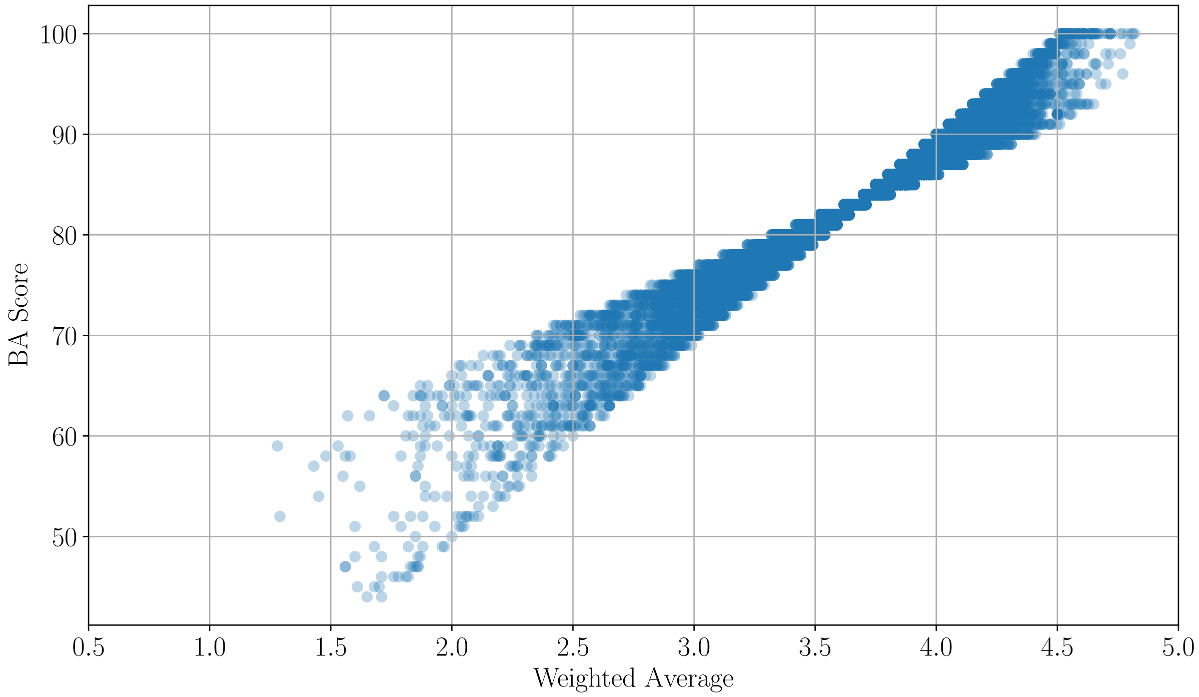


Figure 6: BA Score in function of the weighted average in BeerAdvocate.

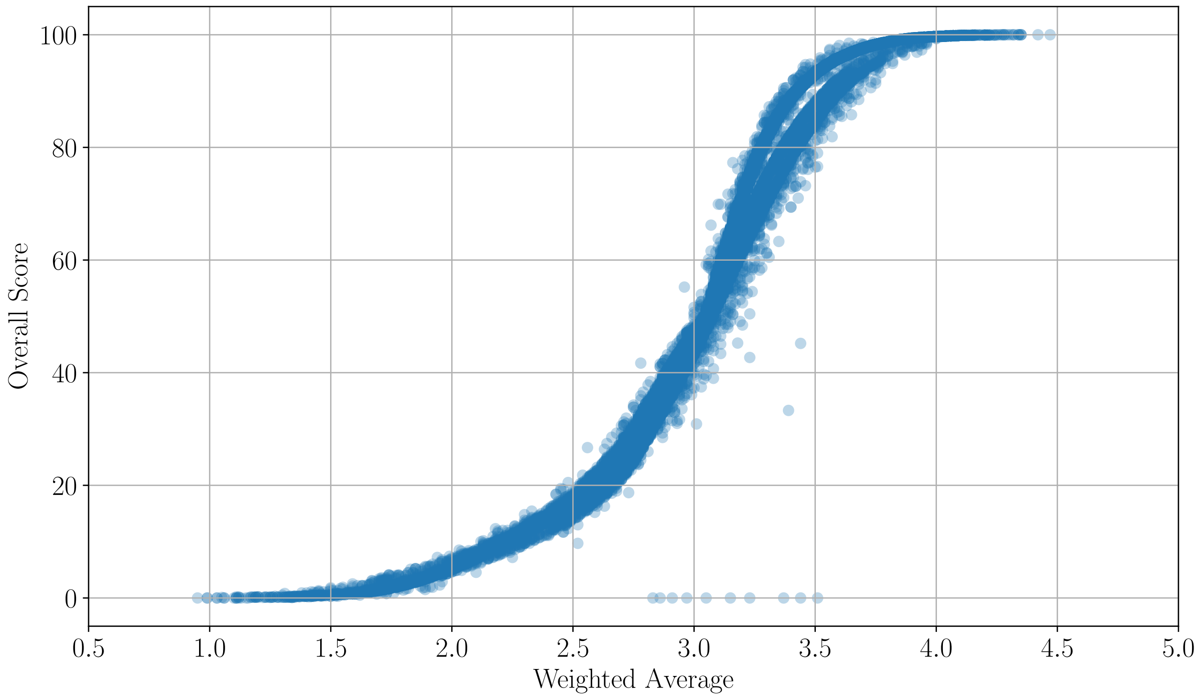


Figure 7: Overall Score in function of the weighted average in RateBeer.

Just by looking at the shapes of thes graphs, it is indisputable that the scores are computed in various ways. In Section Websites, a first explanation was already given: the Overall Score from RateBeer is a *percentile score* while the BeerAdvocate Overall Score is a *weighted point system*. The use of the weighted average is needed to investigate further these scores. However, the modelisation of these averages is quite difficult. Indeed, BeerAdvocate states about the average: "*A weighted average is "an average that takes into account the proportional relevance of each component, rather than treating each component equally." We don't disclose the exact methods used to calculate the Avg, however, we can say that each user's rating history is now included in criteria.*"[14] RateBeer is saying about their weighted average: "*We use a weighted average (Bayesian) to calculate beer and brewer scores.*"[15]. Both of the websites also block the reviews of users with less than ten ratings and if the rating is "*obviously bogus*". RateBeer is giving the Bayesian formula they are using, but after many tests, it was not possible to compute the same weighted average found on their website. There is surely some other components hidden alongside this Bayesian formula.

To continue the analysis of these scores, the simple average of all the reviews is used. Figure 8 for BeerAdvocate and Figure 9 for RateBeer show the main score in function of the mean computed on all the reviews. These graphs are thicker than the previous ones. One of the reasons is, at least for RateBeer, that they are using some Bayesian formula to make the average a little bit closer to the real average of the beer. In other words, if a beer has a few excellent ratings, its average will be lowered a little bit. Therefore, it is expected that the number of reviews plays a major role in modeling these scores. Figure 10 for BeerAdvocate and Figure 11 for RateBeer show the previous graphs with different colors for the points in function of the log of the number of reviews. It is now clear that a beer with a high average rating but few reviews see its score lowered compared to the beers with many reviews.

[14] cited from a post on the forum of Beeradvocate: BeerAdvocate Ratings, Explained

[15] cited from a page on RateBeer: RateBeer Ratings Quality

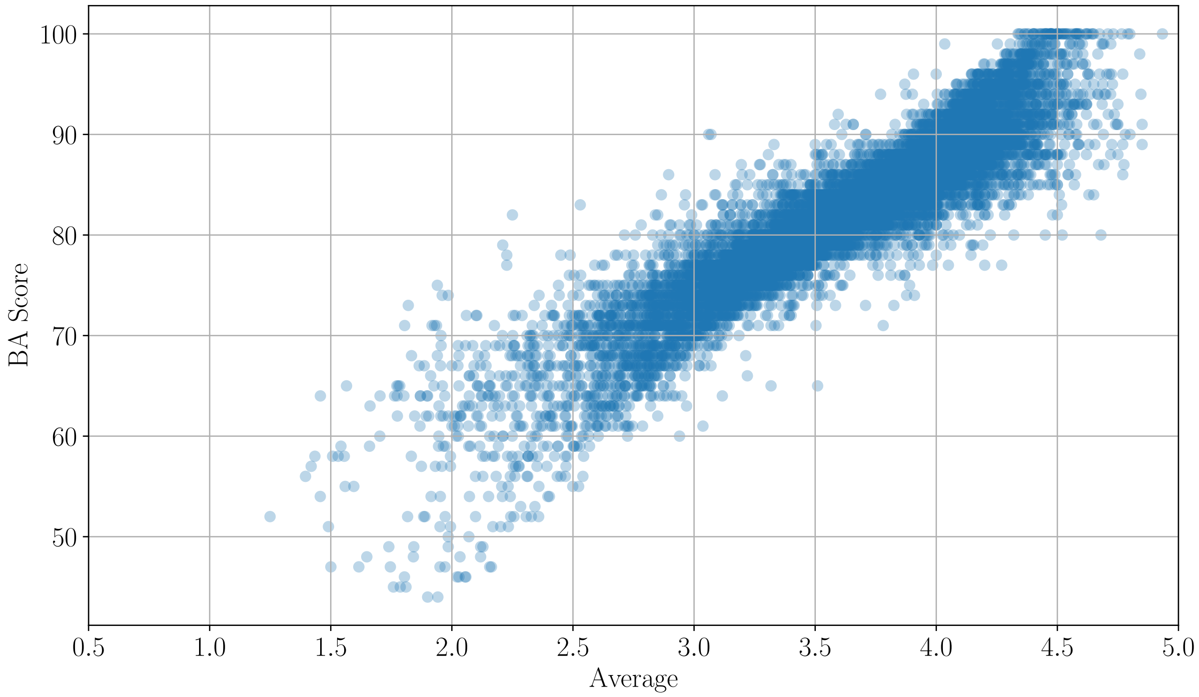


Figure 8: BA Score in function of the average in BeerAdvocate.

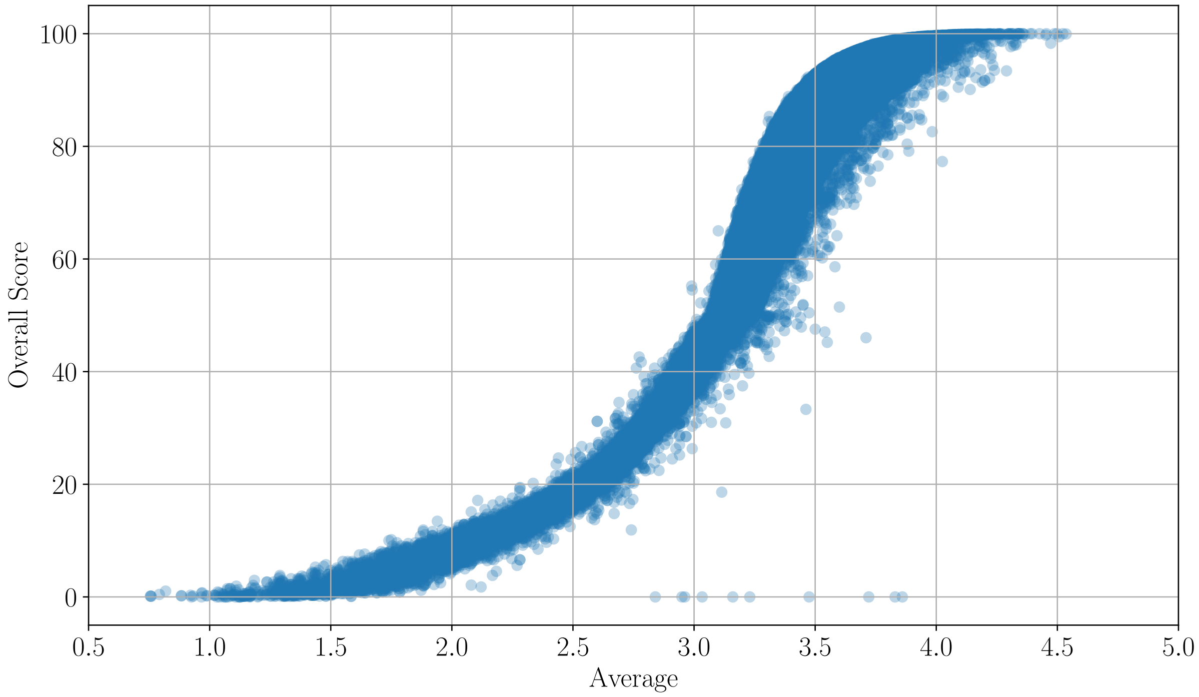


Figure 9: Overall Score in function of the average in RateBeer.

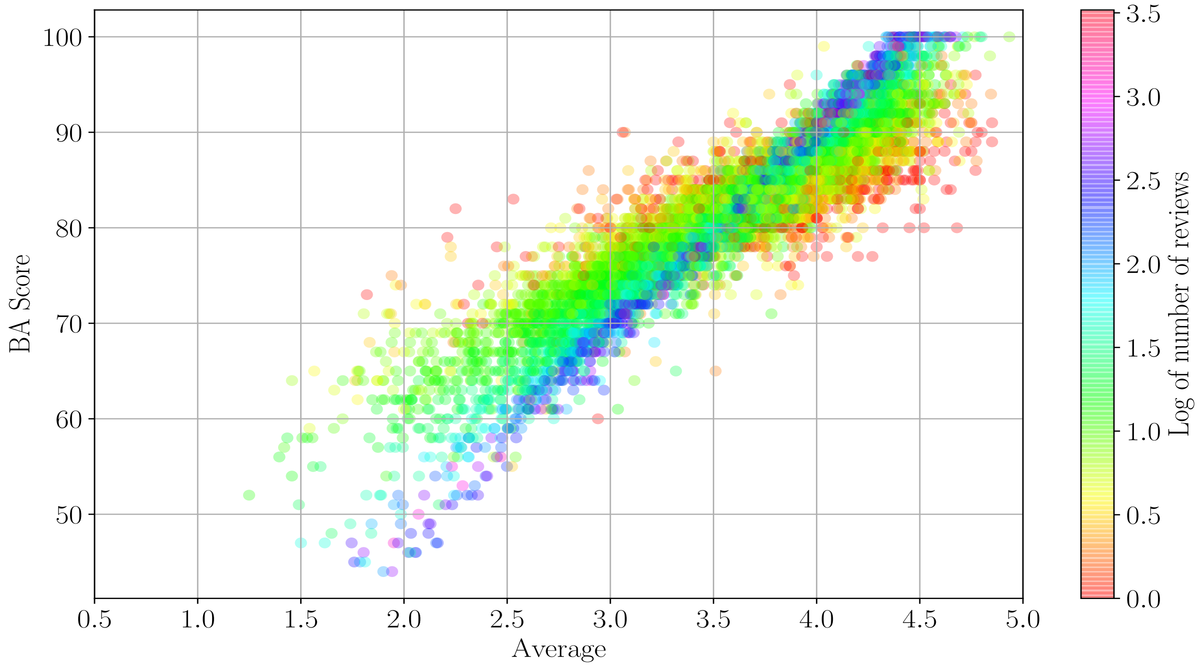


Figure 10: Overall Score in function of the average in BeerAdvocate. Colors are in function of the log of the number of reviews.

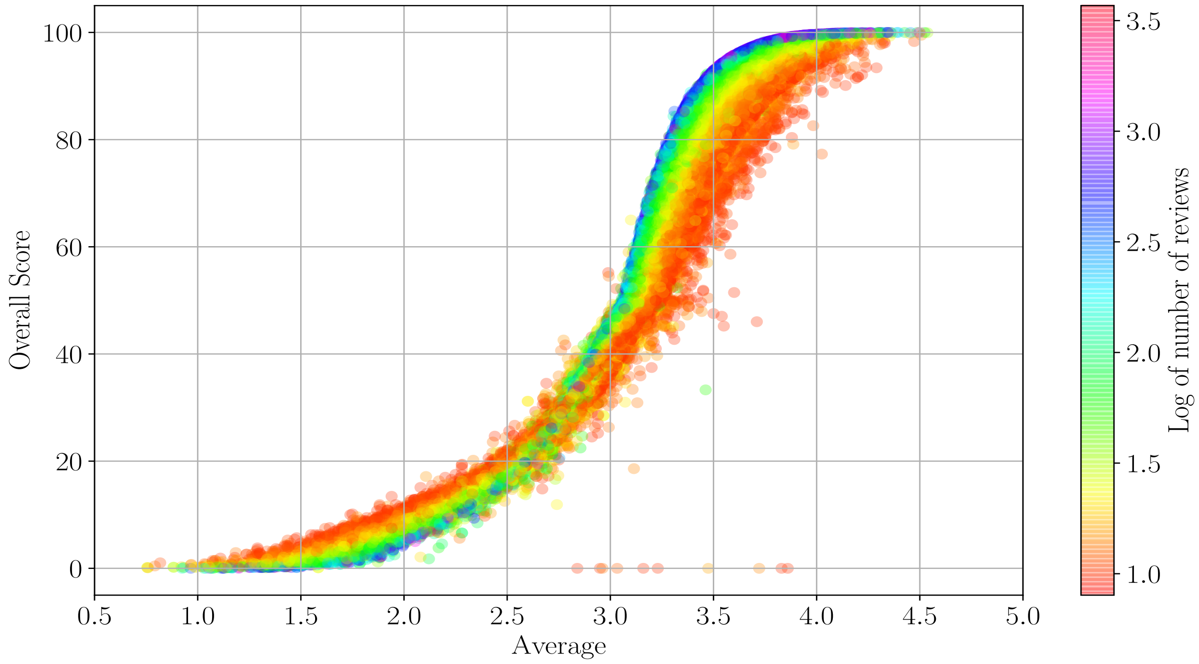


Figure 11: Overall Score in function of the average in RateBeer. Colors are in function of the log of the number of reviews.

The modelisation of these scores would imply to reverse-engineer the whole websites. The aim of this project is to understand how the social effects happen in these kind of online communities, not to find how some developers have built their website. The influence of these scores on the users is, thus, more useful. This is developed in section Influence of the Scores on the Ratings.

Matching

In this section, the algorithm for matching the beers and the users in the two datasets is presented. Indeed, having two datasets on the same kind of product allows extracting the same product in the two different communities. Also, it is known that these two websites are from the U.S. Therefore, it is expected to find some users who are on both websites. It will give an interesting subset of users and beers present in the two communities.

The Beers

The set of beers from BeerAdvocate is defined as $\mathcal{I}_{BA}$ and the set of beers from RateBeer as $\mathcal{I}_{RB}$. Each element i in $\mathcal{I}_X$, X being either BA or RB, is a pair $(b, w)_{i,X}$ where b is the name of the beer and w is the name of the brewery. The values of b and w do not have to be unique. However, the combination of these two strings is unique. Indeed, it is possible to have multiple beers for one brewery. All the pairs also have a pair of IDs used by the websites. It is impossible to match on these IDs since they are different between the two websites. Therefore, the matching is done on the strings. The cardinal of $\mathcal{I}_{BA}$ is 66'055, and the cardinal of $\mathcal{I}_{RB}$ is 112'171. This setup is the perfect example for the Hungarian algorithm [21] (also known as the Kunh-Munkres algorithm). Indeed, an exact matching on the beers would be perfect. Moreover, a beer in one website cannot match two or more beers in the other websites. However, the complexity of the Hungarian algorithm is in $\mathcal{O}(n^3)$, n being here the minimum cardinal. The computation would take too much time. Therefore, an algorithm based on the fact that the number of breweries is smaller than the number of beers has been developed. The set of breweries from BeerAdvocate is defined as $\mathcal{W}_{BA}$ and the set of breweries from RateBeer as $\mathcal{W}_{RB}$. The cardinal of $\mathcal{W}_{BA}$ is 5'730, and the cardinal of $\mathcal{W}_{RB}$ is 7'419. Since these cardinals are smaller than the cardinals of the sets of beers, the matching is first done on the breweries and then on the beers with the matched breweries. This process removes much computation since only around 1/6000 of all the beers are compared.

To match the name of the breweries and the name of the beers, the TF–IDF (Term Frequency–Inverse Document Frequency) algorithm and the cosine similarity are used. All the breweries in BeerAdvocate are matched to the breweries in RateBeer in this order. Simply because there are fewer breweries in BeerAdvocate. Therefore, the corpus used by the TF–IDF algorithm is prepared with all the breweries name from BeerAdvocate. Then, the algorithm vectorizes this corpus into a matrix of TF–IDF features. This is done using the function `TfidfVectorizer` from the package `sklearn.feature_extraction.text` [20]. Then the algorithm also vectorizes the breweries from RateBeer using the corpus of BeerAdvocate with the same function. Finally, the function `cosine_similarity` from the package `sklearn.metrics.pairwise` [20] is used to create a matrix of similarity values. Inside this large matrix, for each brewery in BeerAdvocate, only the two best matches from RateBeer are kept. The difference between the similarity of the best match and the similarity of the second best match is computed. Finally, the algorithm keeps only the pairs (brewery from BeerAdvocate and best match from RateBeer) with a high similarity value and a high difference between the first two similarity values. The second condition is added to make sure that the match is unique. If there is a doubt, the match is simply rejected. This process is summarized in Algorithm 1.

Algorithm 1 Match the Breweries

```
**Define** $\delta_{\text{sim},\mathcal{W}}$, the threshold for the similarity.
**Define** $\delta_{\Delta,\mathcal{W}}$, the threshold for the difference of similarity.

Prepare corpus with all elements of $\mathcal{W}_{RB}$ and $\mathcal{W}_{BA}$. This creates $N_{\mathcal{W}}$ features.
$X_{BA} \in \mathbb{R}^{N_{\mathcal{W}} \times |\mathcal{W}_{BA}|} \Leftarrow$ Vectorize the elements of $\mathcal{W}_{BA}$
$X_{RB} \in \mathbb{R}^{N_{\mathcal{W}} \times |\mathcal{W}_{RB}|} \Leftarrow$ Vectorize the elements of $\mathcal{W}_{RB}$
$M \in \mathbb{R}^{|\mathcal{W}_{BA}| \times |\mathcal{W}_{RB}|} \Leftarrow$ Cosine similarity between $X_{BA}$ and $X_{RB}$
Define $\mathcal{W}_{\text{matched}} = \emptyset$, the set of matched breweries.
for $w$ in $\mathcal{W}_{BA}$ do
  $w_1, w_2 \in \mathcal{W}_{RB} \Leftarrow$ Get the two most similar breweries to $w$ using the matrix $M$
  Compute $\Delta$ the difference of similarity between $w_1$ and $w_2$: $\Delta = \text{sim}_{w_1,w} - \text{sim}_{w_2,w}$.
  if $\text{sim}_{w_1,w} \geq \delta_{\text{sim}}$ and $\Delta \geq \delta_{\Delta}$ then
    Add the pair $(w, w_1) \in (\mathcal{W}_{BA}, \mathcal{W}_{RB})$ to $\mathcal{W}_{\text{matched}}$.
  end if
end for
```

The choice of $\delta_{\text{sim},\mathcal{W}}$ and $\delta_{\Delta,\mathcal{W}}$ is done by checking the results using different values for these two thresholds. Using $\delta_{\text{sim},\mathcal{W}} = 0.8$ and $\delta_{\Delta,\mathcal{W}} = 0.3$ is ideal for the recall. Indeed, it is preferable to lose some matches while being sure that the matches are correct. Therefore, the best recall possible is wanted. With these values, the cardinal of the set $\mathcal{W}_{\text{matched}}$ is 2'338. So, around 30-40% of the breweries are matching between the two websites. Now, the beers from the matched breweries have to be matched as well.

The problem with matching the beer names is that sometimes, the brewery name is in the name of the beer. It can create a false match. Therefore, the original beer names have to be transformed into beer names without the brewery name. Then, the corpus can be prepared for the `TfidfVectorizer` function. Once it is done, the algorithm goes through all matched breweries and gets the beers with the matched breweries. Then, it vectorizes them and computes the similarity matrix. Between the two sets of beers with the matched brewery, the smallest set is selected and the algorithm goes through each beer of this set. The two best matches are retrieved using the similarity matrix, and the difference between these two matches is computed. If there is only one beer in the other set, then the difference is simply equal to the similarity with the first and unique match. Depending on the threshold of the similarity and on the difference between the first two similarities, the beers that match well are kept. At the end of this process, the algorithm has to go through all matched beers to remove duplicates. Indeed, since it is a greedy algorithm, it is possible that some entries are duplicated. It is done to get a higher recall. The whole process is summarized in Algorithm 2.
The choice of $\delta_{\text{sim},\mathcal{B}}$ and $\delta_{\Delta,\mathcal{B}}$ are done by checking the results using different values for these two thresholds, as we did previously for matching the breweries. We found that using $\delta_{\text{sim},\mathcal{B}} = 0.8$ and $\delta_{\Delta,\mathcal{B}} = 0.3$ is ideal for the recall. Using the two previous algorithms leaves us with a set of 16'773 beers that are in both websites.

Algorithm 2 Match the Beers

```
Define: $\delta_{\text{sim},\mathcal{B}}$, the threshold for the similarity.
Define: $\delta_{\Delta,\mathcal{B}}$, the threshold for the difference of similarity.

Tranform the beer names in $\mathcal{B}_{BA}$ and $\mathcal{B}_{RB}$ by removing the brewery names.
Prepare corpus with all element of $\mathcal{B}_{BA}$ and $\mathcal{B}_{RB}$. This creates $N_{\mathcal{B}}$ features.
Define $\mathcal{I}_{\text{matched}} = \emptyset$, the set of matched beers and breweries.
for $(w_{BA}, w_{RB})$ in $\mathcal{W}_{\text{matched}}$ do
  Get $\mathcal{B}_{BA,w_{BA}}$ and $\mathcal{B}_{RB,w_{RB}}$, the sets of beers from BA and RB with matched brewery.
  $X_{BA,w_{BA}} \in \mathbb{R}^{N_{\mathcal{B}} \times |\mathcal{B}_{BA,w_{BA}}|} \Leftarrow$ Vectorize the elements of $\mathcal{B}_{BA,w}$
  $X_{RB,w_{RB}} \in \mathbb{R}^{N_{\mathcal{B}} \times |\mathcal{B}_{RB,w_{RB}}|} \Leftarrow$ Vectorize the elements of $\mathcal{B}_{RB,w}$
  $M_{w_{BA},w_{RB}} \in \mathbb{R}^{|\mathcal{B}_{BA,w_{BA}}| \times |\mathcal{B}_{RB,w_{RB}}|} \Leftarrow$ Cosine similarity between $X_{BA,w_{BA}}$ and $X_{RB,w_{RB}}$
  if $|\mathcal{B}_{BA,w_{BA}}| \leq |\mathcal{B}_{RB,w_{RB}}|$ then
    for $b$ in $\mathcal{B}_{BA,w_{BA}}$ do
      $b_1, b_2 \in \mathcal{B}_{RB,w_{RB}} \Leftarrow$ Get the two most similar beers to $b$ using $M_{w_{BA},w_{RB}}$.
      Compute $\Delta$ the difference of similarity between $b_1$ and $b_2$
      if $b_2 = \emptyset$ then
        $\Delta = \text{sim}_{b_1,b}$
      else
        $\Delta = \text{sim}_{b_1,b} - \text{sim}_{b_2,b}$
      end if
      if $\text{sim}_{b_1,b} \geq \delta_{\text{sim}}$ and $\Delta \geq \delta_{\Delta}$ then
        Add the quadruplet $(b, w_{BA}, b_1, w_{RB}) \in (\mathcal{B}_{BA,w_{BA}}, \mathcal{W}_{BA}, \mathcal{B}_{RB,w_{RB}}, \mathcal{W}_{RB})$ to $\mathcal{I}_{\text{matched}}$
      end if
    end for
  else
    for $b$ in $\mathcal{B}_{RB,w_{RB}}$ do
      $b_1, b_2 \in \mathcal{B}_{BA,w_{BA}} \Leftarrow$ Get the two most similar beers to $b$ using $M_{w_{BA},w_{RB}}$.
      Compute $\Delta$ the difference of similarity between $b_1$ and $b_2$
      if $b_2 = \emptyset$ then
        $\Delta = \text{sim}_{b_1,b}$
      else
        $\Delta = \text{sim}_{b_1,b} - \text{sim}_{b_2,b}$
      end if
      if $\text{sim}_{b_1,b} \geq \delta_{\text{sim}}$ and $\Delta \geq \delta_{\Delta}$ then
        Add the quadruplet $(b_1, w_{BA}, b, w_{RB}) \in (\mathcal{B}_{BA,w_{BA}}, \mathcal{W}_{BA}, \mathcal{B}_{RB,w_{RB}}, \mathcal{W}_{RB})$ to $\mathcal{I}_{\text{matched}}$
      end if
    end for
  end if
end for
Go through $\mathcal{I}_{\text{matched}}$ and remove all the duplicates beers from BA and RB.
```

The Users

The set of users from BeerAdvocate $\mathcal{U}_{BA}$ and from RateBeer $\mathcal{U}_{RB}$ can be retrieved by parsing the data. The location of these users is also parsed. The State is used if a user comes from the U.S. Otherwise the country is used. Matching the usernames is a bit different than matching the beers and breweries. Indeed, the slightest change in the username leads to a different user. Therefore, the threshold for a matched username needs to be higher than the threshold for the matched beers and breweries. A greedy algorithm is also used to match the users. First, it computes the matrix of distances using the function `ratio` from the package `Levenshtein`. Then it goes through all users from RateBeer (since there are fewer users in RateBeer than in BeerAdvocate) and gets the best match in BeerAdvocate. Once all the matches are computed, the algorithm goes through all of them and checks the similarity between the two matches. If it is higher than a given threshold, it keeps it. The algorithm also checks if it is a duplicate. If it is the case, the similarity between the two duplicates is checked, and the best one is kept. In the end, the algorithm just makes sure that there is no duplicate username for both BeerAdvocate and RateBeer. It gives a set of 2'363 users. There is just a small problem. Indeed, some users may have the same username, but they are not the same person. Therefore, the second part of this algorithm matches on the users' location. It simply goes through all the matches and computes the Levenshtein ratio between the two location. If the similarity is higher than a given threshold, the pair is kept. Otherwise, it is simply discarded. This gives a final set of 1'194 users who matched on their username and their location. The whole process is summarized in Algorithm 3. The thresholds δ_{username} and δ_{location} are both set to 0.95 in order to allow for typing errors in the username and the location.

Algorithm 3 Match the Users

```
Define $\delta_{\text{username}}$, the threshold for the similarity of the username.
Define $\delta_{\text{location}}$, the threshold for the similarity of the location.

Define the zero-matrix $M \in \mathbb{R}^{|\mathcal{U}_{BA}| \times |\mathcal{U}_{RB}|}$ for the Levenshtein ratio.
for $u_{BA}$ in $\mathcal{U}_{BA}$ do
  for $u_{RB}$ in $\mathcal{U}_{RB}$ do
    $M(u_{BA}, u_{RB}) = $ Levenshtein.ratio($u_{BA}, u_{RB}$)
  end for
end for
Define $\mathcal{U}_{\text{matched}} = \emptyset$, the set of matched users.
for $u_{RB}$ in $\mathcal{U}_{RB}$ do
  Find $u \in \mathcal{U}_{BA}$ having the highest ratio with $u_{RB}$ using the matrix $M$.
  if $M(u_{BA}, u_{RB}) \geq \delta_{\text{username}}$ then
    Add $(u_{BA}, u_{RB}) \in (\mathcal{U}_{BA}, \mathcal{U}_{RB})$ in $\mathcal{U}_{\text{matched}}$
  end if
end for
for $(u_{BA}, u_{RB})$ in $\mathcal{U}_{\text{matched}}$ do
  Get locations of both users: $l_{u_{BA}}$ and $l_{u_{RB}}$
  Compute $r = $ Levenshtein.ratio($l_{u_{BA}}, l_{u_{RB}}$)
  if $r \leq \delta_{\text{location}}$ then
    Remove pair $(u_{BA}, u_{RB})$ from $\mathcal{U}_{\text{matched}}$
  end if
end for
```

The Beers and the Users

In this subsection, the link between the matched users and the matched beers is done. For each matched users, the set of beers they rated on both websites is compared. If they rated a beer on both websites that matched using the previous algorithm, this pair of ratings is kept. It is pretty straightforward; thus the algorithm is not shown here. In the end, this returns a total of 552 users who rated at least one beer on both websites for a total of 18'109 reviews from the same user on the same beer on both websites. This set of reviews is called $\mathcal{R}_{\text{matched}}$.

These three sets can now be used to perform further analyses on the data. It is compelling to be able to have the same items in two different communities. It is also possible to compare the behavior of the users between both communities and if they belong to one of the communities or both.

Influence of the Scores on the Ratings

In this section, the effect of the score on the ratings is analyzed. The BA score can be modeled by an affine line for beers with a high number of reviews. It's easier to see this in Figure 62 in the Appendix. Therefore the slope is always the same. The Overall Score in RateBeer, on the other hand, has many changes in the slope. It is between flat and almost vertical. Therefore, if the space of average rating is cut into five regions of the same length, there will be a triangle shape for the variance of the RateBeer score and a flat line for the form of the BeerAdvocate score. To remove the outliers, *i.e.* the beers that are not rated on both websites, only the beers in the set $\mathcal{W}_{\text{matched}}$ are used. Only the beers with at least 20 reviews on both websites are used to make sure that the score is displayed when the last user is coming to rate this beer, and influence from the previous reviews could also happen. Thus, there is a total of 2'410 beers. The graphs of the score in function of the average are shown in the appendix, see Figure 64 and Figure 65. The idea is to compare this difference in the variance of the score to the variance of the ratings. Therefore, instead of using all the ratings for the average value on the x-axis, the mean on all the ratings except the last one is used. Then, the standard deviation of the last rating is compared to the standard deviation of the score. Therefore, these two scores are normalized between 0 and 1 to make them comparable. The standardize ratings are also used. Figure 12 shows the graphs of the normalized scores in function of the average standardize ratings without the last one.

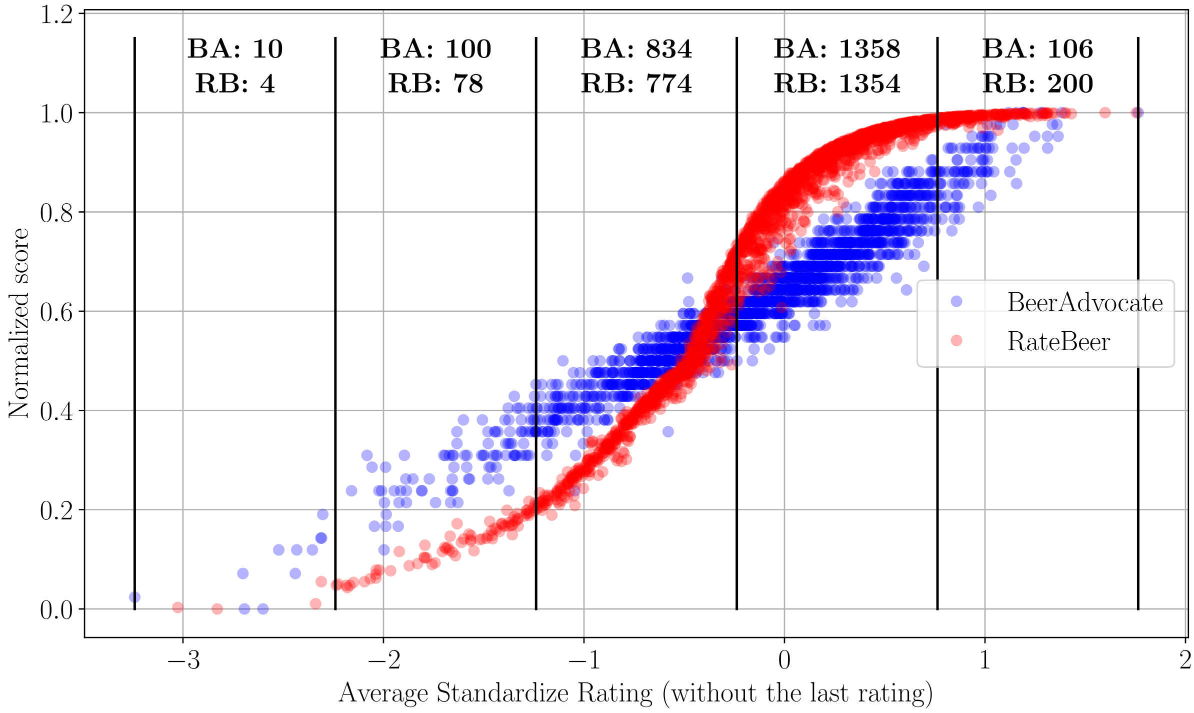


Figure 12: Normalized scores in function of the average standardize ratings (without the last rating). The x axis has been cut in 5 regions of same length.

The standard deviation of the score in function of the mean of all the ratings except the last one is given in Figure 13. The triangle shape for RateBeer is explicit with a peak in the third region. In the mean time, the standard deviation for BeerAdvocate's score is rather flat compared to RateBeer's score.

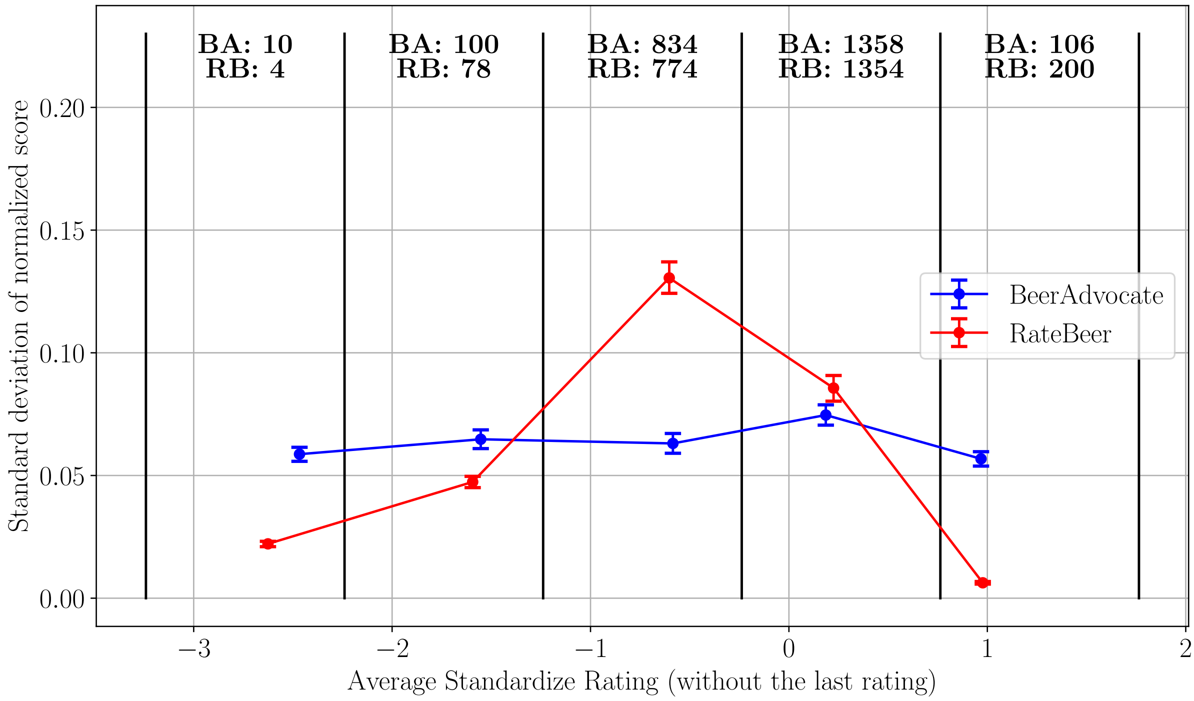


Figure 13: Standard deviation of the score in function of the average standardize ratings (without the last rating). The vertical bars represent the 95% confidence interval. The data have been bootstraped in 10'000 draws of 500 values.

Thus, the hypothesis is: if a user will rate a beer and is highly influenced by the score displayed on the beer's page, then he will tend to follow the indication given by this score. Therefore, the standard deviation of the last rating should follow the same shape as the standard deviation of the scores. Figure 14 shows these results. The shape for the last rating is disparate from the shape of the scores. The fact that the two curves are close means that this an effect independent of the website or the community of the website. Indeed, it seems that users tend to agree on good beers while they do not share the same opinion on the beers with bad ratings.

For this analysis, much data is missing, especially for the beers with a bad average. So, the results are to be taken with a grain of salt. However, if only the third and fourth regions are analyzed, it is clear that the ratings have the same behavior between BeerAdvocate and RateBeer. They do not follow the behavior of the scores. Now, it would be interesting to study the other components that can influence a user, for example, the previous reviews.

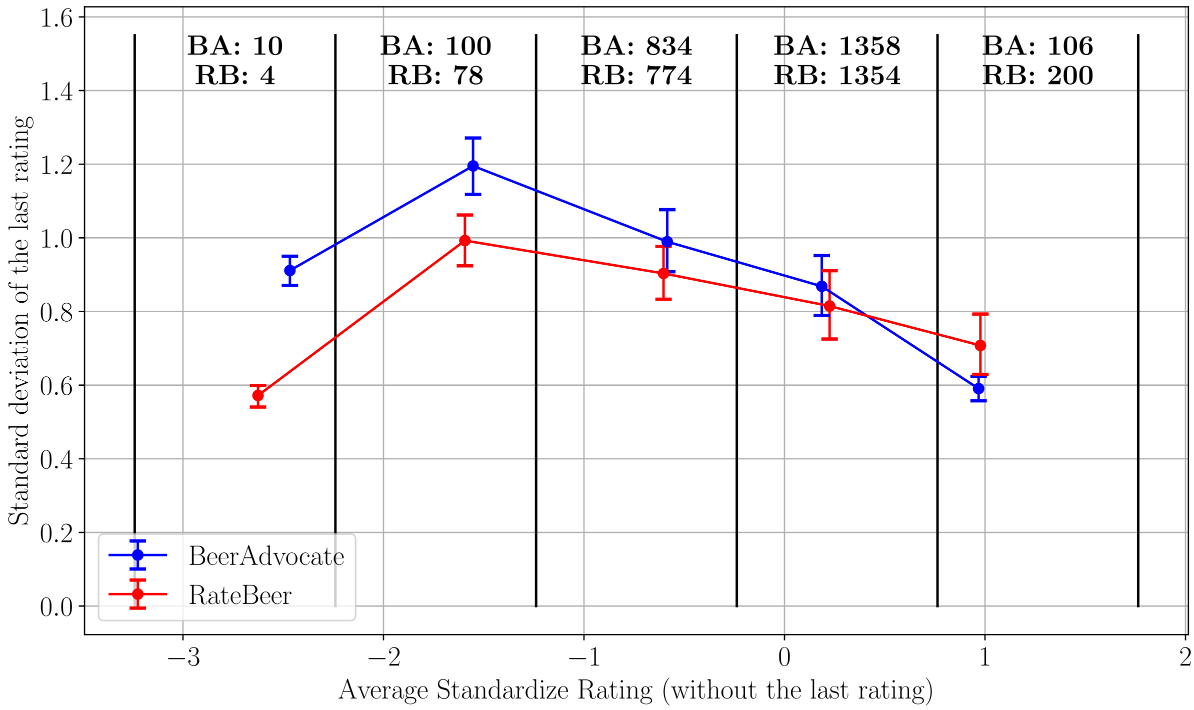


Figure 14: Standard deviation of the last rating in function of the average standardize ratings (without the last rating). The vertical bars represent the 95% confidence interval. The data have been bootstraped in 10'000 draws of 500 values.

Predictability of the Ratings

In this section, the predictability of the users over the years is investigated. The prediction of the ratings given by the users can be useful to create a Machine Learning model for predicting the ratings. It will also provide some valuable insight on the evolution of the communities from both websites over the years. First, it is better to make sure that the big score does not have any effect, even the smallest one. Using the Wayback Machine [18], it was found that the number of reviews required to display the score is ten, in general. For a few periods of time, only five reviews were required. Therefore, the prediction is made on the tenth rating using all the previous ratings. It is also interesting to see if the older ratings influence less than the newer ratings. Therefore, all the beers having at least ten reviews are collected, and they are classified based on the year of the tenth rating's timestamp. The number of beers for BeerAdvocate and RateBeer are given in Table 10.

	BeerAdvocate	RateBeer
2000	0	55
2001	11	607
2002	627	1'545
2003	1'087	1'607
2004	918	2'232
2005	959	2'929
2006	1'128	3'196
2007	1'270	3'744
2008	1'523	3'941
2009	1'784	4'729
2010	2'160	4'793
2011	2'614	5'779
2012	93	171
total	14'174	35'328

Table 10: Number of beers in both datasets per year. The minimum number of reviews per beer is 10. The beer are classified by year on the timestamp of the tenth rating.

The best way to investigate this is to use the model `LinearRegression` from the package `sklearn.linear_model` [20] with only one feature being the i-th rating, $\forall i < 10$. Then, the data are bootstraped to obtain the 95% confidence interval on the RMSE. The data are drawn 500 times. All the reviews in the given year are used and are randomly added to the training (90%) or to the testing set (10%). Finally, the plot of the results per year can be shown to see the evolution of these predictions over the years. The results are given in Figure 15 for BeerAdvocate and Figure 16 for RateBeer. For BeerAdvocate, the average RMSE is rather constant over the years. On the other hand, a drop in the average RMSE can be observed for RateBeer. Indeed, in 2001, the average RMSE was around 1, and in 2011 it was around 0.5. Either the RateBeer users are becoming more predictable over the years, or there is a hidden effect behind this drop of RMSE.

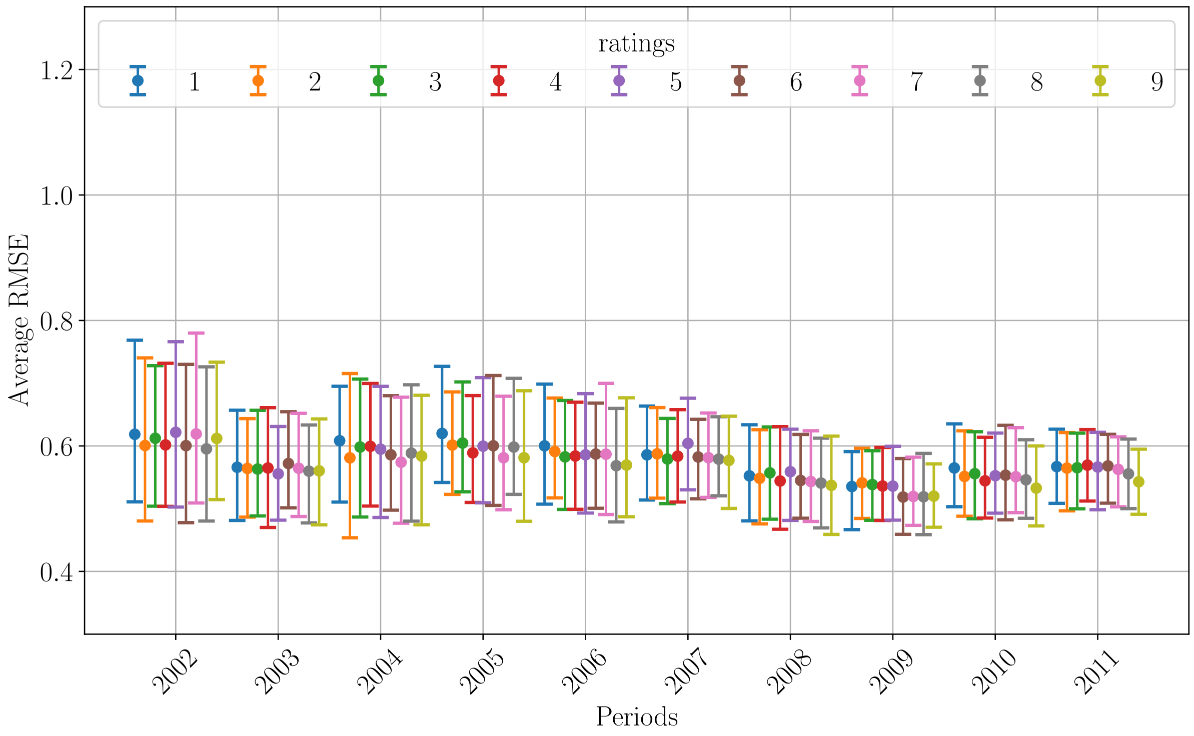


Figure 15: Prediction of tenth rating using all the previous ratings for BeerAdvocate. A linear regression is used for prediction the tenth rating.

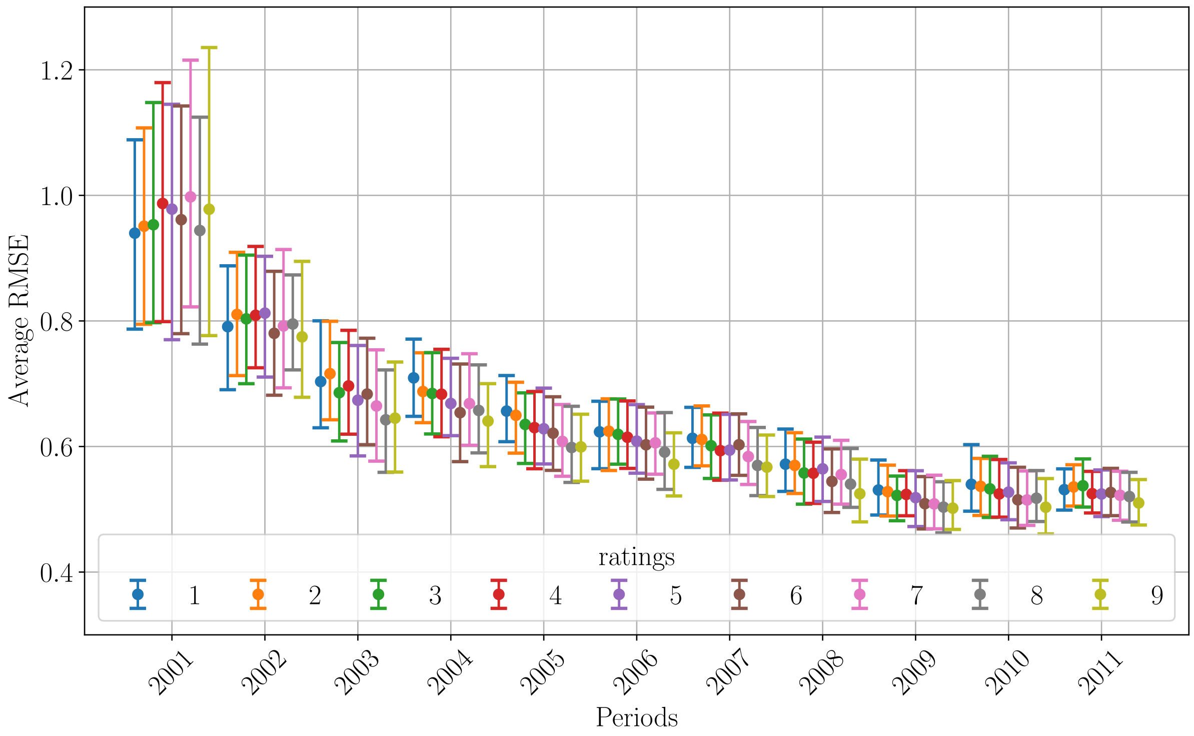


Figure 16: Prediction of tenth rating using all the previous ratings for RateBeer. A linear regression is used for prediction the tenth rating.

The prediction using a linear regression is very dependent on the standard deviation of the data. Therefore, it is always a good idea to plot the average standard deviation of the first ten ratings over the years for BeerAdvocate and RateBeer, see Figure 17. It is clear that there is an important drop for RateBeer. The plots detailed for each rating are given in the Appendix, see Figure 66 for BeerAdvocate and Figure 67 for RateBeer.

It is not clear if this drop of standard deviation is playing a role in the decrease of RMSE seen in the previous figures. Thus, the same process is repeated except that the RMSE is divided by the standard deviation on the reviews of each rating from each year. The results are given in Figure 18 for BeerAdvocate and Figure 19 for RateBeer. The slope is now flat. Therefore, the increase of prediction was indeed induced by the drop of standard deviation.

In the meantime, it is interesting to note that there is a drop in standard deviation for RateBeer, see Figure 17. The effects are present in both datasets, even if it is less strong for BeerAdvocate. Over the years the communities are becoming bigger and bigger. Therefore, the norms of both communities are also more rooted. Therefore, the conclusion is that the users have the tendency to follow the norm of the whole community, even when they are the first one to rate a beer.

The influence of the ratings' positions has not been mentioned yet. Figure 15 for BeerAdvocate and Figure 16 for RateBeer shows that the ninth rating leads to a slightly smaller RMSE than the first rating. This difference is not significant. This difference is also disappearing if the RMSE is divided by the standard deviation, see Figure 18 for BeerAdvocate and Figure 19 for RateBeer. Therefore, the standard deviation seems to be constant between the older and newer ratings.

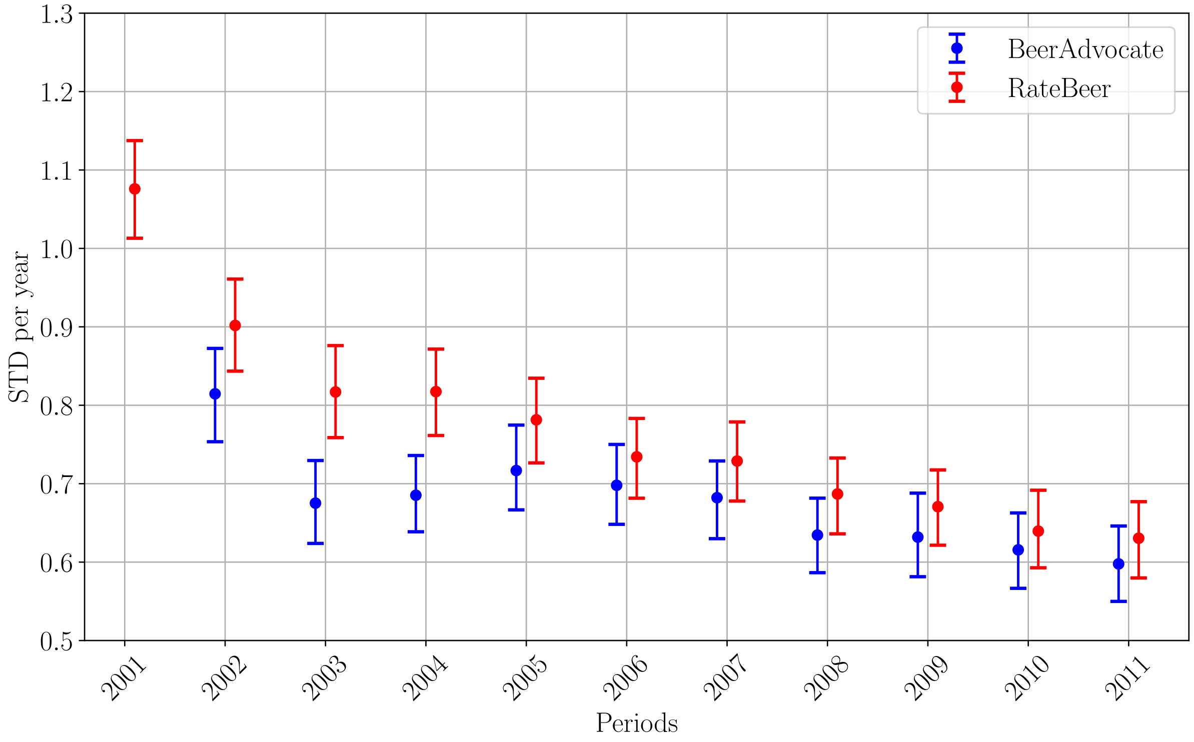


Figure 17: Average Standard deviation on the ten first ratings by year for BeerAdvocate and RateBeer

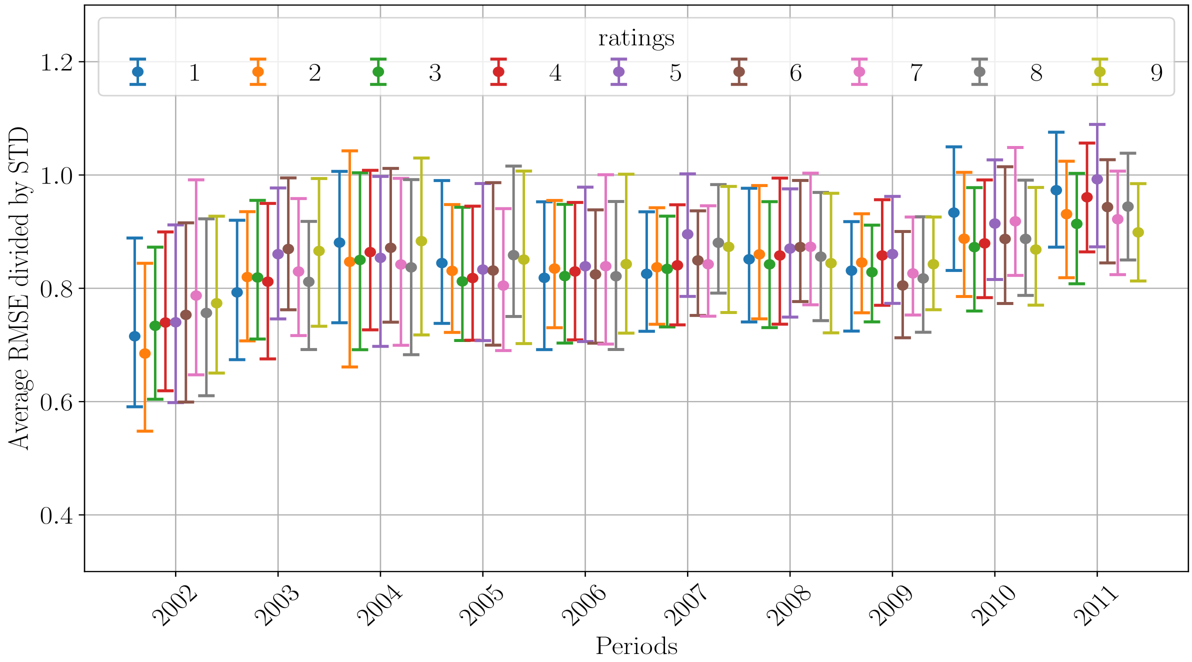


Figure 18: Prediction of tenth rating using all the previous ratings for BeerAdvocate. A linear regression is used for prediction the tenth rating. The average RMSE is divided by the standard deviation of the ratings.

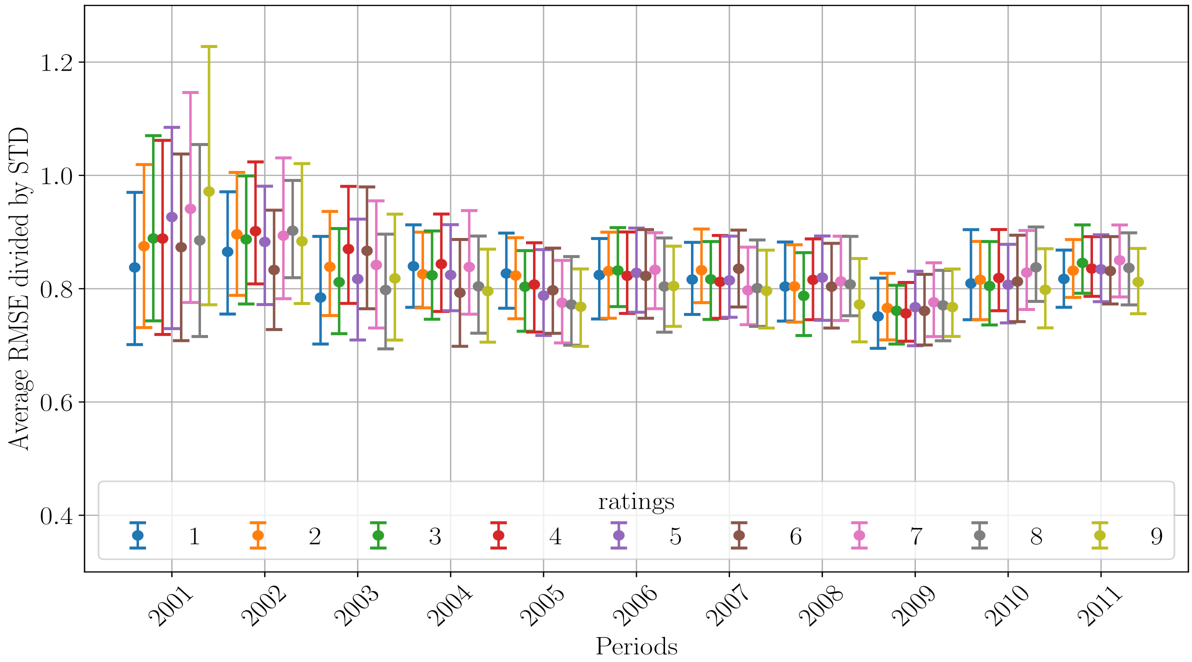


Figure 19: Prediction of tenth rating using all the previous ratings for RateBeer. A linear regression is used for prediction the tenth rating. The average RMSE is divided by the standard deviation of the ratings.

Herding Effects

In the previous section, it was demonstrated that the standard deviation of the first ten ratings is going down over the years. It shows that the norm of the community is becoming more and more important. If more years of data were available, a plateau would be surely seen for the standard deviation. Because even if the users are following the community, they still have their own opinion.

The analysis is moving towards the herding effects on the products themselves instead of the whole website. To explore this, a framework that is well controlled is required. The idea emerged from the paper of Muchnik *et al.* [22]. The aim of their paper is to "*analyze and quantify the effects of social influence on users' ratings and discourse on a social news aggregation Website*". To achieve this, they created three different populations before the users could see the comments or articles. The populations are the "up-treated", the "down-treated", and the "control". They simply added a +1 rating when the comment was created to put it in the "up-treated" population and they added a -1 rating to put it in the "down-treated" population. For the control population, they did not influence the ratings. The idea is to achieve the same randomized experiment with these data. However, it is not possible to influence the ratings in the same way. The only way to do this is to classify the first rating of each beer between High, Medium or Low. The definition of a High, Medium or Low rating is needed to classify the first rating. It is also possible to classify either to use the whole dataset, *i.e.* compare the first rating to the global average, or using the products themselves, *i.e.* compare the first rating to the product average. Since this analysis is on the evolution of the beers ratings, the set $\mathcal{I}_{\text{matched}}$ of matched beers between BeerAdvocate and RateBeer is used. The ratings are standardized as in Equation 1 and only the beers with at least ten reviews are kept since this analysis is done on the first ten ratings. The distribution of difference to the global average is given in Figure 20.

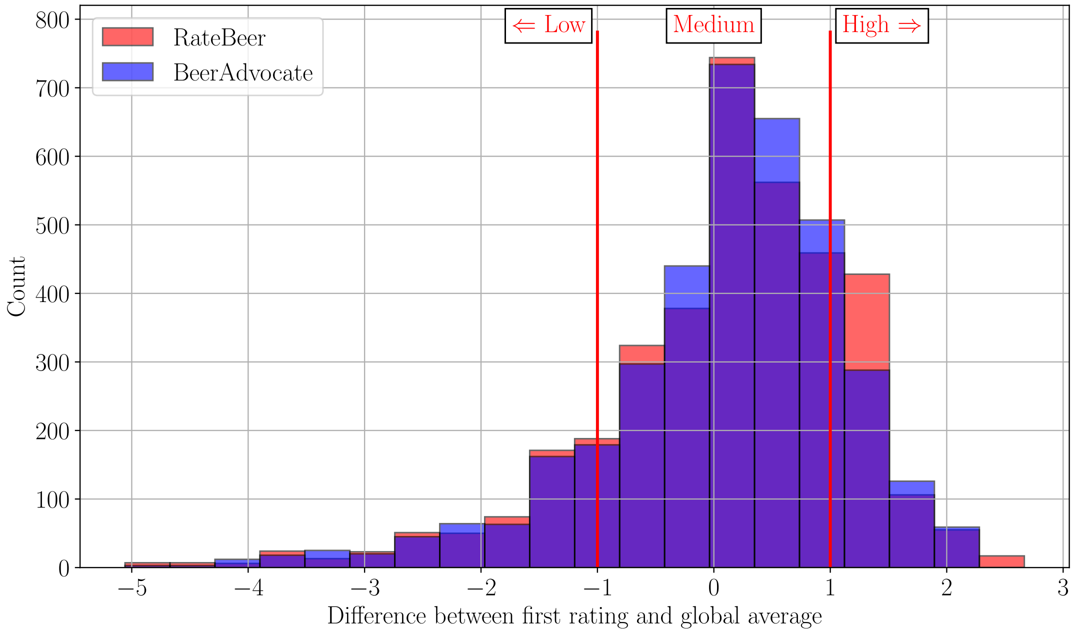


Figure 20: Distribution of difference between the first rating and the global average for BeerAdvocate and RateBeer. The limits for high, medium and low ratings are shown in red.

The peak of differences is around 0 which is the standardize global average. Therefore, it is logical to classify the ratings using the standard deviation. A rating is Low if it is below minus one standard deviation, it is High if it is above one standard deviation, and it is Medium otherwise. It is now possible to classify all the pairs of matched beers between nine different classes (three for BeerAdvocate times three for RateBeer). The matrix of the histogram in 2D with the number of pairs belonging to each region is given in Figure 21.

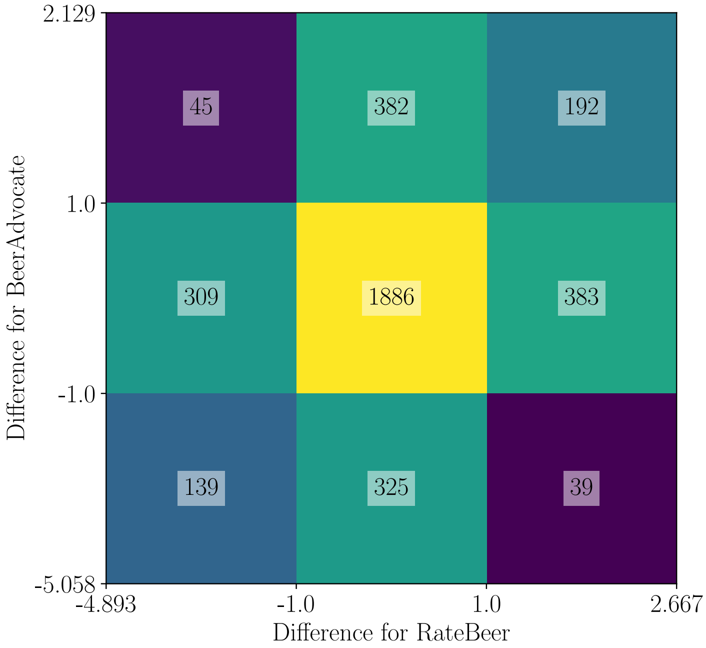


Figure 21: Histogram in 2D for classification of pairs of beers using the difference between the first rating and the global average.

It is now possible to classify the beers between pairs of High (H), Medium (M), or Low (L) first rating. For example, the box with 309 beers and the one with 325 beers are both classified as L-M because the first rating in one of the website is classified as Low and the other one is classified as Medium. The five interesting categories are H-H, H-M, M-M, L-M, and L-L. The analysis will continue on the evolution of the ratings for each of the categories. Figure 22 shows the mean difference to the global average of all the ratings. The data have been bootstrapped by 1000 draws of 500 ratings to compute the 95% confidence interval. There is a correction term after the first rating of the categories H-H (in blue), M-M (in green), and L-L (in red). Indeed, the users certainly tried to make the ratings closer to the inherent quality of the beer. Since the lines are not overlapping, it is possible to infer that the first rating contained the perception of the user as well as the inherent quality of the product. This affirmation is reinforced by the fact that the lines belonging to the categories H-M (in light blue and blue-green) and L-M (in yellow and yellow-brown) are in between the lines of the categories H-H and M-M, and M-M and L-L respectively. Therefore, the difference in classification is implied by the user's opinion. Indeed, if a beer was classified as High on both websites, then it is very likely that its inherent quality is higher than a beer that has received a High and a Medium first rating. However, the most interesting effect is the fact that the line for the beers that have been categorized as High (or Low) in the pair H-M (L-M) is not overlapping with the line of the beer classified as Medium in the same pair. It comes from the fact that the next users are following the opinion of the first user to some extent. Indeed, the beers are the same between these two lines. Thus, the only difference between the ratings of these beers is the first rating. Moreover, since the pairs are composed of a mix between the beers from BeerAdvocate and RateBeer, the effect of the different website is removed.

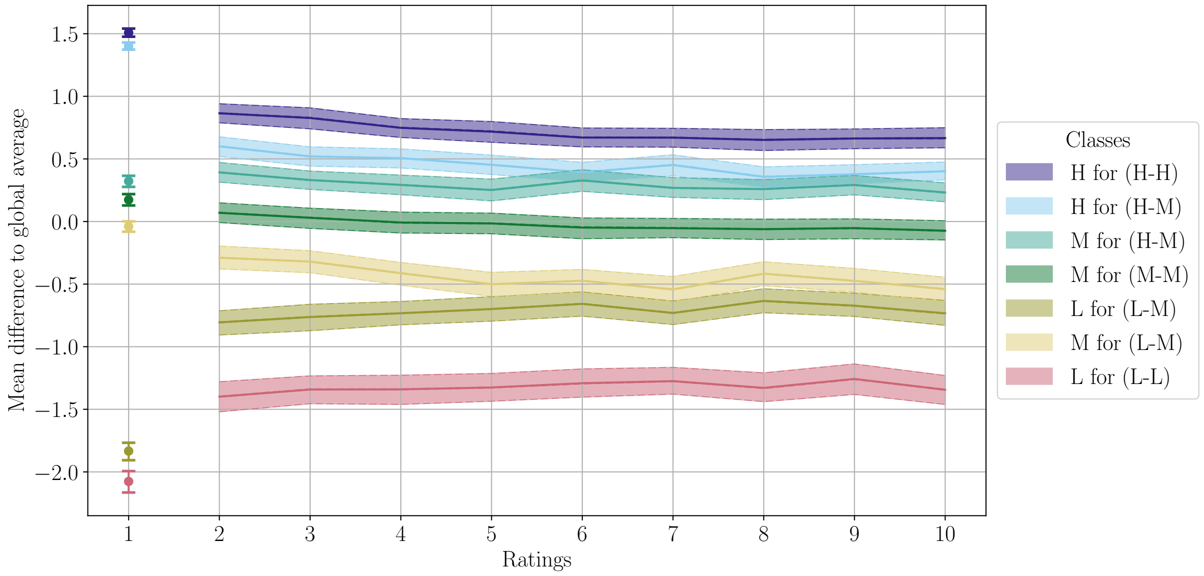


Figure 22: Mean difference to the global average over the ratings. The colors correspond to different categories of classified pairs.

Now that it has been shown that if a user gives a high rating to a beer, the following users will give a slightly better rating, it is normal to ask the following question: Will the users have the tendency to give ratings closer to the product average? If it is the case, then it would simply mean that the very first rating is influencing the average rating of the product. Thus, the average rating of a beer cannot be taken directly as the real inherent quality of this beer.

Therefore, the whole procedure is repeated except that is is using the difference between the i-th rating and the average of the product instead of the global average. The distribution of difference is given in Figure 23. The histogram in 2D used for the classification is given in Figure 24. There are fewer values classified as Low. Therefore, the results with a value classified as Low will be less representative even if the data are bootstrapped, especially the pairs L-L. Figure 25 shows the mean difference to the product average over the ratings. It is clear that the ratings are closer to the product mean as the number of ratings is increased. This means that the users have the tendency to follow the previous users' opinions. They do not directly follow the opinion of the previous users, but they tend to give ratings that are closer to the average of the product. If the information in Figure 22 is added, it is possible to state that the users have the tendency to follow the product average that is influenced by the first rating.

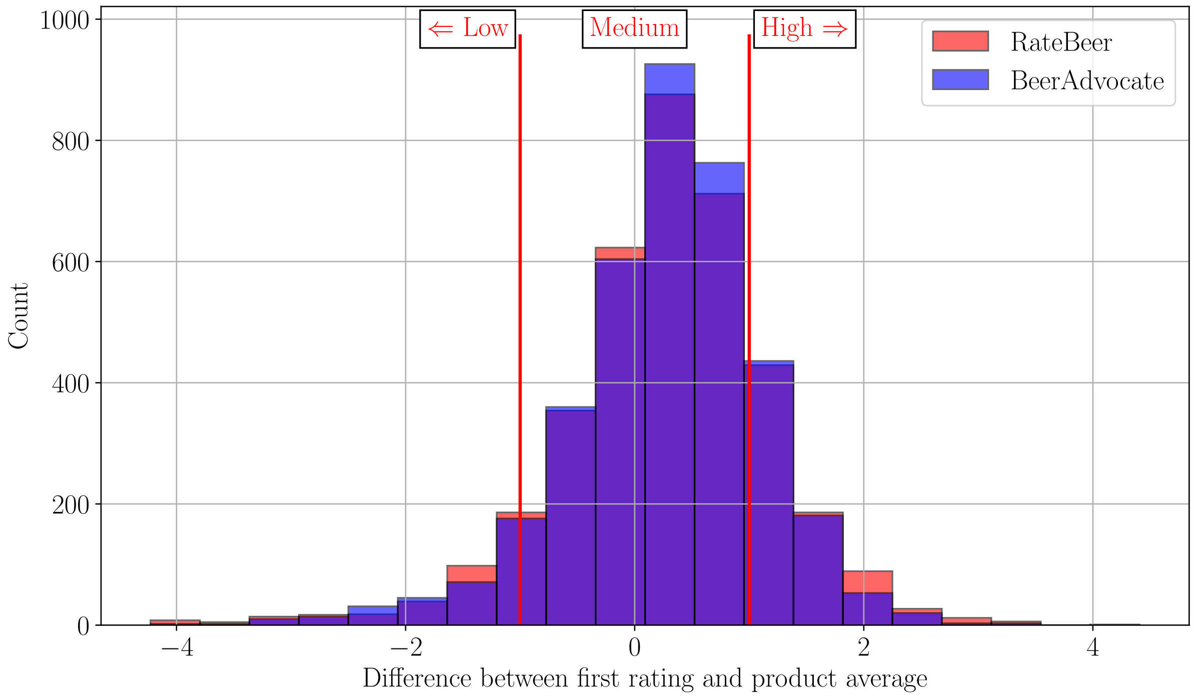


Figure 23: Distribution of difference between the first rating and the product average for BeerAdvocate and RateBeer. The limits for high, medium and low ratings are shown in red.

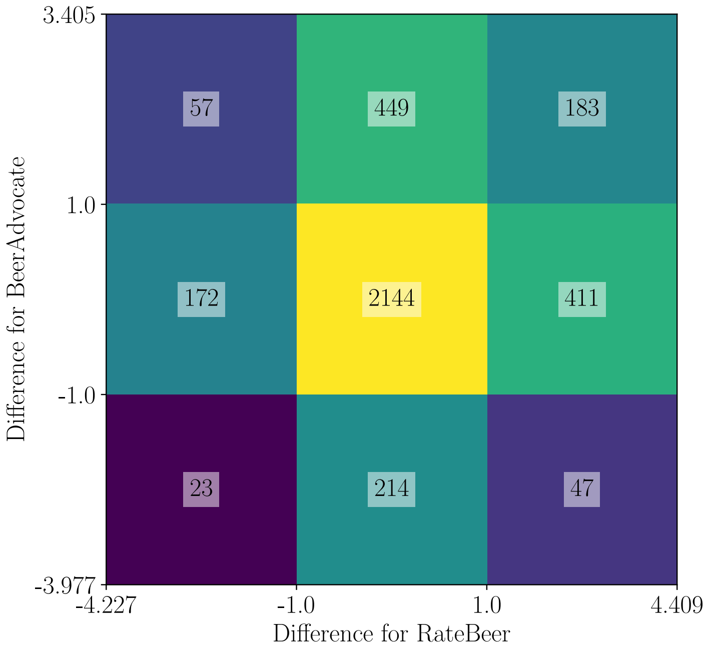


Figure 24: Histogram in 2D for classification of pairs of beers using the difference between the first rating and the product average.

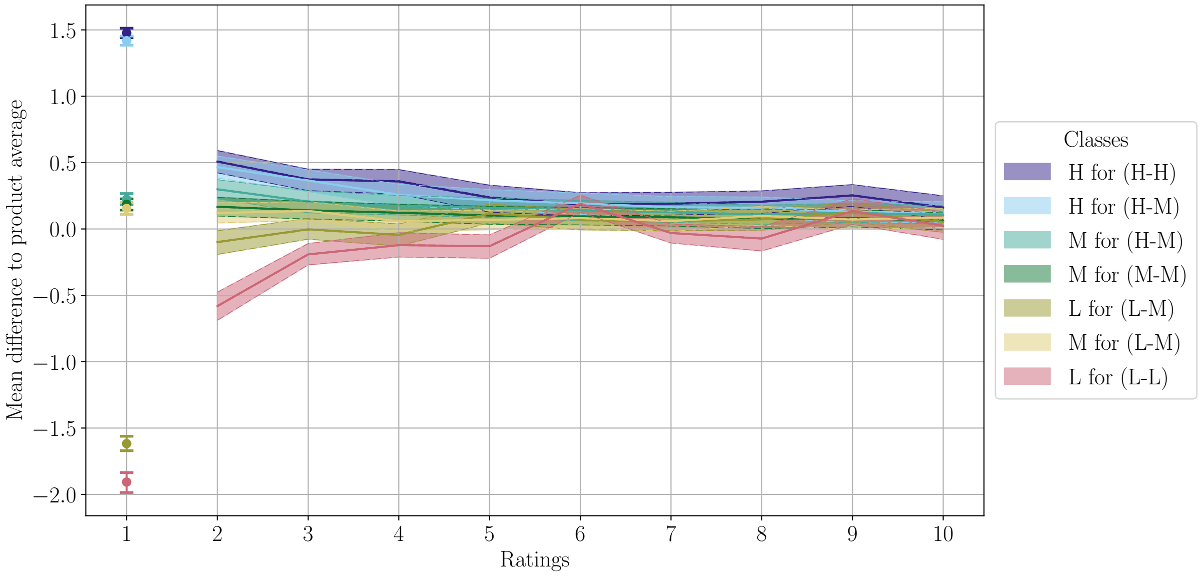


Figure 25: Mean difference to the product average over the ratings. The colors correspond to different categories of classified pairs.

Text Classification and Features Extraction

In this section, something a bit different from the rest of the thesis is explored. Indeed, this section is about text features. The way the users write their reviews can also be used to show if they belong to one community or another. This feeling of belonging to a community helps the users to trust other users in the same community. First, it is needed to extract the text features; then it is possible to predict on which website a given review was written.

Many reviews are required to retrieve the features from them. The equity between the two datasets is also necessary for a good classification algorithm. Therefore, the set of matched beers $\mathcal{I}_{\text{matched}}$ is used. From this set, only the beers with at least ten reviews are selected. It returns 3'942 beers. Then ten reviews per beer are randomly selected for BeerAdvocate and RateBeer. Among these beers, 10% of them are randomly selected to be in the testing set. They will be used later to check the results of the classification. Therefore, there is a total of 70'940 reviews for training and 7'900 reviews for testing, half of them being from BeerAdvocate and the other half from RateBeer.

The model for the text classification and features extraction is based on an article of Benjamin Bengfort [23]. The whole process was modified such that it suits the data and purpose of this section. The model is based on the class `Pipeline` from the package `sklearn.pipeline` [20]. The pipeline is composed of three different steps: the preprocessing, the vectorizer and the classifier. The features are extracted from the vectorizer. For the preprocessing, a class using the package `nltk` [24] has been built. The purpose of the preprocessing is to read the data and transform them such that they become usable by the vectorizer. A document, *i.e.* a string, is given to the preprocessor. The first step is to transform the whole string into tokens. It then transforms the capital letters into lower case letters and removes the punctuation from the tokens. Then, it checks if the given token is in the stopwords; if it is the case, it discards it. Finally, the preprocessor lemmatizes the tokens, and the collection of tokens is returned as a `generator`. An example of a review is given below:

> "*The aroma is nice - sweet, syrupy, and a bit hoppy. The IPA poursdark amberwith a red dark maroon hue. It is malty and hoppy balanced well in flavor with the resiny malts up front and the hops lingering nicely on the tongue.*"

This review is then transformed into the following list of tokens:

```
aroma, nice, sweet, syrupy, bit, hoppy, ipa, poursdark, amberwith,
red, dark, maroon, hue, malty, hoppy, balance, flavor, resiny, malt,
front, hop, linger, nicely, tongue
```

With this example, the work of the preprocessing class is evident. It helps to extract the features after training the whole pipeline.

The second step of the pipeline is the vectorizer. Indeed, to use the common classification techniques in Machine Learning, it is not possible to work with text data. They have to be transformed into "numbers". To do that, the class `TfidfVectorizer` from the package `sklearn.feature_extraction.text` [20] is used. The purpose of this class is to transform a collection of documents or strings into a matrix containing the TF-IDF features. This matrix can be then used by the final step of the pipeline.

The third and last step of the pipeline is the classifier. The algorithm used is implemented in the function `SGDClassifier` from the package `sklearn.linear_model` [20]. This algorithm is here to get the labels, preliminarily transformed into 0 and 1, using the matrix of features from the vectorizer. The default parameters of the different functions were not changed. A long and tedious grid search on all the different parameters available for the vectorizer and the classifier can be applied. However, the results are already satisfying without changing any of these parameters.

The results are given in Table 11 for the classification report and in Figure 26 for the confusion matrix.

	Precision	Recall	F1-score	Support
BeerAdvocate	0.85	0.87	0.86	3'950
RateBeer	0.86	0.84	0.85	3'950
avg / total	0.85	0.85	0.85	7900

Table 11: Classification report provided by the function `classification_report` in the package `sklearn.metrics` on the test data after training on the train data.

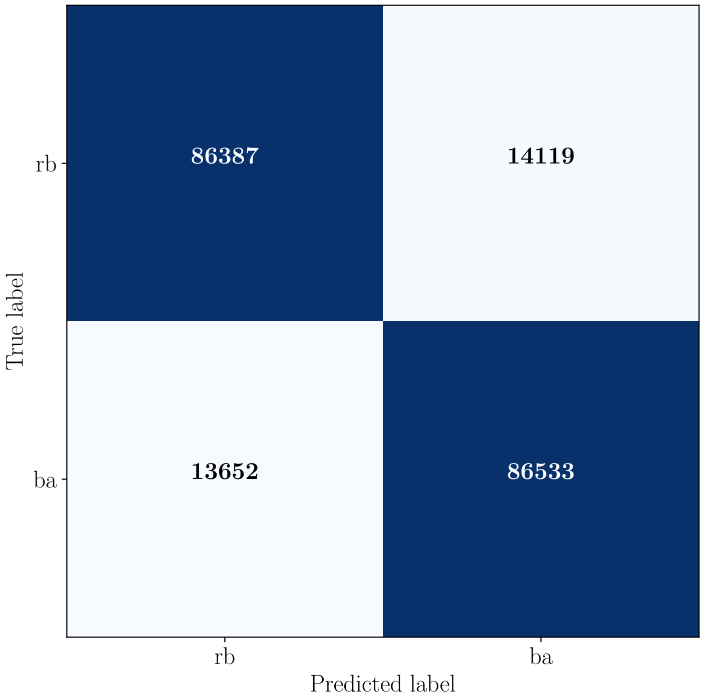


Figure 26: Confusion matrix on the test data after training on the 70'940 matched reviews. The results return 85.443% of correct predictions.

These results are excellent. Indeed, it is clear that even in the text, the community is well developed and far from each other. The interesting thing now is to have a look at the text features, *i.e.* the tokens. Indeed, with these features, it is possible to find an explanation for the good results. The twenty most useful features for each website are shown in Table 12.

	BeerAdvocate		RateBeer	
#	Score	Features	Score	Features
1	-6.229	smell	6.704	aroma
2	-5.807	drinkability	4.467	bottle
3	-5.696	mouthfeel	3.698	draft
4	-5.175	glass	2.233	rating
5	-4.736	carbonation	2.149	palate
6	-4.008	lace	1.943	flavor
7	-3.891	finger	1.782	draught
8	-3.803	retention	1.779	white
9	-3.692	pint	1.720	beige
10	-3.440	beer	1.490	pours
11	-3.273	drinkable	1.437	courtesy
12	-3.243	lacing	1.388	rate
13	-3.128	review	1.349	finish
14	-2.867	leave	1.330	rat
15	-2.748	one	1.297	lightly
16	-2.337	style	1.250	lively
17	-2.230	overall	1.219	bitter
18	-2.223	follow	1.213	soft
19	-2.166	tulip	1.182	nose
20	-2.153	inch	0.959	gold

Table 12: Twenty feature with the highest absolute weights for both BeerAdvocate and RateBeer trained on the 70'940 matched reviews. BeerAdvocate's weights are negative because the label is 0 and the label for RateBeer is 1.

The features are fascinating. For example, the terms `drinkability` and `drinkable`, which are close semantically speaking, have a strong impact on the reviews. It can be interesting to search from where this term comes. If it is written somewhere on the website or if some users came with these words and the community accepted them. In Table 12, two pairs of words are in green and red. These two pairs have the same meaning, but the words are different because they are written in the reviews. It can be seen in Figure 53 in the Appendix that in the review, BeerAdvocate uses the terms `smell` and `feel`. RateBeer uses the terms `aroma` and `palate` in Figure 54 in the Appendix. It shows that the framework, *i.e.* the website, influences the users in some way. Some other words are also interesting to stress out. For example, the word `beer` is the tenth most useful feature for BeerAdvocate. That seems odd because both websites are talking about beer reviews. To make sure that these features are not only linked to these specific subsets of data, one million random reviews are taken from both BeerAdvocate and RateBeer. In these two million reviews, 10% of them are kept for the testing set. The rest is going in the training set. The same pipeline is used, and the model is trained on these data. The results found are fascinating. First, the classification report after training on the 1.8 millions reviews is given in Table 13. The confusion matrix of this model is given in Figure 27. The results are slightly better in average. However, it is quite close to the model trained on the subset of matched data. Also, since the model is specifically trained to learn features, it should not overfit on the training data. Therefore, ten thousand reviews from the training set of 1.8 million reviews are selected, and the classification is done on these reviews. The classification report is given in Table 14. The results are similar to the classification on the testing set. Therefore, this means that the model is not overfitting on training data. Thus, it is possible to use this trained model on the training data without any loss of information.

	Precision	Recall	F1-score	Support
BeerAdvocate	0.87	0.86	0.87	100’506
RateBeer	0.86	0.87	0.87	100’185
avg / total	0.87	0.87	0.87	200’691

Table 13: Classification report provided by the function `classification_report` in the package `sklearn.metrics` on the test data after training on 1.8 millions reviews.

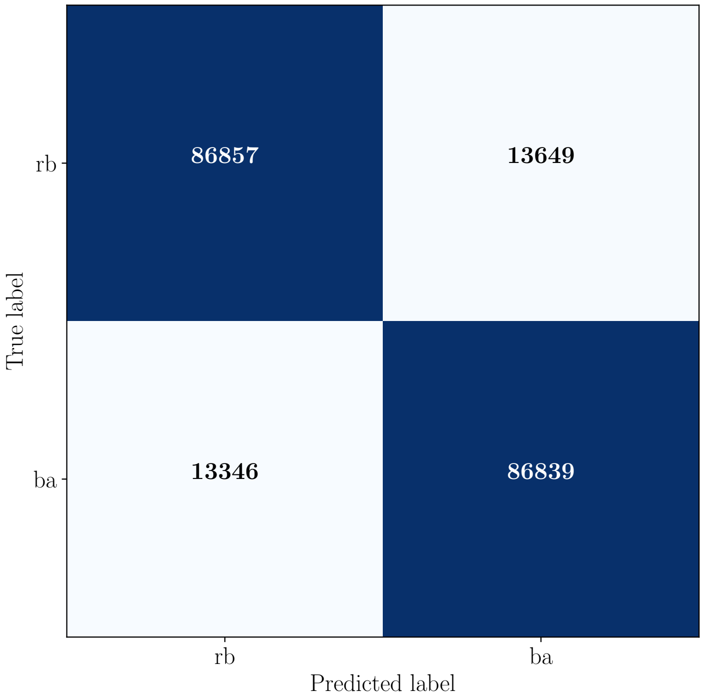


Figure 27: Confusion matrix on the test data after training on 1.8 millions reviews. The results return 86.162% of correct predictions.

	Precision	Recall	F1-score	Support
BeerAdvocate	0.86	0.86	0.86	4’995
RateBeer	0.86	0.86	0.86	5’005
avg / total	0.86	0.86	0.86	10’000

Table 14: Classification report provided by the function `classification_report` in the package `sklearn.metrics` on a subset of the train data after training on the train data.

The twenty most important features for the model trained on the 1.8 million reviews are given in Table 15. Out of the twenty previous features (in Table 12), nineteen from BeerAdvocate and fifteen from RateBeer are still present. It is also interesting to note that the feature `flavour` and `colour` entered the list of twenty most useful features in RateBeer. It means that British English is favored in RateBeer compared to BeerAdvocate.

	BeerAdvocate		RateBeer	
#	Score	Features	Score	Features
1	-6.811	`smell`	6.785	`aroma`
2	-6.082	`drinkability`	4.218	`bottle`
3	-5.373	`mouthfeel`	2.509	`draft`
4	-4.958	`glass`	2.462	`palate`
5	-4.878	`carbonation`	2.291	`rating`
6	-4.559	`lace`	1.735	`flavor`
7	-4.137	`finger`	1.708	`draught`
8	-3.970	`retention`	1.637	`beige`
9	-3.525	`pint`	1.599	`white`
10	-3.463	`lacing`	1.560	`flavour`
11	-3.441	`drinkable`	1.539	`finish`
12	-3.437	`beer`	1.517	`rate`
13	-3.306	`review`	1.498	`rat`
14	-2.903	`leave`	1.305	`courtesy`
15	-2.528	`style`	1.260	`brewpub`
16	-2.449	`one`	1.234	`bitter`
17	-2.403	`abv`	1.160	`il`
18	-2.286	`tulip`	1.084	`cask`
19	-2.134	`foam`	1.071	`colour`
20	-2.127	`inch`	1.043	`soft`

Table 15: Twenty feature with the highest absolute weights for both BeerAdvocate and RateBeer trained on 1.8 millions reviews. BeerAdvocate's weights are negative because the label is 0 and the label for RateBeer is 1. The feature in red are the features that were not the most important one when training on the subset of matched reviews in Table 12.

There is still one question that can be asked. Indeed, the websites provide directly the features `smell` and `mouthfeel` (`feel`) for Beeradvocate and `aroma` and `palate` for RateBeer in their rating systems. Therefore, it is interesting to know if it is possible to classify the reviews using only these four features. The test data from the two million reviews are greedily classified by adding a weight of +1 for the two features in RateBeer and -1 for the two features in BeerAdvocate. Then, if the sum of weight is positive, the review is classified as coming from RateBeer; if it is negative, it is classified as coming from BeerAdvocate; if it is zero, it is undefined. The confusion matrix is given in Figure 28.

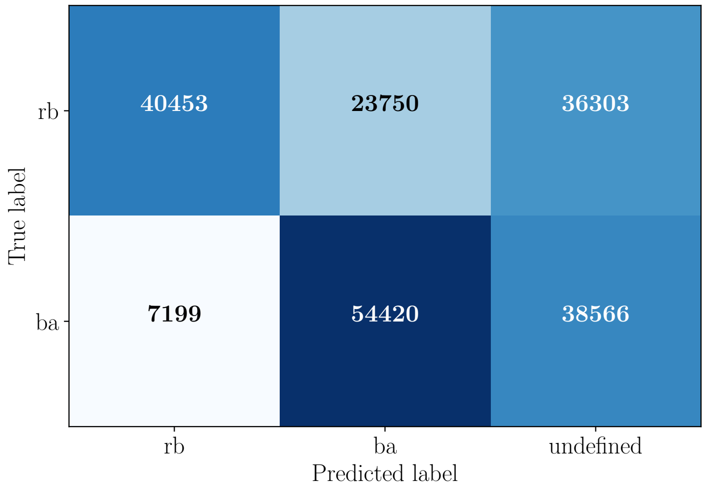


Figure 28: Confusion matrix on the test data using only the two features present in the rating system of each website. The results return 47.273% of correct predictions.

The results are not good. It is not possible to predict more than half of the reviews correctly. That means that using only these two features is not the solution. Therefore, the affiliation of the users to one of the community by the way they write is not just due to the influence of the rating systems. It is reasonable to state that the users are looking at the previous reviews to get some inspiration for their reviews or how to use the right vocabulary.

Communities

In this section, the users and the way they belong to their community is investigated. The idea is to use the model developed in Section Text Classification and Features Extraction to classify the users. It gives an idea about how close the users are to their communities. Then, a further investigation on the users in the set $\mathcal{U}_{\text{matched}}$ can be made.

Text and Community

First, the sets of all the users from BeerAdvocate $\mathcal{U}_{\text{BA}}$ and RateBeer $\mathcal{U}_{\text{RB}}$ are used. From these two sets, all the matched users in $\mathcal{U}_{\text{matched}}$ are removed. Then, the reviews of these users are classified using the model developed in the previous section, see Section Text Classification and Features Extraction. If the user has a majority of reviews classified as BeerAdvocate, then he belongs to the BeerAdvocate community on the writing aspect. If he has a majority of reviews classified as RateBeer, then he belongs to RateBeer. If the number of reviews is the same between BeerAdvocate and RateBeer, then this user belongs to neither of the communities. The confusion matrix is given in Figure 29.

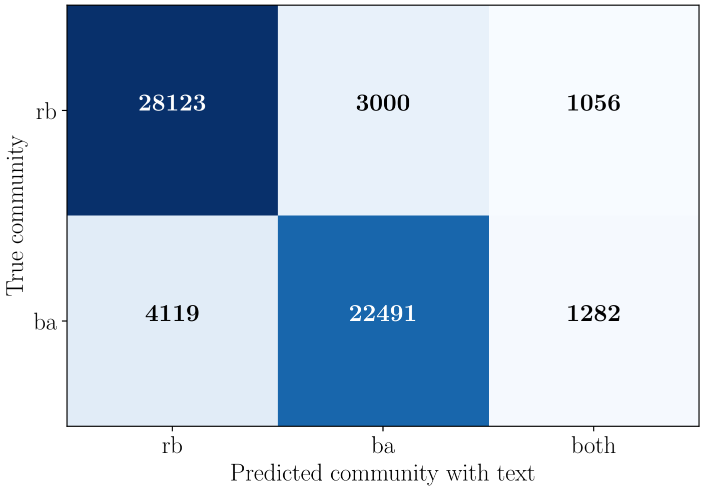


Figure 29: Confusion matrix on the users present in only one website. The prediction on the community has been done using the model developed in Section Text Classification and Features Extraction. The final prediction has been done using the number of reviews classified in each website. The predictions return 84.257% of correct prediction.

The results are pretty good. Indeed, more than 84% of the users can be linked to their communities by the way they write. It is also possible to apply the model for the text classification to the reviews of the users who are in both communities $\mathcal{U}_{\text{matched}}$. This classification can also be linked to the joining date and the number of beers rated on both websites. For the joining date, if a user joined both websites in less than a month, it is classified as coming from both communities. Otherwise, the first joining date is linked to its community. For the number of beers, if the absolute difference in the number of reviews is less than 10% of the minimum number of reviews from one of the two websites, then the user is classified as coming from both communities. Otherwise, the website in which the user rated more beer is his community. The results are given in the confusion matrix in Figure 30.

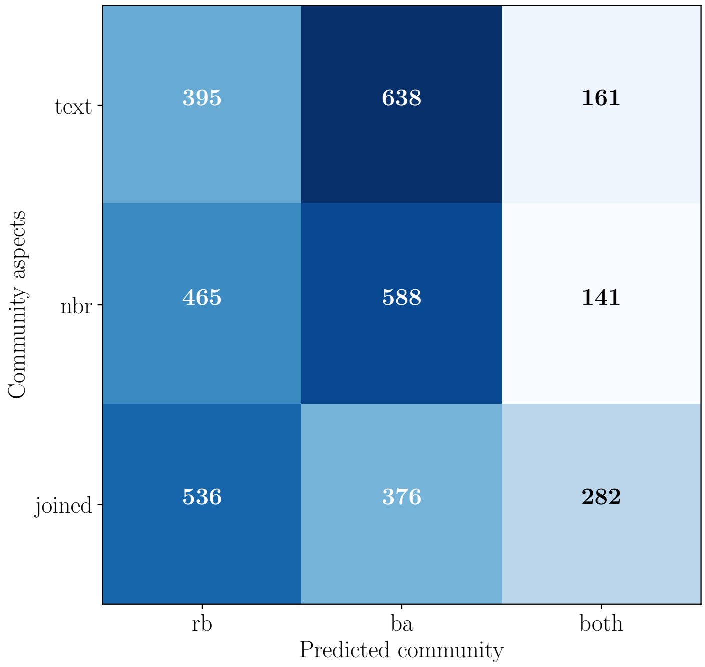


Figure 30: Confusion matrix on the users present in both websites. The predictions have been done on the reviews using the model developed in Section Text Classification and Features Extraction, on the joining date and on the number of reviews done.

The results in Figure 30 are difficult to interpret. The numbers are not consistent between each row. However, the percentage of users who are either in BeerAdvocate or RateBeer for the three categories is 8.710%. This proportion is minuscule. It shows that the users who are in both datasets are taking inspiration from both of them, especially if they spend much time on both.

Users in both Communities

The previous subsection has shown that if the users are spending time on both websites, then it is hard to link them to a specific community. However, it is possible to investigate their behavior in more details. Therefore, the set of matched users and matched beers $\mathcal{R}_{\text{matched}}$ is used, and some results are inferred from this set.

The first investigation is the difference in the ratings between the two websites on the same beer. The logical hypothesis if that the difference of ratings between the beers on both websites is close to 0. The difference between the mean rating of each of the matched beers between BeerAdvocate and RateBeer is used to have a comparison to make. The results are given in Figure 31 for the histograms and in Table 16 for some statistics. We note that the users have the tendency to give a rating on both websites closer to the mean of the given beer. However, it is also interesting to see that this difference is not zero in median and average. It confirms again the fact that each community has a norm. However, this could also come from the differences in the rating systems. However, since around 20% of the differences are equal to 0 or very close to 0, the users who wanted to give the same rating could do it.

	Same beer, same user	Same beer, mean value
Max	3.480	1.944
Median	**0.100**	**0.364**
Min	-1.500	-1.365
Mean	**0.143**	**0.364**
STD	0.247	0.286

Table 16: Difference between BeerAdvocate and RateBeer. The left column is computed using the ratings of a given user on a given beer on both websites. The right column is computed using the average ratings of a given beer present in both websites.

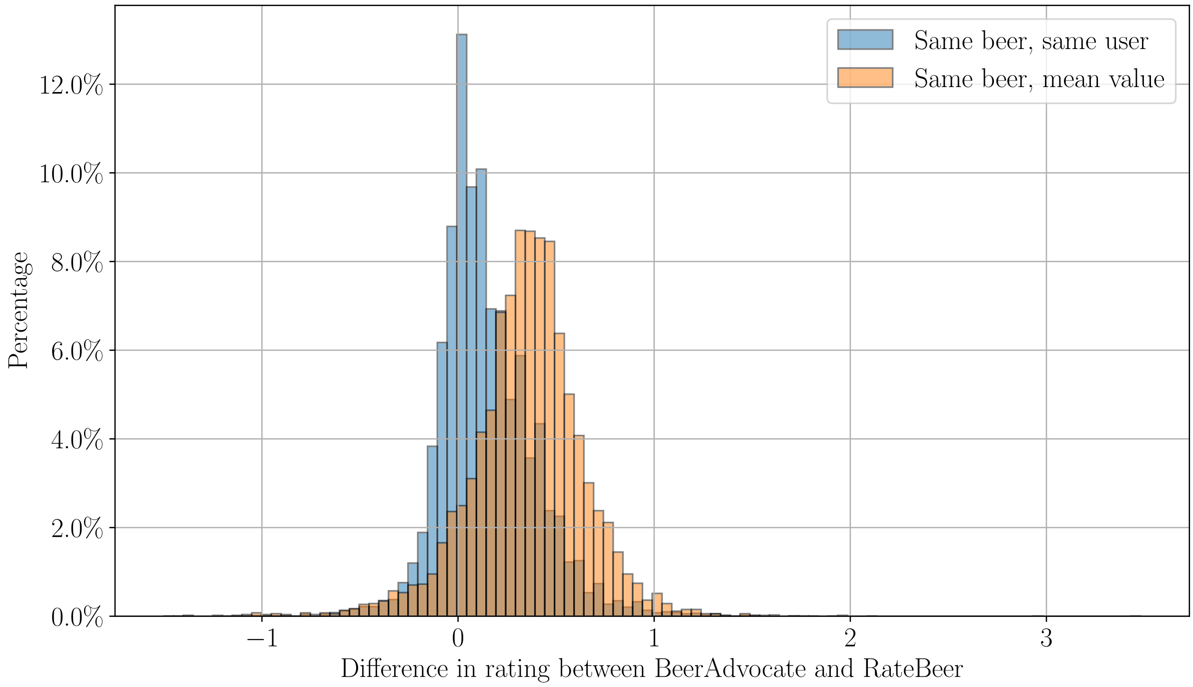


Figure 31: Distribution of difference between BeerAdvocate and RateBeer. The blue histogram is computed using the ratings of a given user on a given beer on both websites. The orange histogram is computed using the average ratings of a given beer present in both websites. Both histograms have 100 bins.

The text of the reviews is also interesting. Indeed, it was just shown that the users in both communities have the tendency to give close ratings between the same beer in BeerAdvocate and RateBeer. Since it is not exactly 0, the texts should not be the same between both websites. Therefore, the function `ratio` from the package `Levenshtein` is used between the reviews of each user on the same beer between BeerAdvocate and RateBeer. This ratio is also compared with the ratio between two random reviews, one from BeerAdvocate and one from RateBeer. The results are given in Figure 32 for the histograms and in Table 17 for some statistics. The results are fascinating. Indeed, the average ratio is around 0.3 for the random reviews. It is an expected number since the users are talking about beers on both websites; they are using the same kind of vocabulary. On the other hand, the ratio is extremely high for the users who wrote about the same beers on both websites. Indeed, the median is almost at 1. It means that around 50% of the users simply copy-pasted their reviews between both websites without changing a letter in their reviews.

	Same beer, same user	Random reviews
Max	1.000	0.519
Median	**0.992**	**0.364**
Min	0.000	0.000
Mean	**0.925**	**0.346**
STD	00.154	0.073

Table 17: Levenshtein ratio between reviews from BeerAdvocate and RateBeer. The left column is computed using the reviews of a given user on a given beer on both websites. The right column is computed using random reviews, one from each website.

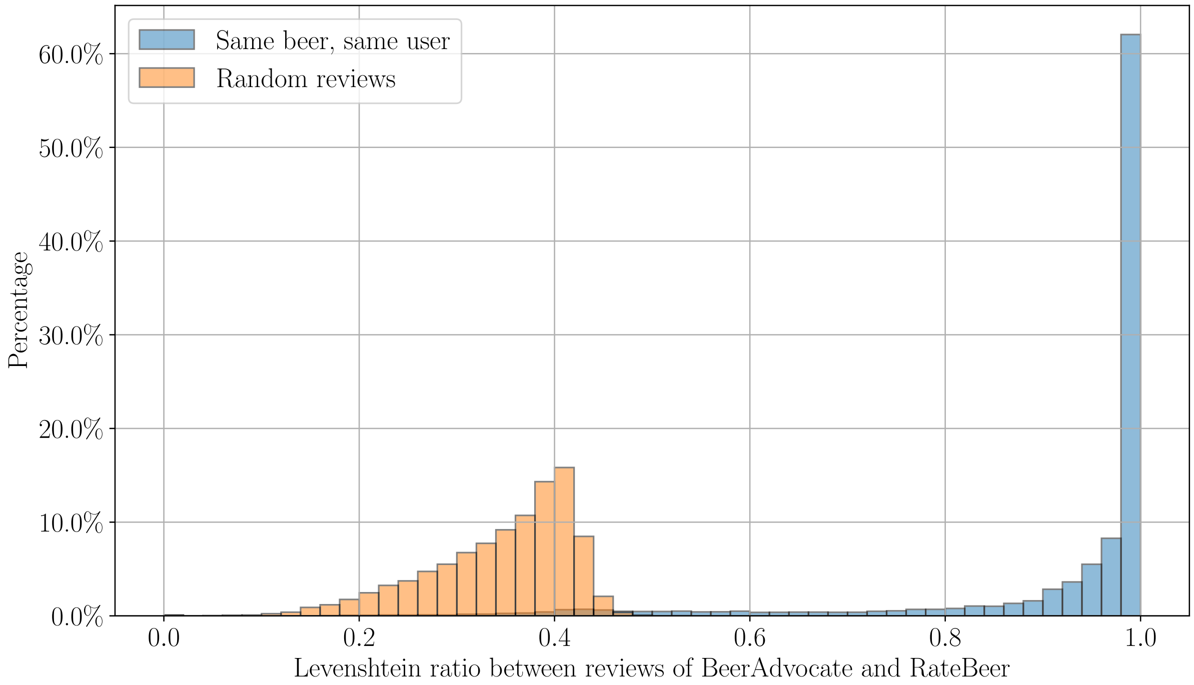


Figure 32: Distribution of Levenshtein ratio between reviews from BeerAdvocate and RateBeer. The blue histogram is computed using the reviews of a given user on a given beer on both websites. The orange histogram is computed using random reviews, one from each website. Both histograms have 50 bins.

Linking this result with the previous one is interesting. Indeed, the users have the same opinion expressed on both websites. It means that if they liked the beer on one website, they also liked it on the second website. However, they do not give the same rating on both websites. It is true that the rating systems are different and it would require a few tries before finding the right combination leading to the same final rating. However, only 4.77% of the users give the exact same rating, and 38.71% of the users give two ratings with a difference smaller than 0.1. It shows that most of the users give different ratings for the same beer on both websites. Therefore, it implies that the users have only one opinion and they are simply trying to adapt it to the community norm. In Subsection Examples of Reviews in the Appendix, three pairs of reviews are shown, each pair from a different user and a different beer. These three examples are the typical examples of what can happen when the users are reviewing the beers on both websites. The first two reviews, in Figure 68, show the reviews of a user called `roborb` on the beer called *Duck-Rabbit Paul's Day Off.* This user posted his reviews on the same day, 11th of April 2009. He simply copy-pasted the text and tried to give very similar ratings. This the example of someone who tries to be fair between both websites. Because of this fairness, he is fifteen percent under the average rating of this beer on BeerAdvocate. The second pair of reviews, in Figure 69, show the reviews of a user called `redave` on the beer called *Indian Brown Ale.* This user also posted both of his reviews on the same day, 19th of May 2006, and copy-pasted the text. However, the two ratings are different. We see that for BeerAdvocate, this user gives a review that is twelve percent under the average. It is possible to do the calculus for RateBeer as well. When the screenshot was taken, the review of this beer was 3.73/5. His rating is fifteen percent under the average. We see that this deviation is consistent between the two websites. Therefore, this user is trying to adapt his rating to the norms of the communities. The last pair of reviews, in Figure 70, show the reviews of a user called `kindestcut` on the beer called *Saint Arnold Brown Ale.* The behavior of this user is entirely different from the two previous users. He did not post his reviews on the same day, and he did not simply copy-paste the texts. The function `ratio` from the package `Levenshtein` returns a ratio of 0.64, which is much higher than the comparison between random reviews, see Figure 32. Therefore, he took his text and modified it. He first posted the review on BeerAdvocate and seemed to appreciate this beer. His rating is eight percent above the average rating. However, a few months later, he wrote his review on RateBeer. He changed his opinion from a “very good beer” to a “good beer”. This is really interesting because his rating on RateBeer is still higher than the average rating of this beer. But this user clearly changed his mind about this beer and that is why he did not copy-paste his review from BeerAdvocate.

After finding out that the users tend to copy-paste their reviews on both websites, it is interesting to investigate the sets of beers they rate on both websites. One of the most useful index to analyze the disparity of two sets is the Jaccard index. For two sets $\mathcal{A}$ and $\mathcal{B}$, it is defined as:

$$J(\mathcal{A},\mathcal{B}) = \frac{|\mathcal{A}\cap\mathcal{B}|}{|\mathcal{A}\cup\mathcal{B}|} = \frac{|\mathcal{A}\cap\mathcal{B}|}{|\mathcal{A}|+|\mathcal{B}|-|\mathcal{A}\cap\mathcal{B}|} \tag{3}$$

An alternative version of the Jaccard index called the minimum Jaccard index can be used. For two sets $\mathcal{A}$ and $\mathcal{B}$, it is defined as:

$$J_{\min}(\mathcal{A},\mathcal{B}) = \frac{|\mathcal{A}\cap\mathcal{B}|}{\min\left(|\mathcal{A}|,|\mathcal{B}|\right)} \tag{4}$$

The minimum Jaccard index is used because it happens that the number of beers rated on one of the website is huge. However, if all the beers in the second website have already been rated in the first website, then the Jaccard index will be really small. However, the minimum Jaccard index will be high. These two similarity metrics can give valuable insights on the users' behavior. The results of these two metrics on the 552 users in $\mathcal{R}_{\text{matched}}$ are given in Figure 33 for the histograms and in Table 18 for some statistics. The Jaccard index is low in average. However, the minimum Jaccard index is really high, the median being 1. Therefore, the conclusion is that the users have the tendency to write a lot of reviews on one of the two websites and then they rate some of the beers they already rated on the other websites. That is why a lot of the users were classified between `ba` and `rb` instead of `both` in Figure 30.

	Jaccard Index	Minimum Jaccard Index
Max	1.000	1.000
Median	**0.200**	**1.000**
Min	0.001	0.111
Mean	**0.342**	**0.898**
STD	0.347	0.202

Table 18: Difference between BeerAdvocate and RateBeer. The left column is computed using the ratings of a given user on a given beer on both websites. The right column is computed using the average ratings of a given beer present in both websites.

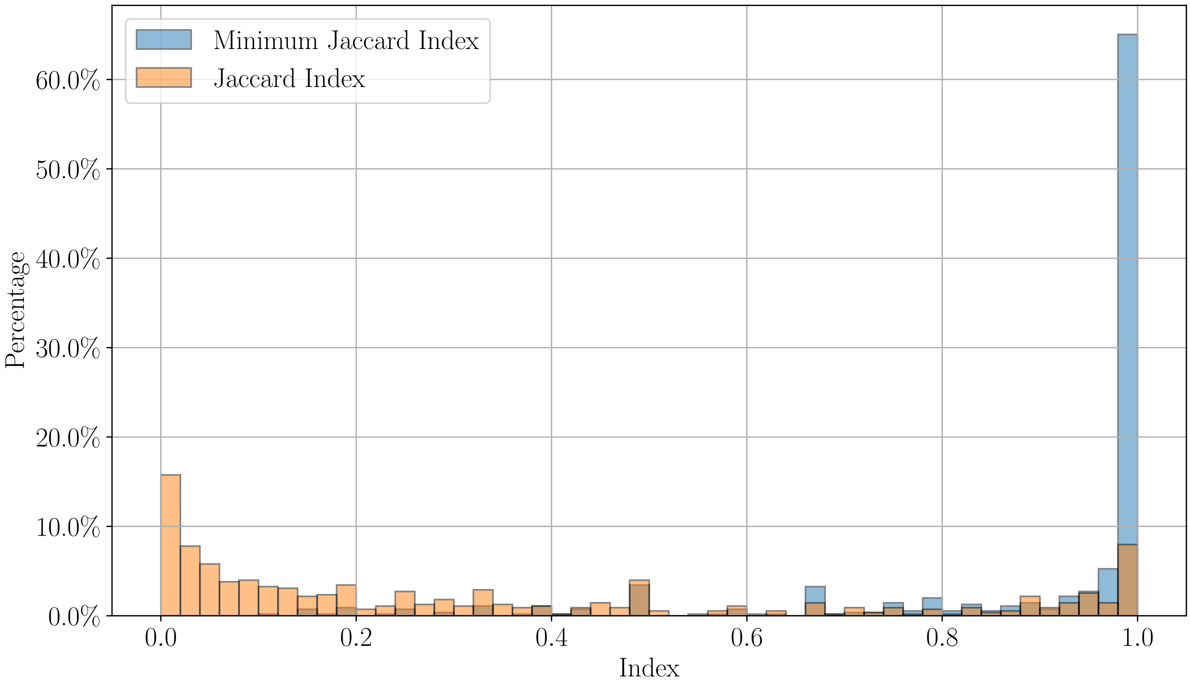


Figure 33: Distribution of Levenshtein ratio between reviews from BeerAdvocate and RateBeer. The blue histogram is computed using the reviews of a given user on a given beer on both websites. The orange histogram is computed using random reviews, one from each website. Both histograms have 50 bins.

Finally, it is interesting to investigate the correlation between the number of beers rated on both websites by a user and the average Levenshtein ratio for his reviews. The results are given in Figure 34. The correlation is quite difficult to see. However, the users rating a high number of beers between both websites have the tendency to copy-paste their reviews more often. Indeed, if a user rarely rates beers on both websites, the he has the time to write a new review.

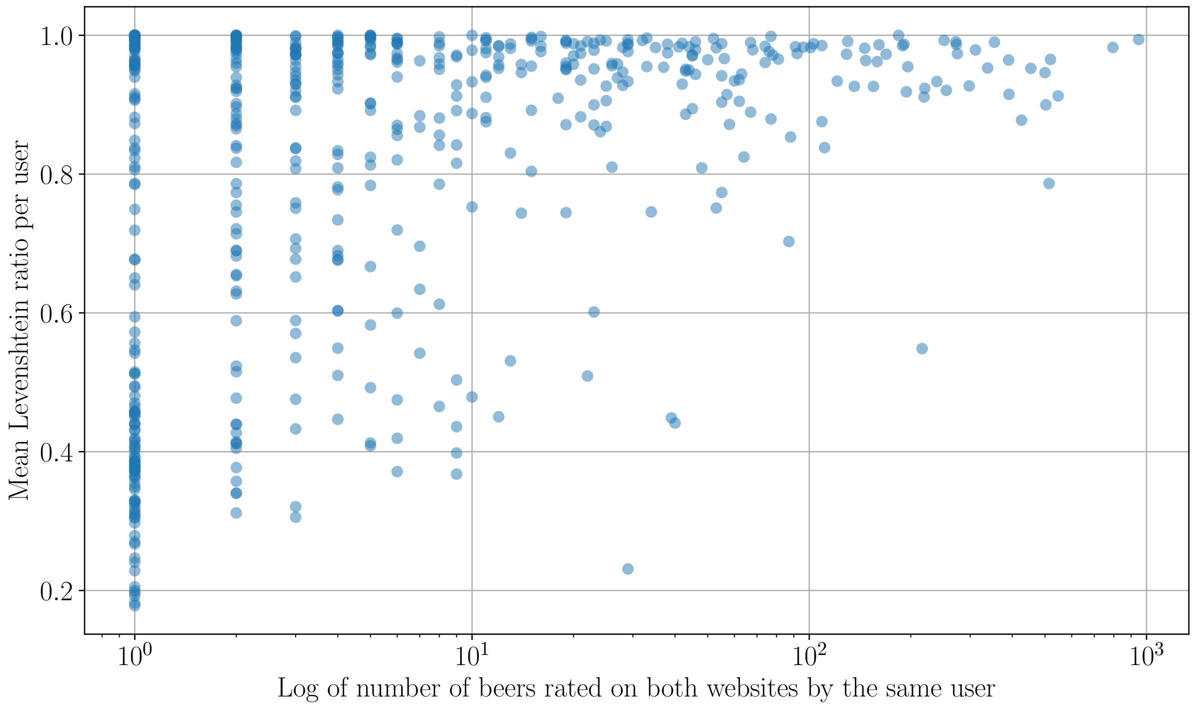


Figure 34: Average Levenshtein ratio in function of the number of beers rated on both websites. Each point correspond to one of the 552 users in $\mathcal{R}_{\text{matched}}$.

All of the results in this section show that either the users belong to only one community and they adapt their behavior to the community or they belong to both, and they take influences from both of them. It is interesting to see that their opinion is the same since they are mostly copy-pasting their reviews. However, the ratings they give differ, most of the time, between the two communities. So, these users are either trying to adapt their ratings to the communities norms, or they have some troubles with the different scaling of the two websites.

Same User, same Beer, two Websites

In the previous section, see Section Communities, the behavior of the users in both communities was investigated. In this section, the analyses concentrates on the difference in ratings. The purpose of this section is to find some of the reasons that could lead to this difference. Then, a Machine Learning algorithm is used to predict the ratings of the users on the second websites given their ratings on the first website.

Correlations of the Ratings

In this section, the 18'109 pairs of reviews from the 552 users in the set $\mathcal{R}_{\text{matched}}$ are used. It is important to make sure that the reviews are the same, *i.e.* the reviews when the user changed his mind have to be discarded. However, it would not be smart to remove the reviews based on the ratings. Therefore, all the reviews having a Levenshtein ratio less than 0.8 between the texts on both websites are removed. The final set of is composed of 15'968 reviews. Figure 35 shows the final rating on RateBeer in function of the final rating on BeerAdvocate. In a perfect world, the ratings should only be on the red line. However, the points have the tendency, in average, to be under the red line. It means that the BeerAdvocate ratings are a bit higher than the RateBeer ratings. However, these ratings are not random. Indeed, the Pearson's correlation coefficient is above 0.9.

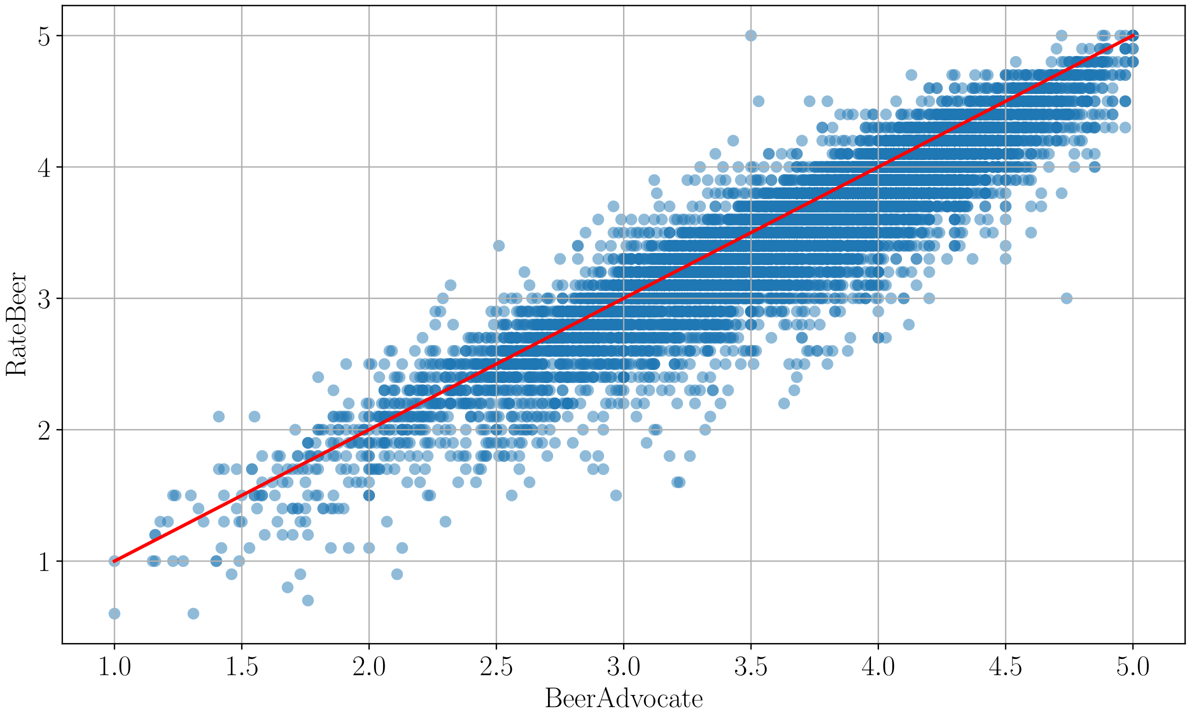


Figure 35: RateBeer rating in function of the BeerAdvocate rating for the users with the same review between the two websites. The red line corresponds to the same rating between RateBeer and BeerAdvocate. Pearson's correlation coefficient is 0.923.

For each of the aspects, the correlation matrices are given in the Appendix, Subsection Correlation of the Aspects. Only the Pearson's correlation coefficient are recalled in Table 19. The ratings are well correlated with a minimum coefficient of 0.794. It is interesting to note that all of the aspects are less correlated than the final rating. The two highest correlations are for aroma and taste. These two aspects are the closest between both websites regarding the number of possible ratings, see Table 3. The least correlated aspect is overall. In Table 3, this is the most different between the two websites compared to the others (different in the number of possibilities and scale). Therefore, this result indicates that the scaling is a major factor regarding correlation when a user is rating a beer on both websites.

Aspects	Pearson's r
Apperance	0.818
Aroma	0.862
Overall	0.794
Palate	0.816
Taste	0.886

Table 19: Pearson's correlation coefficient between the ratings on BeerAdvocate and on RateBeer for all the aspects. The ratings are taken from the users rating the same beer on both websites.

While analyzing the matrices in the Appendix in more details, one can note that the overall aspect, see Figure 73, is highly uncorrelated for the high and low ratings. Indeed, the users have the tendency to give a lower rating on RateBeer when they gave the maximum rating on BeerAdvocate. It is known that giving a 20 out of 20 is much harder than giving a 5 out of 5. It is the same for the low ratings. In Figure 75 for the taste, there is a strange combination. Indeed a user gave a 1 on BeerAdvocate and a 10 on RateBeer for the same beer. The ratings of each of the aspect for this review are given in Table 20. In this example, the user gave the same ratings for both Appearance and Aroma.

Aspects	BeerAdvocate	RateBeer
Apperance	4.0	4.0
Aroma	1.5	1.5
Overall	1.0	0.25
Palate	3.5	3.0
Taste	1.0	5.0
Final rating	1.55	2.1

Table 20: Example of a review with one of the aspects being extremely uncorrelated. The ratings of RateBeer are scaled with a maximum of 5.

The text of this review is given below. A specific part of the text is highlighted in bold.
"*Well-bodied and glassy ruby red look. One finger of rose-white head. Nice, deliberate carbonation. Looks really nice. Very very sour aroma, but in a metallic, cherry medicine way. Pure cough syrup. Downright unappetizing. Hint of grapefruit/citrus. Severely sour malt taste. Puckery but not in a good way. Like a sweet tart dipped in Robitussin. Sour, cloying saccharine cherry.* **Not a good tasting beer.** *The feel is creamy and fizzy. Good enough. Not a good beer, sorry to say. I could finish about half and then I had to dump the rest down the drain. Too bad really.*"

This user did not like the taste of this beer. Moreover, since this review is the same on both websites, it is difficult to understand why this user gave a 5 for the taste on RateBeer. This review was first posted on BeerAdvocate and 141 days later on RateBeer. It is possible that this user was trying to get a similar rating on both websites. However, he was influenced by the fact that the ratings are higher on RateBeer. Therefore, he had to make a choice on which aspects to increase. This behavior leads to the last subsection Predicting the Ratings on the Second Website where a Machine Learning model is used to predict the ratings of the users on the second website after the user gave his ratings on the first website.

Randomness between the two Websites

Before using any Machine Learning model, a preliminary investigation of the randomness in these ratings is conducted. The Machine Learning model will help with this, but it is also possible to verify the randomness in another way. A Multinomial distribution is matched to the differences between both websites for each aspect to inject some randomness in the ratings the users give. The differences are computed by subtracting the rating on RateBeer to the rating on BeerAdvocate. The Multinomial distributions are given in the Appendix, see Subsection Multinomial Distributions of the Aspects' Differences. They are computed by counting the differences and creating the probabilities of each of the differences. Then, the function `multinomial` from the package `numpy.random` [19] is used. Then, for each of the aspect, the difference is randomly drawn, and the randomized final rating is computed for BeerAdvocate. The distribution of the final rating with the real values on BeerAdvocate and the one computed using the Multinomial distributions are given in Figure 36.

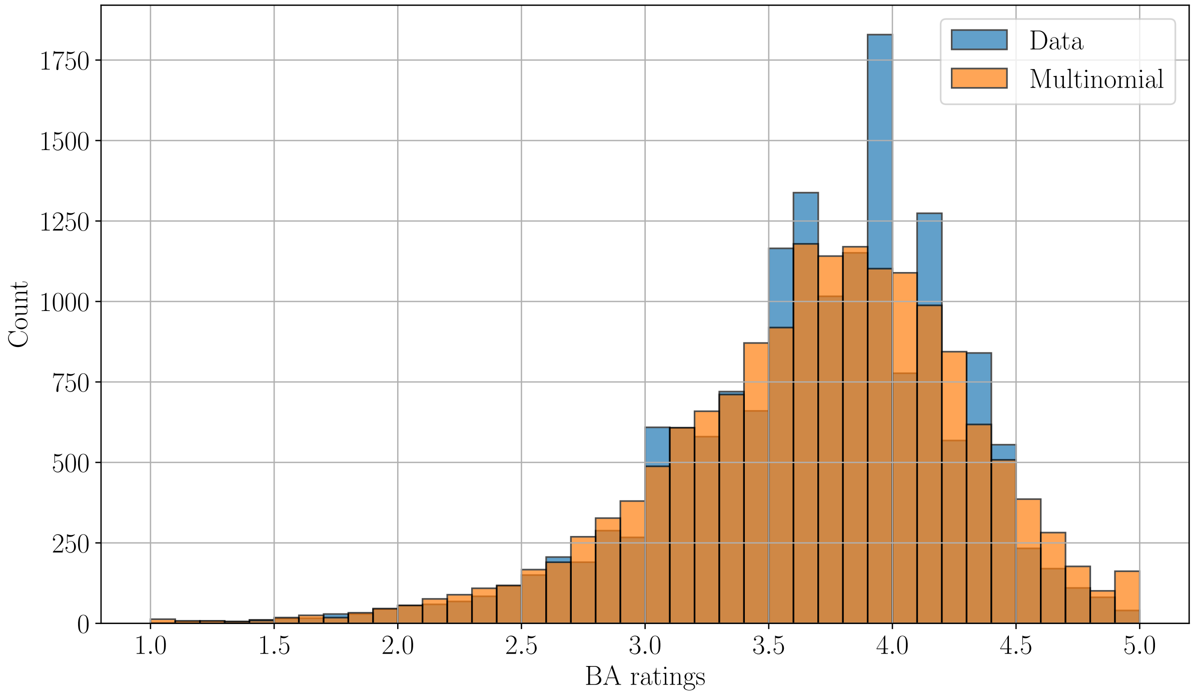


Figure 36: Distribution of the final rating on BeerAdvocate with the distribution computed using the ratings on RateBeer and the Multinomial distributions for the differences.

The shape of the two distributions are quite close. The distribution computed with random differences is a bit smoother than the real distribution. The next step is to calculate the boxplots of the differences between the actual ratings on BeerAdvocate and the random ratings. The average difference should be close to 0. The boxplot of the differences between the ratings on RateBeer and BeerAdvocate should be close to the boxplot of the differences between the ratings on RateBeer and the random ratings for BeerAdvocate. They should have a similar average. These three boxplots are given in Figure 37.

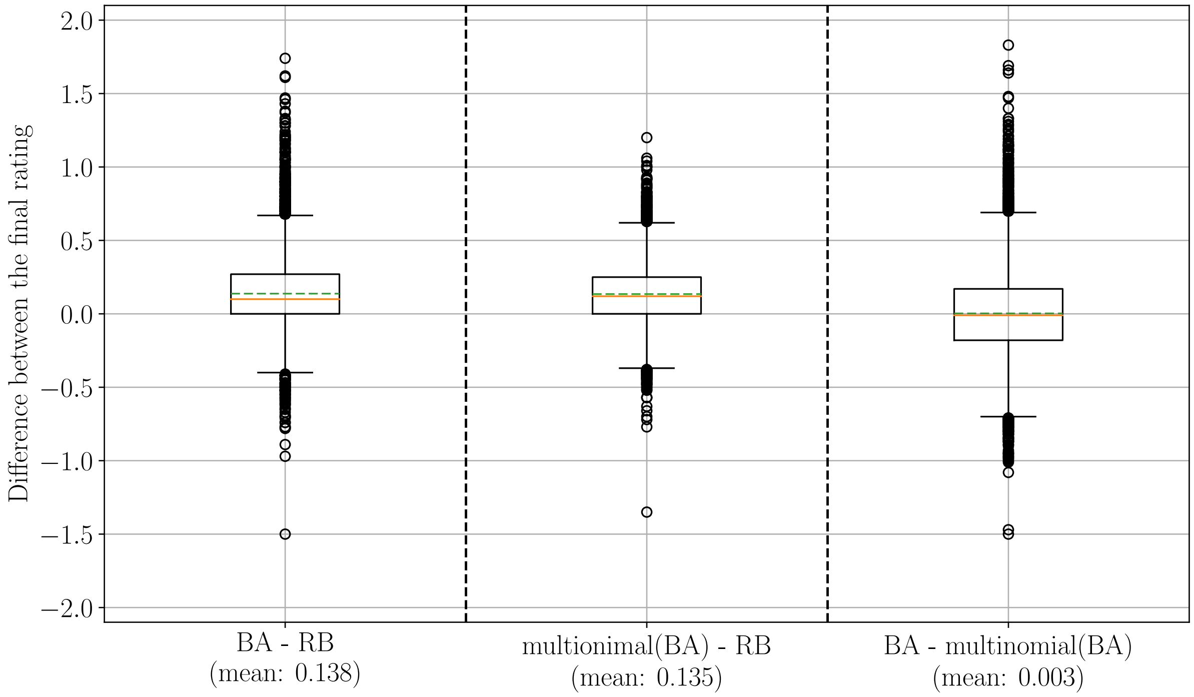


Figure 37: Boxplots of differences between different ratings. The boxplot on the left gives the differences between the ratings on RateBeer and BeerAdvocate. The boxplot in the middle gives the differences between the ratings on RateBeer and the random ratings from BeerAdvocate. The boxplot on the right gives the differences between the ratings on BeerAdvocate and the random ratings from BeerAdvocate.

The results are congruent with our hypotheses. Indeed, the first two boxplots (left and middle) have a similar average difference. The one in the middle is a little bit less wide than the one on the left. It is surely due to the smoothness of the distribution seen in Figure 36. For the boxplot on the right, the average is close to 0, even if the differences can be quite high with the biggest difference being around 2.
It is fascinating to see that the same kind of results can be achieved while the differences between the two websites are randomly drawn. It implies that if only the final ratings is used, some information is lost because of the randomization on the aspects. Therefore, it is important to work with the aspects for the Machine Learning model.

Time Differences

For the Machine Learning model, it is imperative to be consistent with the direction of the ratings. In other words, if a user rated first on RateBeer and then on BeerAdvocate, the Machine Learning model has to work in this specific direction. Therefore, the number of reviews going in each direction is computed. Table 21 gives these numbers for each direction.

Directions	Number of reviews
BeerAdvocate ⇒ RateBeer	3'847
RateBeer ⇒ BeerAdvocate	12'121

Table 21: Number of reviews per direction. The total number of reviews is 15'968.

There is a huge inequality. However, it is impossible to do anything about this except taking the same number of points in each group for the Machine Learning model. However, it could be better to restrict the time difference between the two reviews to less than 24 hours. It is done to make sure that the users still remember what they wrote on the first website. Table 22 gives the number of reviews for each direction that have been done in less than a day.

Directions	Number of reviews
BeerAdvocate ⇒ RateBeer	209
RateBeer ⇒ BeerAdvocate	5'042

Table 22: Number of reviews per direction with a time difference of less than a day.

The problem here is that there is not enough data for the direction "BeerAdvocate ⇒ RateBeer". It is quite odd to see that for one of the directions there is around 40% of the reviews that have been written on both websites in less than a day, and for the other direction, this percentage is only at around 5%. Figure 38 gives the boxplots of the average number of days for each user between the ratings on both websites for the two directions.

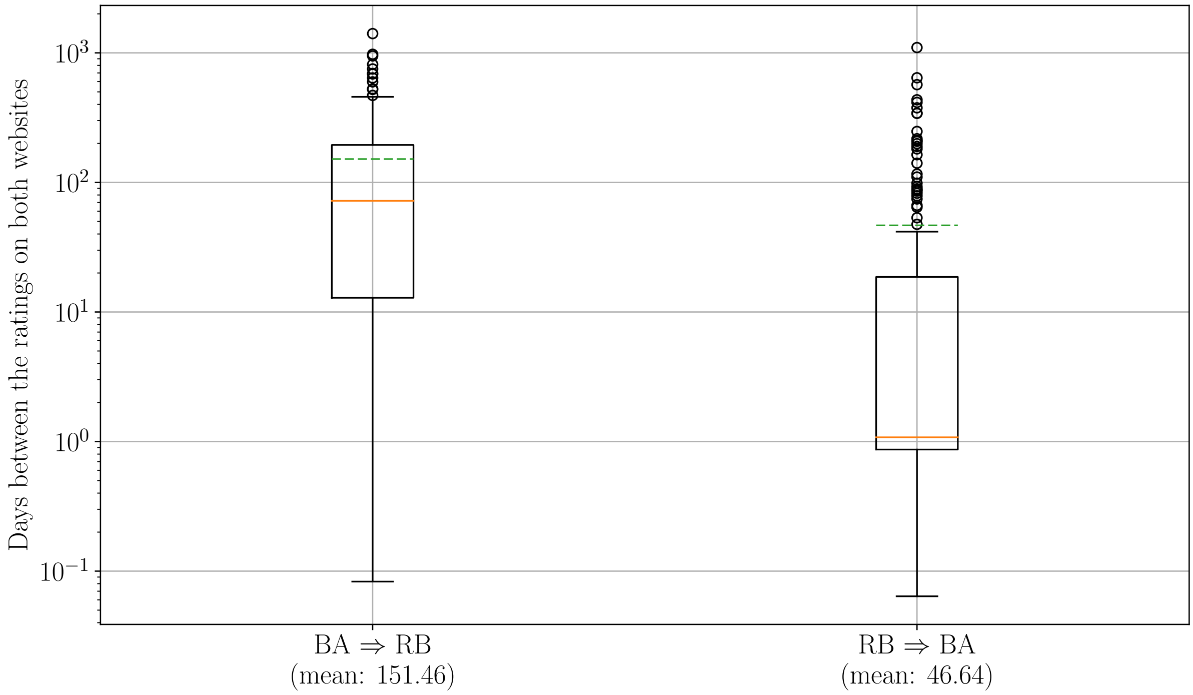


Figure 38: Average number of days between the ratings on both websites for each direction. Each point correspond to the average number of days for a user.

There is a big difference here. Indeed, it would have been more logical to have the same number of days between the ratings in each direction. However, this is not the case. Finding an explanation for this can be quite tricky. One of the most plausible reason is that the users in the direction "BeerAdvocate ⇒ RateBeer" are joining much later and therefore they are rating many beers the same day. Figure 39 gives the number of days between joining the two websites for both directions.

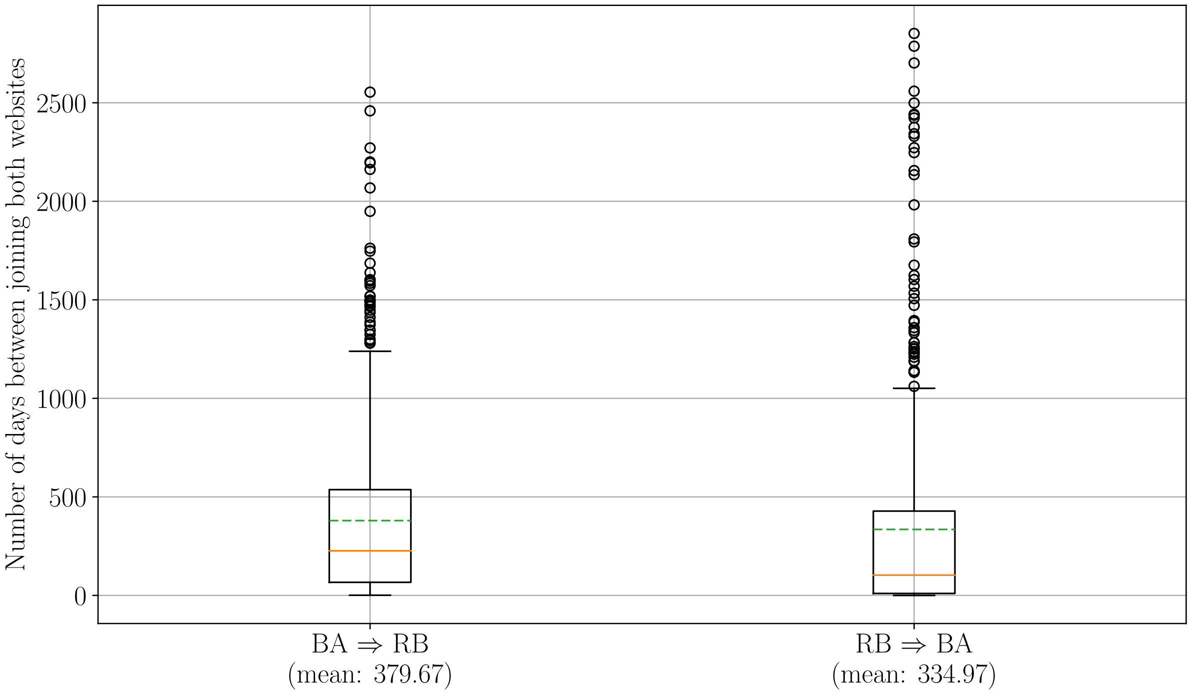


Figure 39: Average number of days between joining both websites for the two directions.

This time, there is a small difference in the number of days between joining the websites. The users in the direction "RateBeer ⇒ BeerAdvocate" tend to take a little bit less time before joining BeerAdvocate in average. However, the average for both directions is around a year which is quite long. Therefore, it is not possible to use the big time difference to explain the numbers in Table 22. It is then useful to have a look at the number of beers rated per day, see Figure 40. On average, the users rate one beer per day. The days when a user did not rate any beer were not taken into account. It is interesting to see that the users in the direction "BeerAdvocate ⇒ RateBeer" have a small tendency to rate a bit more beers per days than the users joining in the other direction. It could give the beginning of the explanation for the low number of reviews given on both websites in less than a day for this direction, see Table 22.

This analysis can continue for a long time, but this is not the most interesting aspect of these data. Therefore, it is better to make sure that the whole span of time can be used instead of restricting on less than a day. Figure 41 shows the boxplots of the differences of the ratings between BeerAdvocate and RateBeer for each aspect. The direction was not taken into account. For each aspect, three box plots are given. The first one is done with reviews posted on both websites in less than a day (5'251 ratings), the second one is done with reviews posted after more than a day and less than a week (4'215 ratings), and the third one is done with reviews posted with a time difference bigger than a week (6'502 ratings). The boxplots are quite similar between the time periods. Therefore, the predictions should be similar between all the time frames. The results are given in the next subsection.

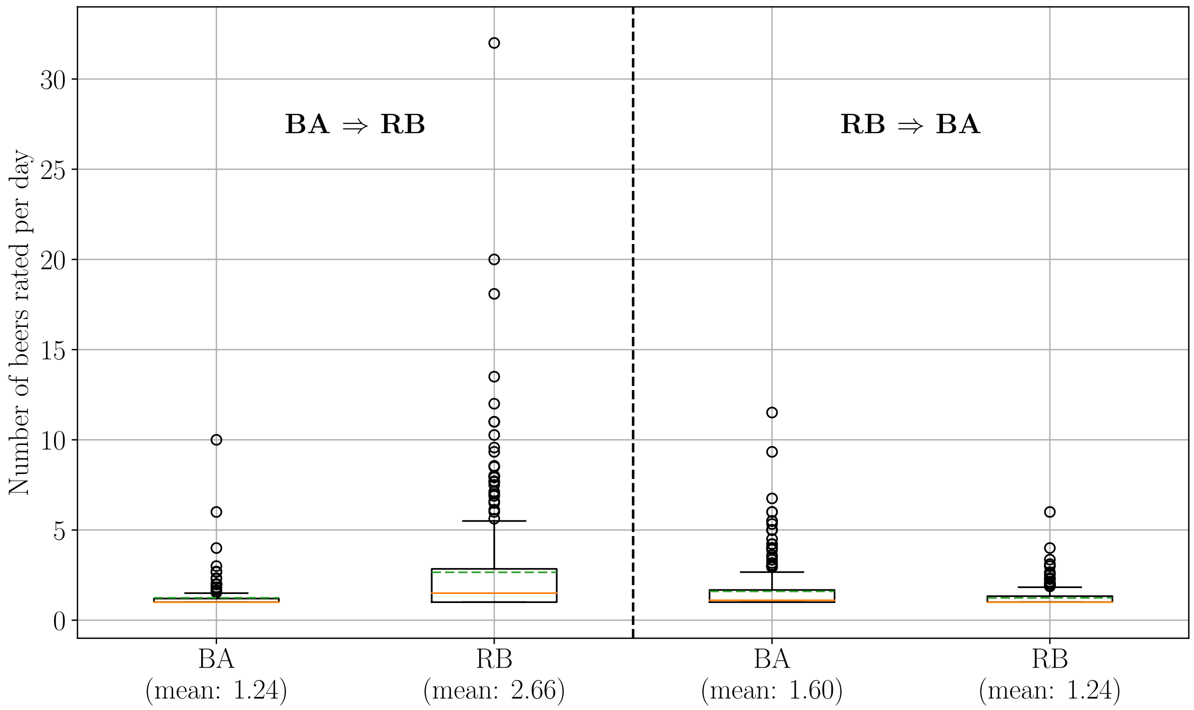


Figure 40: Average number of beers per user rated on both websites for the two directions.

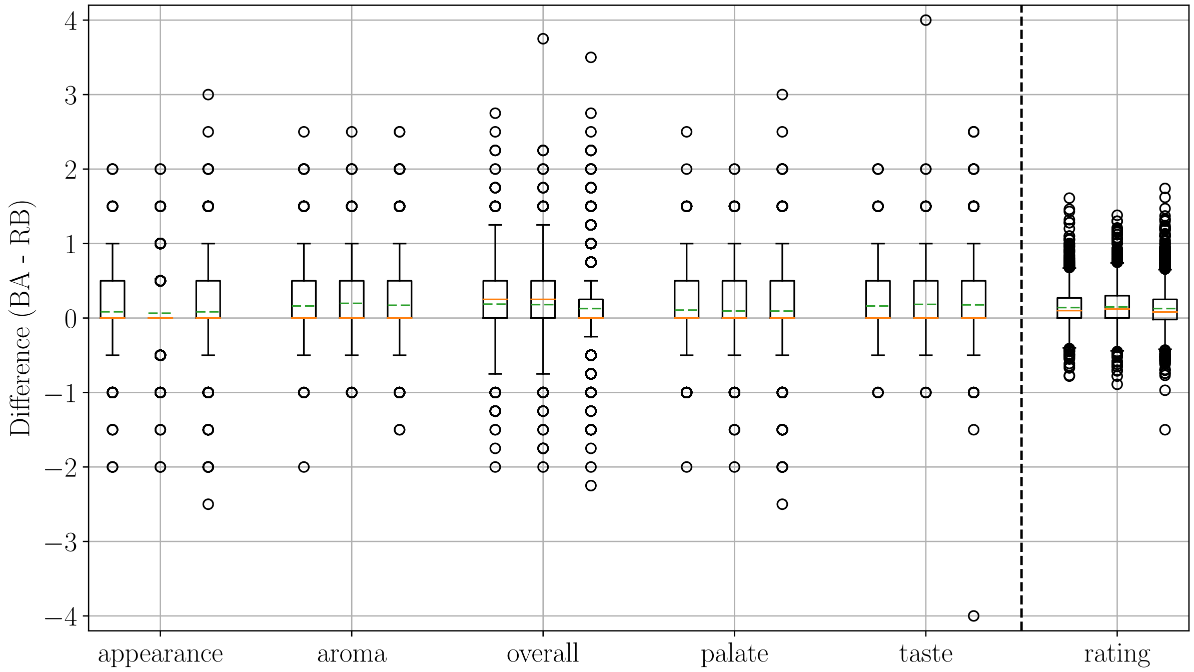


Figure 41: Differences of ratings for all the aspects. Each aspect has three boxplots linked to the time between the ratings on both websites, in order: less than a day, more than a day and less than a week, and more than a week.

Predicting the Ratings on the Second Website

For the final part of the results, the long-awaited Machine Learning model, one a bit more complex than the standard linear regression, is used. The aim of this subsection is to compare the predictions in the direction “RateBeer ⇒ BeerAdvocate” to the direction “BeerAdvocate ⇒ RateBeer”. But, first, it is important to remove the influence of some specific users. Therefore, all the users who have not rated the beers in both directions are eliminated. This leaves a total of 14’035 reviews, 10’758 for the direction “RateBeer ⇒ BeerAdvocate” and 3’277 for the direction “BeerAdvocate ⇒ RateBeer”.

Before starting with the Machine Learning model, a hypothesis has to be stated. Figure 42 shows the differences for each of the aspects for both directions.

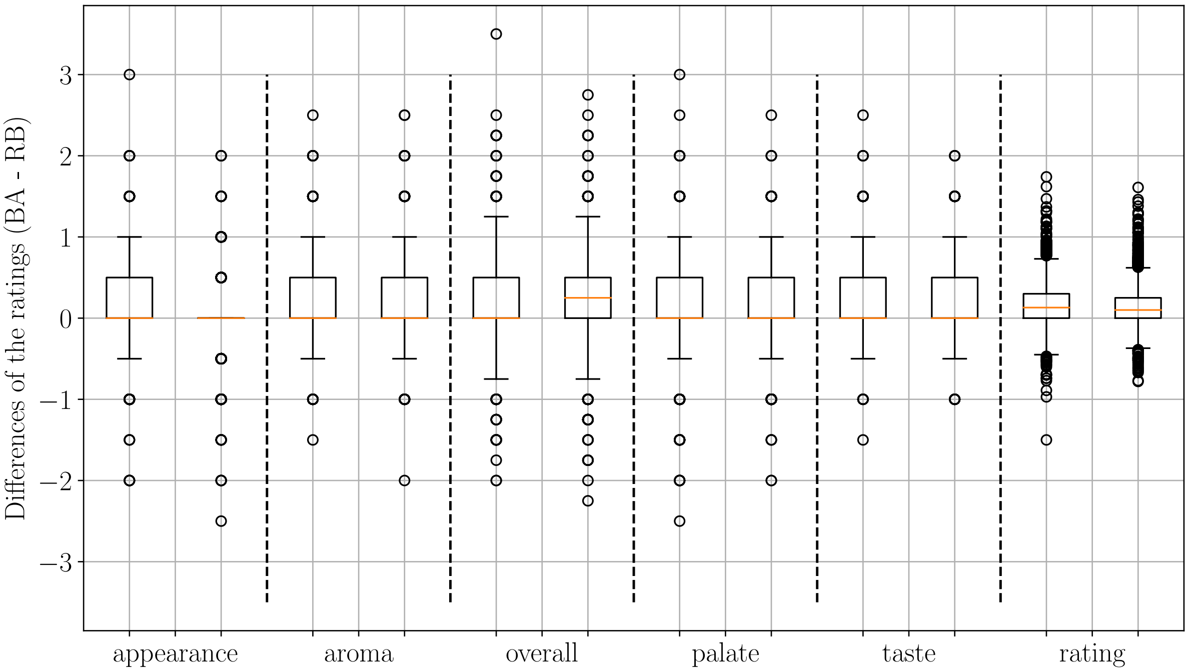


Figure 42: Differences of ratings for all the aspects. Each aspect has two boxplots. The first one is for the direction “BeerAdvocate ⇒ RateBeer”, the second is for the direction “RateBeer ⇒ BeerAdvocate”.

The boxplots for the differences are quite similar between the two directions. Therefore, the hypothesis is that there will be a similar prediction error for both directions.

First, a good Machine Learning model is nothing without useful and clever features. These features have been chosen according to what the users could see and according to his experience on both websites. The explanation, as well as the presentation of these features, are given in Table 23.
For the Machine Learning model, a Random Forest Regressor was chosen. It has been implemented in the function `RandomForestRegressor` from the package `sklearn.ensemble` [20]. The only parameter that has been changed is the number of estimators, `n_estimators`, which corresponds to the number of Decision Tree in the forest. The data have been bootstrapped 200 times by randomly selecting 3’000 ratings every time.

Type/Level	Description	Number
Time	• The time difference between the rating on both website in days.	**1**
Beer	• The ratings on all the aspects on the first website for the given beer.	**5**
	• The final rating on the first website.	**1**
	• The average rating of the beer on the first website.	**1**
	• The average rating of the beer on the second website.	**1**
User	• The average ratings on all the aspect given on the first website by this user.	**5**
	• The average final rating given on the first website by this user.	**1**
	• The average ratings on all the aspect given on the second website by this user.	**5**
	• The average final rating given on the second website by this user.	**1**
History Website	• The last 5 ratings for the five aspects on the second website.	**25**
	• The last 5 final ratings on the second website.	**5**

Table 23: Features used in the Machine Learning model to predict the ratings of the second website given the ratings on the first website. There is a total of 51 features.

The first model is done with the direction "RateBeer ⇒ BeerAdvocate" and only the reviews posted in less than a day on both websites are used. It gives a total of 4'653 reviews. The results are used as the reference to compare them to other models trained on other sets of data. The results are given in Figure 43. The first five boxplots are the aspects. Then, the prediction of the final ratings is shown in the sixth boxplot. The seventh boxplot represents the RMSE of the final rating computed using the predictions of the five aspects.
For the aspects, the predictions follows the correlations given in Table 19. Indeed, overall is the most difficult aspect to predict while aroma and taste are the easiest. It is also interesting to note that the computed final rating is a bit better than the predicted final rating. Finding a specific explanation for this is difficult.

Now, the randomness of the users has to be tested. Indeed, it is possible that the users have the tendency to give random differences between the ratings on the first and the second websites. Therefore, the differences are computed for a given aspect on all the reviews. Then this array of differences is shuffled. Finally, the BeerAdvocate ratings are computed using the RateBeer ratings plus one of the random differences for each aspect. The correct distributions of each of the aspects are kept with this method. Finally, these new ratings are used by the Machine Learning model with the same features as before. The results are given in Figure 44.

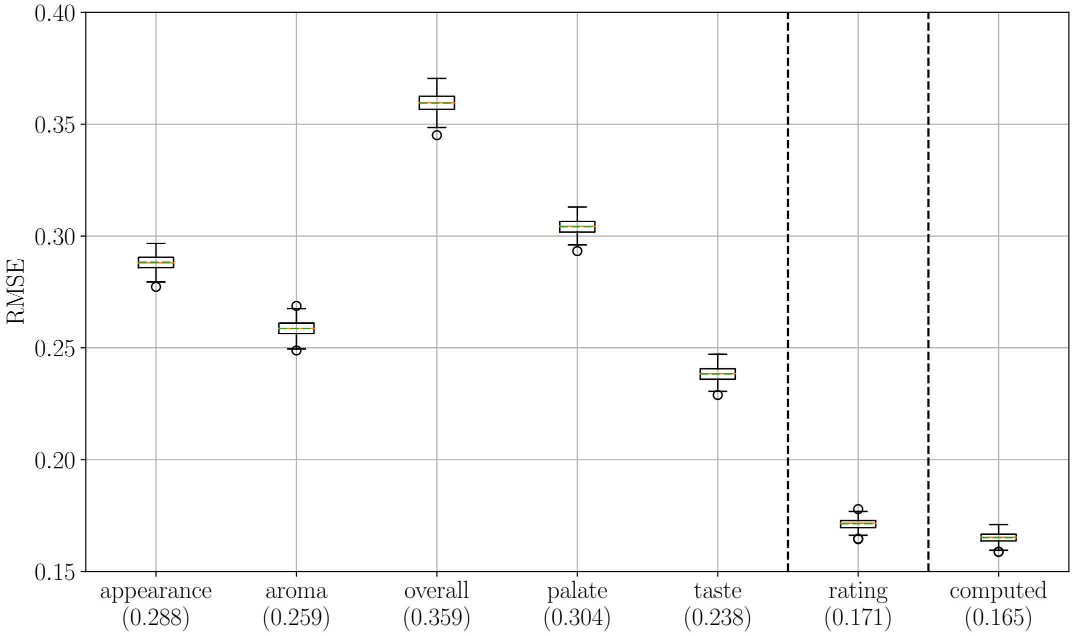


Figure 43: RMSE for the prediction of the ratings in the direction “RateBeer ⇒ BeerAdvocate”. The reviews were posted in less than a day on both websites. The values in parentheses are the average value. Last boxplot is the computation of the final rating using the prediction of all the aspects.

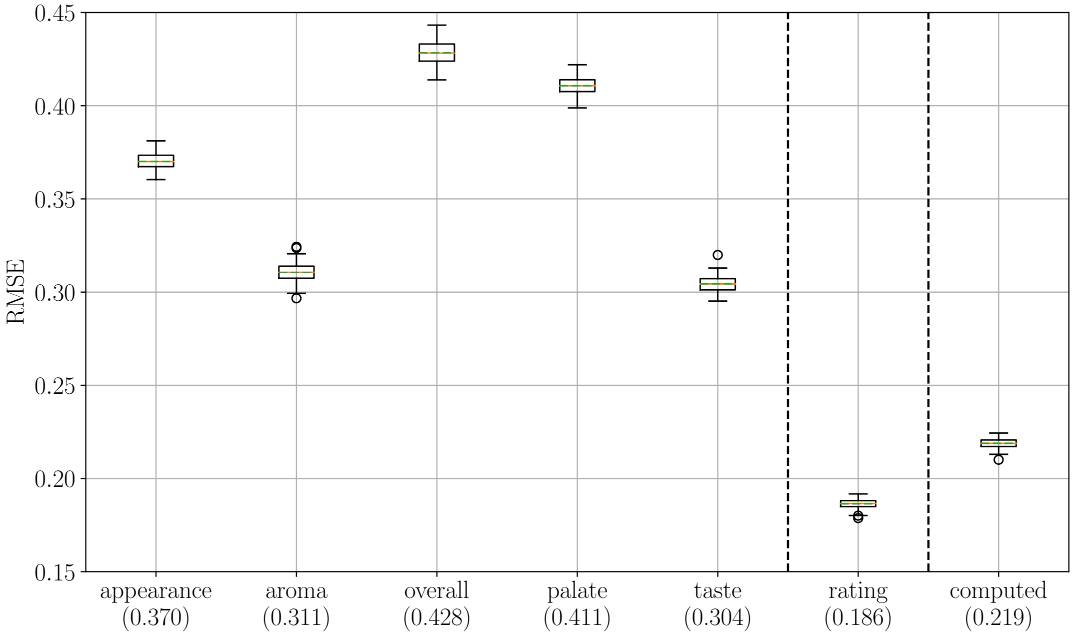


Figure 44: RMSE for the prediction of the ratings in the direction “RateBeer ⇒ BeerAdvocate”. The reviews were posted in less than a day on both websites. The ratings on BeerAdvocate were computed by randomizing the differences with the RateBeer ratings. The values in parentheses are the average value. Last boxplot is the computation of the final rating using the prediction of all the aspects.

The predictions in Figure 44 are worse than the predictions in Figure 43. Overall is the aspect that worsens the less. That means that this is the most random aspects regarding the differences between both websites. On the other hand, for the other aspects, the randomness increases quite a bit the RMSE, meaning that the users are using some logic while giving different ratings. It is also interesting to note the prediction of the final rating did not suffer that much from the random differences. It has already been shown in Subsection Randomness between the two Websites. The computed rating, logically, has a higher RMSE due to the higher RMSE values for each of the aspects.

To complete the randomness study, the Machine Learning model can be used again but this time with random values as the ratings on BeerAdvocate. They are simply drawn randomly between all the possible values on BeerAdvocate, see Table 3. This time, the Machine Learning model should not correctly work since there is no logic in the way the BeerAdvocate ratings are done. The results are given in Figure 45.

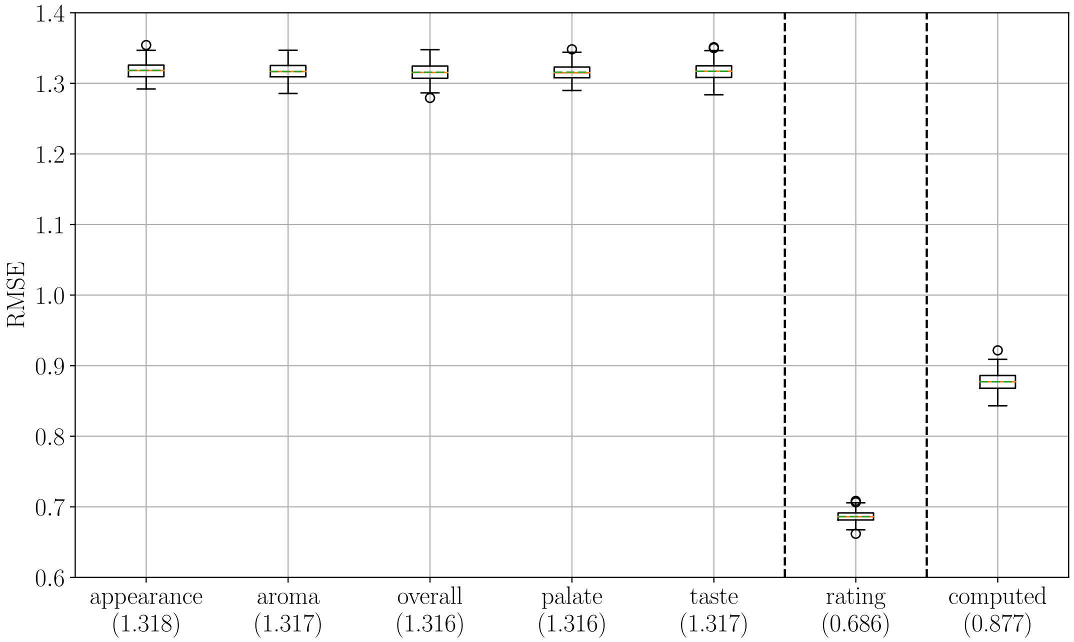


Figure 45: RMSE for the prediction of the ratings in the direction "RateBeer ⇒ BeerAdvocate". The reviews were posted in less than a day on both websites. The ratings on BeerAdvocate were computed by randomly drawing them between all the possible values on BeerAdvocate. The values in parentheses are the average value. Last boxplot is the computation of the final rating using the prediction of all the aspects.

The RMSE values are a lot higher than the values on the two previous graphs. All of the aspects have the same RMSE which is logical. It is interesting to note that the prediction of the final rating[16] is much better than the prediction of the aspects. It means that the aggregation of the aspects leads to a drop in the randomization. It is simply because different combinations of ratings on the aspect will lead to the same final rating.

The results of the first three tests with the Machine Learning model are summarized in Table 24. The increase for the average RMSE is also given. Once again, the random differences

[16] The final rating was computed using the random ratings of the aspects.

lead to results that are worse than the unchanged ratings but not as bad as the random ratings. The prediction on the final rating is the closest between the unchanged ratings and the ratings with random differences. The reason has already been discussed earlier in this subsection.

Types	Reference	Random differences		Random ratings	
appearance	0.288	0.370	+28.4%	1.318	+357.3%
aroma	0.259	0.311	+20.1%	1.317	+409.0%
overall	0.359	0.428	+19.2%	1.316	+266.1%
palate	0.304	0.411	+35.9%	1.316	+332.6%
taste	0.238	0.304	+27.7%	1.317	+452.6%
rating	0.171	0.186	+8.8%	0.686	+300.5%
computed	0.165	0.219	+32.5%	0.877	+431.1%

Table 24: Summary of the results obtained with the Machine Learning model on the test with randomness. The direction for the prediction is "RateBeer ⇒ BeerAdvocate". All reviews have been posted in less than a day.

For all the previous tests, a separated model was trained for each of the aspects. It is interesting to know if the prediction on all the aspects and the final rating at the same time give similar results. Therefore, the same unchanged data, as in Figure 43, are used, and only one model is trained. The results are shown in Figure 46.

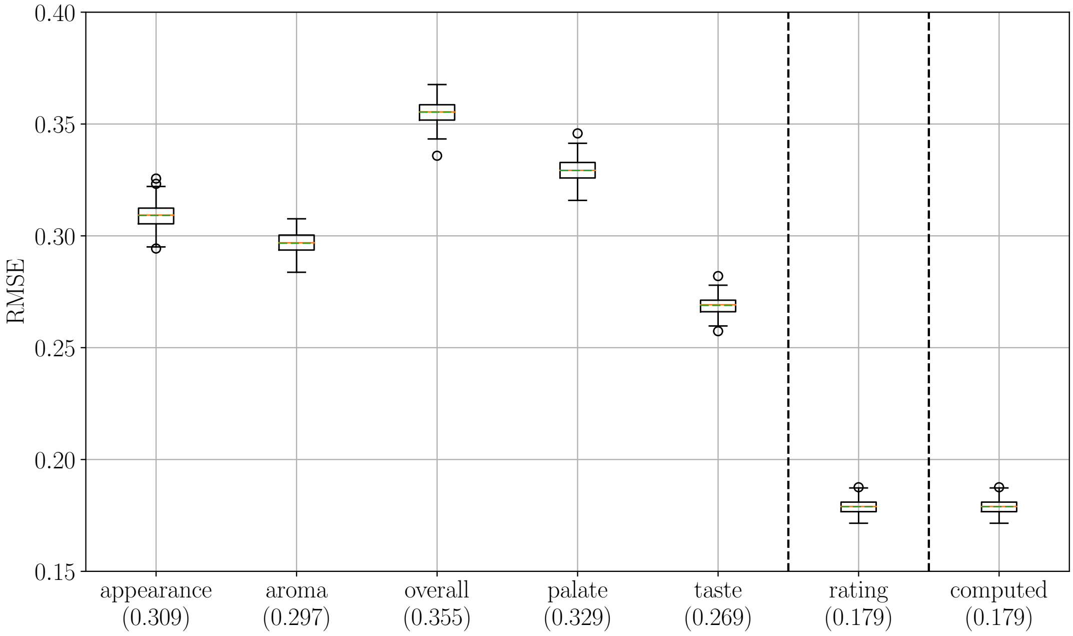


Figure 46: RMSE for the prediction of the ratings in the direction "RateBeer ⇒ BeerAdvocate". The reviews were posted in less than a day on both websites. All the aspects as well as the final rating are predicted using only one model. The values in parentheses are the average value. Last boxplot is the computation of the final rating using the prediction of all the aspects.

These results are a bit less good than for the separated predictions except for the aspect overall. It is possible that overall is used to change the final rating of a beer to make it closer to the rating on the first website. Therefore, it would benefit to get the information about the other aspects. Table 25 gives the comparison with the separated predictions.

Types	Reference	One model for all	
appearance	0.288	0.309	+7.2%
aroma	0.259	0.297	+14.7%
overall	0.359	0.355	-1.1%
palate	0.304	0.329	+8.2%
taste	0.238	0.269	+12.8%
rating	0.171	0.179	+4.4%
computed	0.165	0.179	+8.3%

Table 25: Summary of the results obtained with the Machine Learning model on the test with the prediction of all the aspects at the same time. The direction for the prediction is "RateBeer ⇒ BeerAdvocate". All reviews have been posted in less than a day.

Since now, only the data in the direction "RateBeer ⇒ BeerAdvocate" could be used. However, we want to compare the results in both directions. Therefore, the next step is to check if it is possible to use the reviews that were posted on both websites with a time difference bigger than a day. Therefore, the remaining 6'105 reviews that have been posted on BeerAdvocate more than a day after being posted on RateBeer are used with the Random Forest. The results are given in Figure 47.

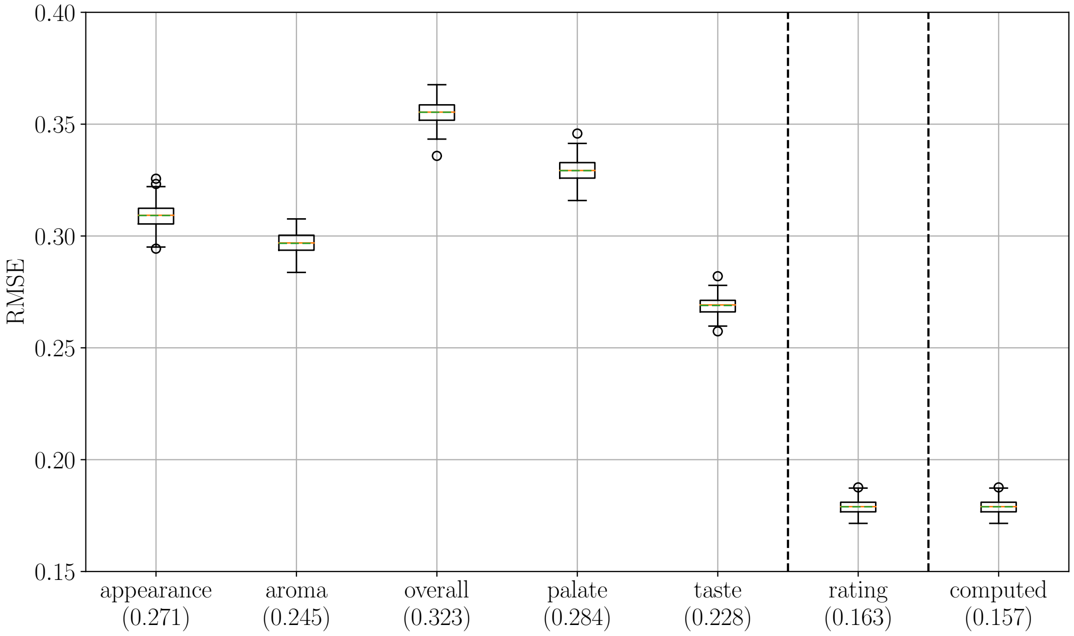


Figure 47: RMSE for the prediction of the ratings in the direction "RateBeer ⇒ BeerAdvocate". The reviews were posted in more than a day on both websites. The values in parentheses are the average value. Last boxplot is the computation of the final rating using the prediction of all the aspects.

The results are quite similar to the results in Figure 43 except that they are all better. It is astonishing since one can think that after a long time the users could forget about the ratings they put on the first website. However, it is also possible that they went back to the first website to check their ratings before posting them on the second website. Table 26 gives the comparison between the results with a time difference smaller than a day and bigger than a day.

Types	Reference	More than a day	
appearance	0.288	0.271	-6.1%
aroma	0.259	0.245	-5.2%
overall	0.359	0.323	-10.3%
palate	0.304	0.284	-6.8%
taste	0.238	0.228	-4.5%
rating	0.171	0.163	-4.9%
computed	0.165	0.157	-5.1%

Table 26: Summary of the results obtained with the Machine Learning model of the test with a different time difference. The direction for the prediction is "RateBeer ⇒ BeerAdvocate".

The results in the previous table show that the predictions are a bit better when the user tends to spend more time before rating the beer on the second website. It is quite surprising as said previously. The good news is that there is no reason to restrict the data on a time difference to make the predictions. Thus, it is possible to compare the two directions. Firstly, the predictions using all the data for the direction "RateBeer ⇒ BeerAdvocate" are computed. The results are given in Figure 48.

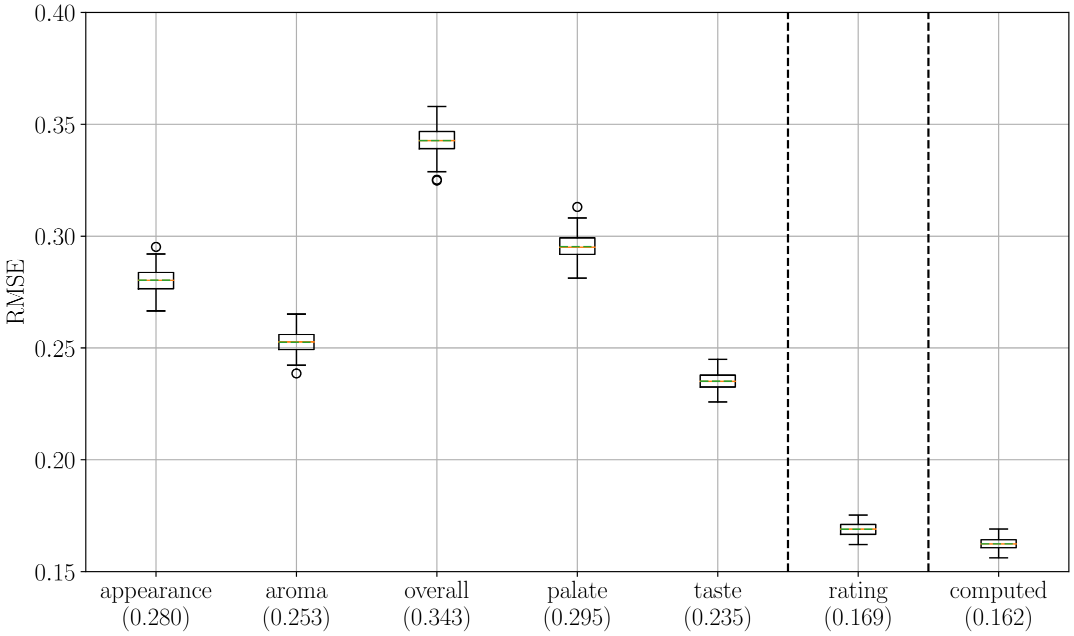


Figure 48: RMSE for the prediction of the ratings in the direction "RateBeer ⇒ BeerAdvocate". The values in parentheses are the average value. Last boxplot is the computation of the final rating using the prediction of all the aspects.

The results are close to the reference which was expected. They are a bit better due to the newly added data that were better than the reference. The comparison of the average values is given in Table 27.

Types	Reference	All data	
appearance	0.288	0.280	-2.8%
aroma	0.259	0.253	-2.4%
overall	0.359	0.343	-4.6%
palate	0.304	0.295	-2.9%
taste	0.238	0.235	-1.4%
rating	0.171	0.169	-1.4%
computed	0.165	0.162	-1.7%

Table 27: Summary of the results obtained with the Machine Learning model of the test with a subset and all the data. The direction for the prediction is “RateBeer ⇒ BeerAdvocate”.

It is finally time for the comparison between the two directions. The results for the direction “BeerAdvocate ⇒ RateBeer” are given in Figure 49.

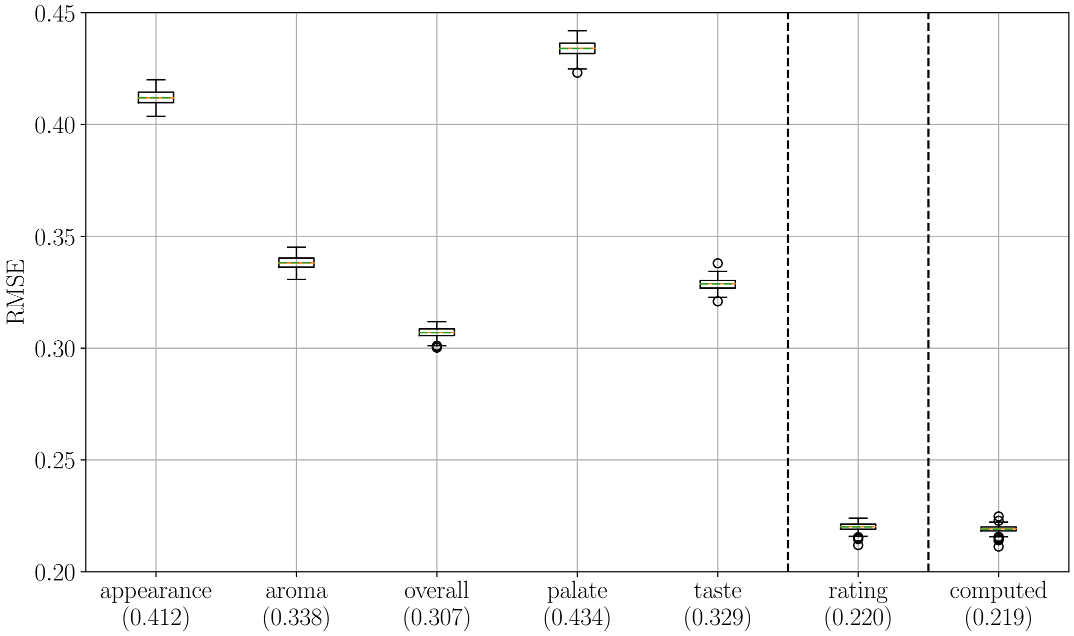


Figure 49: RMSE for the prediction of the ratings in the direction “BeerAdvocate ⇒ RateBeer”. The values in parentheses are the average value. Last boxplot is the computation of the final rating using the prediction of all the aspects.

The predictions in this direction are not good compared to the other direction. Overall is the only aspect for which the predictions do not become worse. It even becomes better. For the other aspects, the results are similar to the test with the random differences. Table 28 shows the comparison between the two directions with all the data.

Types	RB ⇒ BA	BA ⇒ RB	
appearance	0.280	0.412	+47.0%
aroma	0.253	0.338	+33.9%
overall	0.343	0.307	-10.4%
palate	0.295	0.434	+47.0%
taste	0.235	0.329	+39.8%
rating	0.169	0.220	+30.2%
computed	0.162	0.219	+34.9%

Table 28: Summary of the results obtained with the Machine Learning model with the two directions for the predictions.

Finding an explanation at this point is difficult. Most of the time, Machine Learning models have the tendency to work better when there is a small standard deviation in the data. Therefore, it is good to check if the ratings of the different aspects are more consistent in a direction than in the other. Figure 50 shows the boxplots of the Euclidian distance in the space of dimension 5 with the five aspects as coordinates. A smaller distance implies that the ratings are closer to each other within a review.

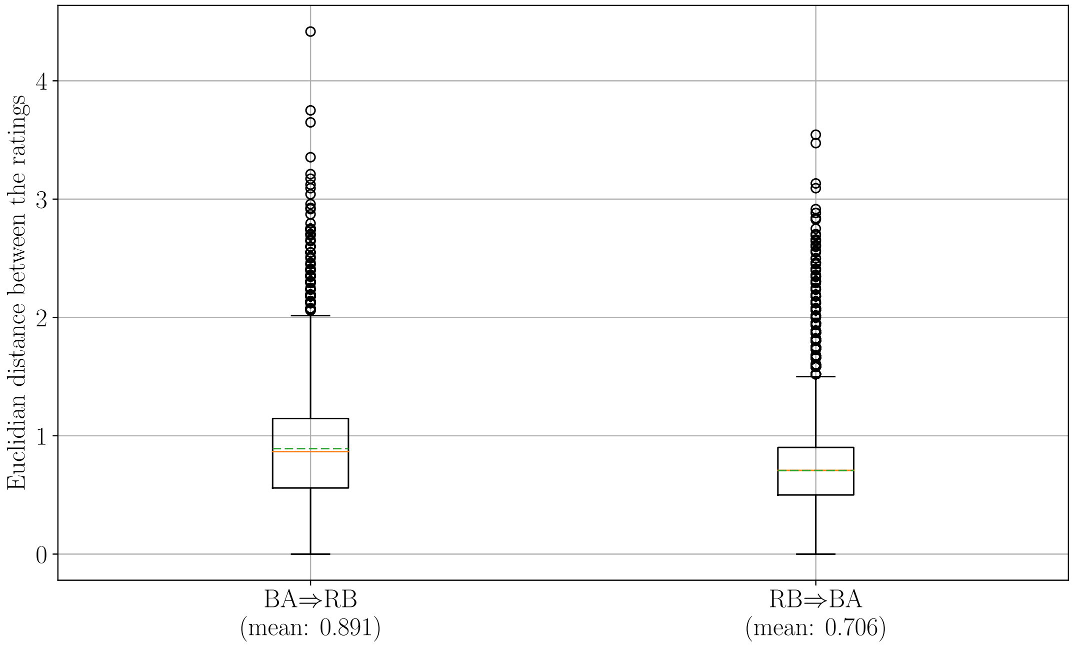


Figure 50: Euclidian distance in the space of dimension 5 with the five aspects as coordinates for both directions.

Tthe average distance is a little bit higher (around 20%) in the direction “BeerAdvocate ⇒ RateBeer” than in the direction “RateBeer ⇒ BeerAdvocate”. This fact can only explain a part of this difference. It surely comes from many different factors. The first figure in this subsection, see Figure 42, could never give the idea of such disparate results between the two directions. For example, the appearance has smaller whiskers, but the difference is almost the same everywhere. Therefore, it is important to finish with the comparison of these results with a dummy model: the predicted ratings are the ratings given on the first website. The results for the direction “RateBeer ⇒ BeerAdvocate” are shown in Figure 51 and the results for the direction “BeerAdvocate ⇒ RateBeer” are given in Figure 52.

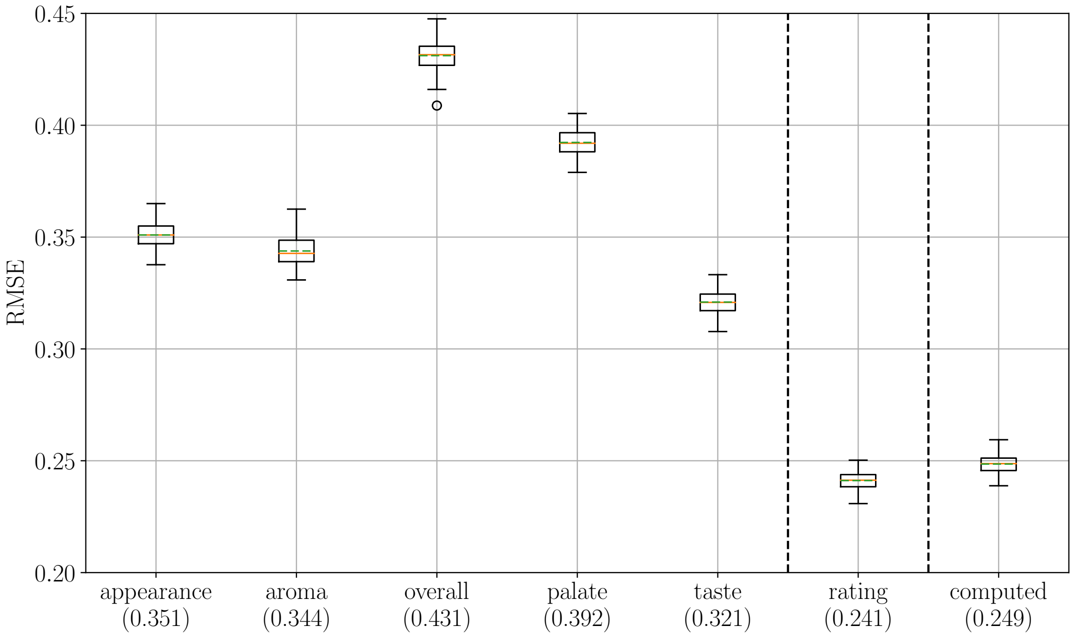


Figure 51: RMSE for the prediction with a dummy model of the ratings in the direction “RateBeer ⇒ BeerAdvocate”. The values in parentheses are the average value. Last boxplot is the computation of the final rating using the prediction of all the aspects.

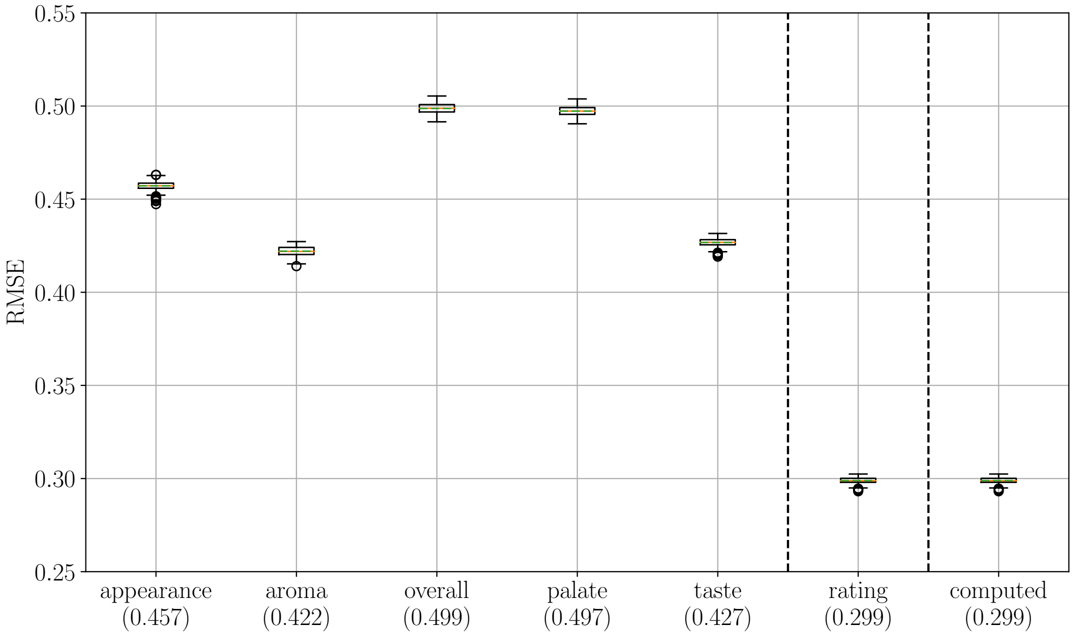


Figure 52: RMSE for the prediction with a dummy model of the ratings in the direction “BeerAdvocate ⇒ RateBeer”. The values in parentheses are the average value. Last boxplot is the computation of the final rating using the prediction of all the aspects.

These two dummy models perform worse than the RandomForest which is quite reassuring. Indeed, it would be strange to see a dummy model having better performance than an advanced Machine Learning model. Table 29 shows the increase between the Machine Learning model and the dummy model for both directions. The model for the direction “RateBeer ⇒ BeerAdvocate” improves a little bit more the results compared to the dummy model than the other direction. The dummy model for the direction “BeerAdvocate ⇒ RateBeer” is worse than the dummy model for the other direction. It simply means that the users have the tendency to give ratings that are further away from the first ratings for this direction.

Types	RB ⇒ BA	dummy		BA ⇒ RB	dummy		
appearance	0.280	0.351	+25.2%	0.412	0.457	+11.0%	+30.3%
aroma	0.253	0.344	+36.1%	0.338	0.422	+24.8%	+22.8%
overall	0.343	0.431	+25.8%	0.307	0.499	+62.5%	+15.7%
palate	0.295	0.392	+32.8%	0.434	0.497	+14.6%	+26.8%
taste	0.235	0.321	+36.5%	0.329	0.427	+29.8%	+33.0%
rating	0.169	0.241	+42.7%	0.220	0.299	+35.9%	+24.0%
computed	0.162	0.249	+53.0%	0.219	0.299	+36.5%	+20.3%

Table 29: Summary of the results obtained with the Machine Learning model with the two directions for the predictions compared to the dummy models. Last column corresponds to the increase of the dummy model for the direction “BA ⇒ RB” compared to the dummy model for the direction “RB ⇒ BA”.

There is no simple nor clear explanation why the hypothesis given at the beginning of this subsection was not fulfilled. Indeed, even if the scales and weightings are different between the two websites, this relation is bijective. Indeed, only the users who rated the beers in both directions have been kept. More analysis is required to answer this question that still stays open fully. However, no one should forget that humans are not machines and sometimes it is not possible to find any rational explanation.

Conclusion

In this thesis, the data of BeerAdvocate and RateBeer websites have been studied deeply. Even if the starting idea of this project, *i.e.* creating a Machine Learning model based on the model shown in the paper of Wang & Wang [1], had to be aborted, some very interesting discoveries have been made. Working with human-made data is very interesting and quite difficult. Everybody can learn a lot because there is no simple equations nor theory to describe the behavior of humans. With all of these analyses, the power of using and comparing two datasets has been demonstrated. Indeed, many of those analyses could have been done with only one of the two datasets. But the comparison between them lead to some very interesting results. The good news is that the analyses of these two datasets are far from being finished.

In this paragraph, the different results obtained in this thesis are recalled. For the first part of the results, it has been shown that the two websites are quite different. Indeed, the scales and weightings of the rating systems are different. They also display a score that is very different. It has been shown that the ratings themselves have a different average, even on the beers that are reviewed on both websites. At the end of this first part, the matching algorithms have been presented. The matched data have been used for the next two parts of this thesis.
The second part of the results, using the matching beers, started with a small analysis on the influence of the scores on the ratings. This analysis concluded that the users are not that influenced by the scores, perhaps because of the different scaling of these scores. We state here that this score is mostly here for the people who are looking for good beers instead of reviewing them. Then, the next section was about the prediction of the tenth rating using the previous ones. Due to a decrease of the standard deviation on the ratings every year, the predictions were not conclusive. Finally, a specific environment comparing the beers on both websites was created for studying the herding effects. With this analysis, two fascinating results were found: the average rating of a beer is influenced by the early ratings, and the users have the tendency to follow this crowd judgment. It is one of the strongest results concerning the fact that the users have the tendency to follow the average opinion.
Finally, the third and last part of the results, using the matched users and matched beers, started with a little bit of Natural Language Processing. The model developed in this section successfully classified the reviews between the two websites, from a subset of the reviews written by the matching users and from a subset of all the reviews. This result shows that both communities have the tendency to build their own vocabulary. And this is not only due to the fact that the websites are displaying different words with the same meaning as demonstrated in the last part of this section. Then, using the model developed in the previous section, the users who are in both communities were classified based on their writing and their joining date. The results gave the information that these users have the tendency to take inspiration from both communities. Then, the investigation continued on the users who are in both communities with different statistics. For example, it has been shown that these users have the tendency to give closer ratings between both websites than the average of these beers. But they do not give the exact same ratings. They also have the tendency to copy-paste their reviews between both websites. The final part of this thesis is about understanding why these users have the tendency to give different ratings between both websites even if their opinion is the same. Firstly, the correlation between the ratings has been investigated. It was interesting to see that the correlation is linked to the scaling of the ratings. Then, it has been shown that some users gives really strange ratings between both websites. Thus, the analysis continued on the randomness of the differences between the ratings on both websites. Using a multinomial distributions for each of the aspects, new ratings were created. The final rating was not affected by these random

differences. The final part of the results is about using a Random Forest to predict the ratings in both directions, *i.e.* "RateBeer ⇒ BeerAdvocate" and "BeerAdvocate ⇒ RateBeer". The same users have the tendency to give more similar ratings in the direction "RateBeer ⇒ BeerAdvocate" than in the other direction. The Euclidian distance in the space of the five aspects also showed the same result. However, the features as well as the Machine Learning model helped to improve the predictions against a dummy model, giving a results around 30% better. The complete logic of the users could not be found with this analysis. More analyses are required in order to fully grasp the users behavior.

We do not want to conclude this thesis with just a collection of results. They can be used in a clever way to create better online rating systems. Throughout the creation of such a website, the developer needs to think if he wants to create a strong community or if he wants to create a neutral website. The results collected in this thesis can help with that. First, we state that rating different aspects of a product is a very clever way of proceeding. Indeed, in the paper of Hu *and al* [4] it is shown that the five-stars rating system of Amazon has the tendency to create this "*Brag-and-Moan*" model. The use of a more complex rating system will discourage some of the users to do this. If the developer wants to build a strong community, he has to think about the vocabulary he will use on his website. Indeed, the text features could be use to separate accurately BeerAdvocate and RateBeer communities. This feeling of belonging is very important for a user to stay longer in the community. On the other hand, using a neutral vocabulary could help the community to stay neutral and avoid developing their own norms. This is a trade-off that the creator needs to think about. We also think that it is very important for the developer to check which websites exist on the same products. Indeed, the community will have members coming from other existing communities. Some users like to be part of many different communities. Therefore, it is important for the developer to give the possibility for these users to change between them easily. For example, it could be interesting to have different scales on the user's choice. He could use a rating system with the given weights as in RateBeer or use one having the same scales for each aspect and hidden weights as in BeerAdvocate. We think that being able to see the final rating in real time, as in RateBeer, is very useful. Especially if some users have the tendency to visit different communities. Even if the results with the Machine Learning model were better in the direction "RateBeer ⇒ BeerAdvocate" than in the direction "BeerAdvocate ⇒ RateBeer"[17] We still think that this feature can really help some users. If the developer wants to reduce the herding effects and the influence of the early ratings on the community norms, we think that the early ratings should not be displayed to the users, at least not before the user rated the product. If the website is created this way, then the early influences will totally disappear. It is also possible that the herding effects will be lessen since the early ratings can be disparate. Finally, showing a big score on the web page can be interesting. We also think that having a different scale than the ratings for this score is smart. The idea of using the percentiles as on RateBeer is handy for the casual users who just want to test a great product. On the other hand, the scaling of the score on BeerAdvocate is quite strange. We think that a little bit more transparency for the users would be better.
All of these advice are subjective and come from our experience while working with these data during the whole project. They do not represent the perfect way to create a website based on an online rating system. We just wanted to give a human and useful meaning to all of these results obtained in thesis. These kind of analyses take a lot of time, but everybody can learn a lot from them. That is the reason we wanted to give our opinion on how it is possible to make a better online rating system.

[17] Only RateBeer displays such a feature.

As stated earlier, the analyses on these two datasets are far from being finished. For example, only a small analysis of the texts was conducted. And there is still a lot to do on this. Some papers already worked with the texts. But they just checked how some specific words appeared or disappeared, see the paper of Danescu *et al.* [17]. It can be interesting, for example, to investigate how the communities build their vocabulary over the years. A network analysis of the users could be fascinating as well. Perhaps, some of the users tend to influence the other users. And it is possible that some parts of the graphs will have a high density of users who like specific types of beer. It is also possible to continue working with different Machine Learning models in order to predict many different things. As stated in the last part of the results, we could not grasp the whole logic of the users. Many analyses are still required in order to fully understand these users. Since a good data scientist always want more data, finding other websites on beer reviews in the U.S. would be interesting. It would be then possible to compare the results obtained in this thesis with the new datasets. But there is something that needs to be done first: the data have to be scraped again. Indeed, they have been scraped in early 2012. Therefore, five years and a half of data are missing. As explained in Subsection Datasets, more than five millions reviews can be added if the data are scrapped again in 2017.

References

[1] T. Wang and D. Wang, "Why amazon's ratings might mislead you: The story of herding effects," Big Data Journal, 2014.

[2] M. Glenski and T. Weninger, "Rating effects on social news posts and comments," CoRR, vol. abs/1606.06140, 2016.

[3] Y. Zhang, T. Lappas, M. Crovella, and E. D. Kolaczyk, "Online ratings: Convergence towards a positive perspective?," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4788–4792, May 2014.

[4] N. Hu, P. A. Pavlou, and J. Zhang, "Can online reviews reveal a product's true quality?: Empirical findings and analytical modeling of online word-of-mouth communication," in Proceedings of the 7th ACM Conference on Electronic Commerce, EC '06, (New York, NY, USA), pp. 324–330, ACM, 2006.

[5] S. M. Mudambi and D. Schuff, "What makes a helpful online review? a study of customer reviews on amazon.com," MIS Q., vol. 34, pp. 185–200, Mar. 2010.

[6] A. V. Banerjee, "A simple model of herd behavior*," The Quarterly Journal of Economics, vol. 107, no. 3, p. 797, 1992.

[7] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, Patterns of Cascading Behavior in Large Blog Graphs, pp. 551–556.

[8] L. R. Anderson and C. A. Holt, "Information cascades in the laboratory," The American economic review, pp. 847–862, 1997.

[9] S. A. Myers and J. Leskovec, "Clash of the contagions: Cooperation and competition in information diffusion," in Data Mining (ICDM), 2012 IEEE 12th International Conference on, pp. 539–548, IEEE, 2012.

[10] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," Journal of Marketing Research, vol. 43, no. 3, pp. 345–354, 2006.

[11] M. Luca, "Reviews, reputation, and revenue: The case of yelp. com," 2016.

[12] F. Zhu and X. Zhang, "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," Journal of marketing, vol. 74, no. 2, pp. 133–148, 2010.

[13] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection." http://snap.stanford.edu/data, June 2014.

[14] J. J. McAuley, J. Leskovec, and D. Jurafsky, "Learning attitudes and attributes from multi-aspect reviews," CoRR, vol. abs/1210.3926, 2012.

[15] J. J. McAuley and J. Leskovec, "From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews," in Proceedings of the 22Nd International Conference on World Wide Web, WWW '13, (New York, NY, USA), pp. 897–908, ACM, 2013.

[16] E. Guàrdia-Sebaoun, V. Guigue, and P. Gallinari, "Latent trajectory modeling: A light and efficient way to introduce time in recommender systems," in Proceedings of the 9th ACM Conference on Recommender Systems, RecSys '15, (New York, NY, USA), pp. 281–284, ACM, 2015.

[17] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts, "No country for old members: User lifecycle and linguistic change in online communities," in Proceedings of the 22nd international conference on World Wide Web, pp. 307–318, ACM, 2013.

[18] Internet Archive Wayback Machine. https://archive.org/web/, 1996.

[19] E. Jones, T. Oliphant, P. Peterson, et al., "SciPy: Open source scientific tools for Python." http://www.scipy.org/, 2001–.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[21] H. W. Kuhn, "The hungarian method for the assignment problem," Naval Research Logistics Quarterly, vol. 2, no. 1-2, pp. 83–97, 1955.

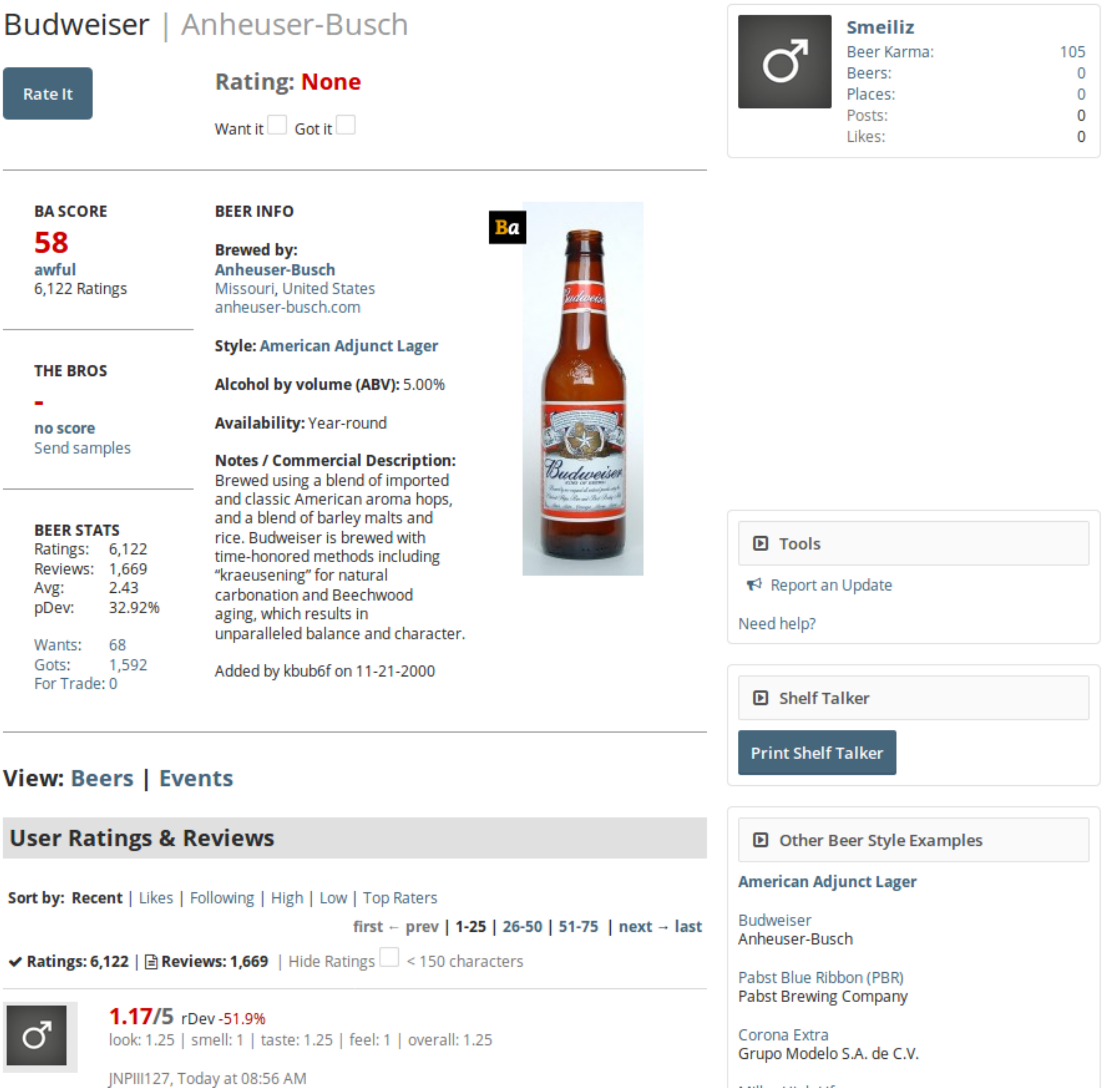
[22] L. Muchnik, S. Aral, and S. J. Taylor, "Social influence bias: A randomized experiment," Science, vol. 341, no. 6146, pp. 647–651, 2013.

[23] B. Bengfort, "Text Classification with NLTK and Scikit-Learn." http://bbengfort.github.io/tutorials/2016/05/19/text-classification-nltk-sckit-learn.html, 2016.

[24] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. O'Reilly Media, Inc., 1st ed., 2009.

Appendix

Websites



Budweiser | Anheuser-Busch

Rate It

Rating: None

Want it □ Got it □

Smeiliz

Beer Karma:	105
Beers:	0
Places:	0
Posts:	0
Likes:	0

BA SCORE
58
awful
6,122 Ratings

THE BROS
-
no score
Send samples

BEER STATS

Ratings:	6,122
Reviews:	1,669
Avg:	2.43
pDev:	32.92%
Wants:	68
Gots:	1,592
For Trade:	0

BEER INFO

Brewed by:
Anheuser-Busch
Missouri, United States
anheuser-busch.com

Style: American Adjunct Lager

Alcohol by volume (ABV): 5.00%

Availability: Year-round

Notes / Commercial Description:
Brewed using a blend of imported and classic American aroma hops, and a blend of barley malts and rice. Budweiser is brewed with time-honored methods including “kraeusening” for natural carbonation and Beechwood aging, which results in unparalleled balance and character.

Added by kbub6f on 11-21-2000

Tools

Report an Update

Need help?

Shelf Talker

Print Shelf Talker

View: Beers | Events

User Ratings & Reviews

Sort by: Recent | Likes | Following | High | Low | Top Raters

first ← prev | 1-25 | 26-50 | 51-75 | next → last

✓ Ratings: 6,122 | Reviews: 1,669 | Hide Ratings □ < 150 characters

1.17/5 rDev -51.9%
look: 1.25 | smell: 1 | taste: 1.25 | feel: 1 | overall: 1.25

JNPIII127, Today at 08:56 AM

Other Beer Style Examples

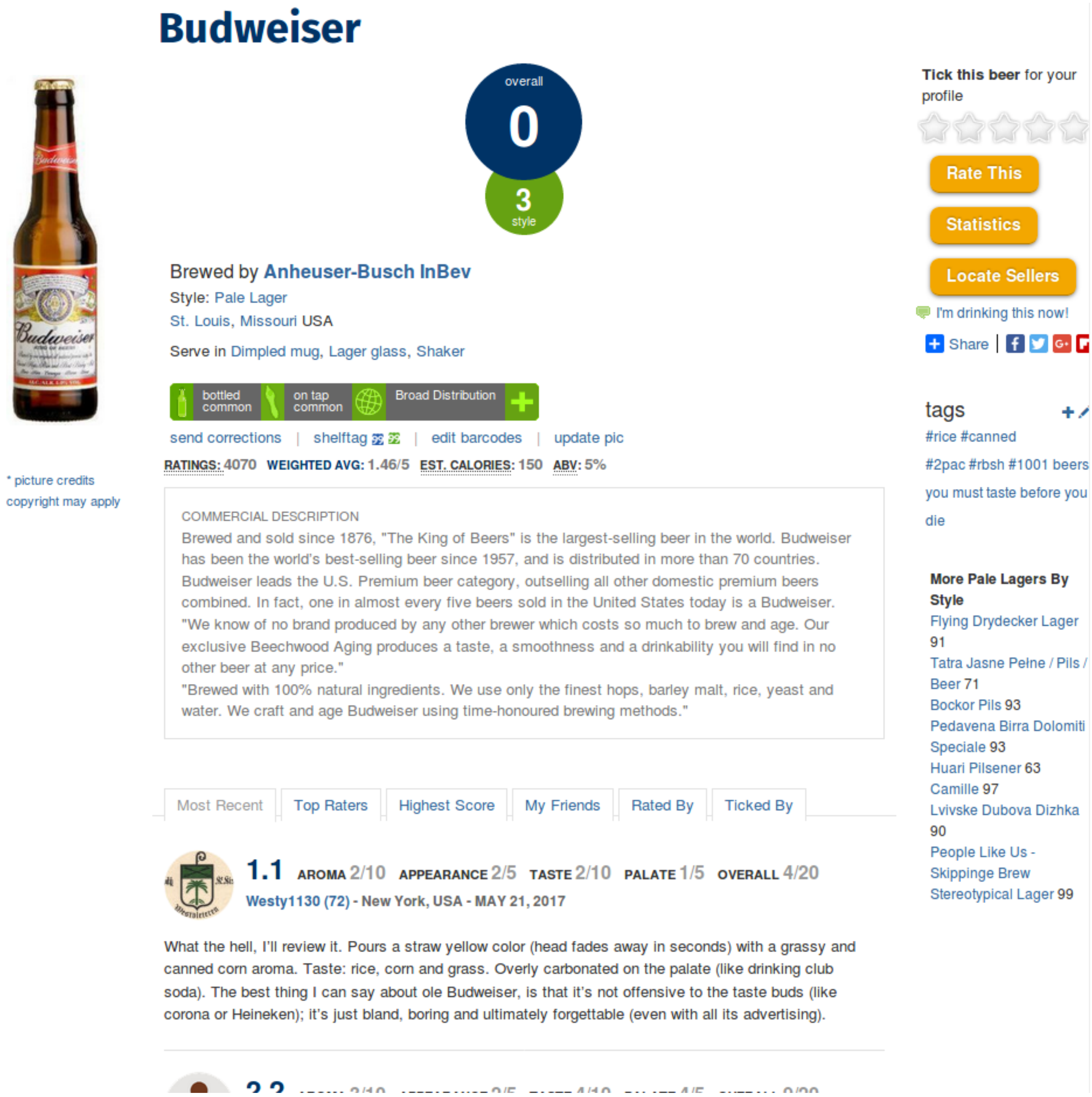
American Adjunct Lager

Budweiser
Anheuser-Busch

Pabst Blue Ribbon (PBR)
Pabst Brewing Company

Corona Extra
Grupo Modelo S.A. de C.V.

Figure 53: Webpage of the Budweiser beer on BeerAdvocate. Screenshot taken on the 23rd of May 2017 at 3:40 pm. Link to the webpage.



Budweiser

overall **0**

3 style

Brewed by **Anheuser-Busch InBev**

Style: Pale Lager

St. Louis, Missouri USA

Serve in Dimpled mug, Lager glass, Shaker

bottled common | on tap common | Broad Distribution

send corrections | shelftag | edit barcodes | update pic

RATINGS: 4070 **WEIGHTED AVG:** 1.46/5 **EST. CALORIES:** 150 **ABV:** 5%

* picture credits
copyright may apply

COMMERCIAL DESCRIPTION

Brewed and sold since 1876, "The King of Beers" is the largest-selling beer in the world. Budweiser has been the world's best-selling beer since 1957, and is distributed in more than 70 countries. Budweiser leads the U.S. Premium beer category, outselling all other domestic premium beers combined. In fact, one in almost every five beers sold in the United States today is a Budweiser. "We know of no brand produced by any other brewer which costs so much to brew and age. Our exclusive Beechwood Aging produces a taste, a smoothness and a drinkability you will find in no other beer at any price."

"Brewed with 100% natural ingredients. We use only the finest hops, barley malt, rice, yeast and water. We craft and age Budweiser using time-honoured brewing methods."

Most Recent | Top Raters | Highest Score | My Friends | Rated By | Ticked By

1.1 **AROMA** 2/10 **APPEARANCE** 2/5 **TASTE** 2/10 **PALATE** 1/5 **OVERALL** 4/20

Westy1130 (72) - New York, USA - MAY 21, 2017

What the hell, I'll review it. Pours a straw yellow color (head fades away in seconds) with a grassy and canned corn aroma. Taste: rice, corn and grass. Overly carbonated on the palate (like drinking club soda). The best thing I can say about ole Budweiser, is that it's not offensive to the taste buds (like corona or Heineken); it's just bland, boring and ultimately forgettable (even with all its advertising).

Tick this beer for your profile

Rate This

Statistics

Locate Sellers

I'm drinking this now!

Share

tags

#rice #canned

#2pac #rbsh #1001 beers you must taste before you die

More Pale Lagers By Style

Flying Drydecker Lager 91

Tatra Jasne Pełne / Pils / Beer 71

Bockor Pils 93

Pedavena Birra Dolomiti Speciale 93

Huari Pilsener 63

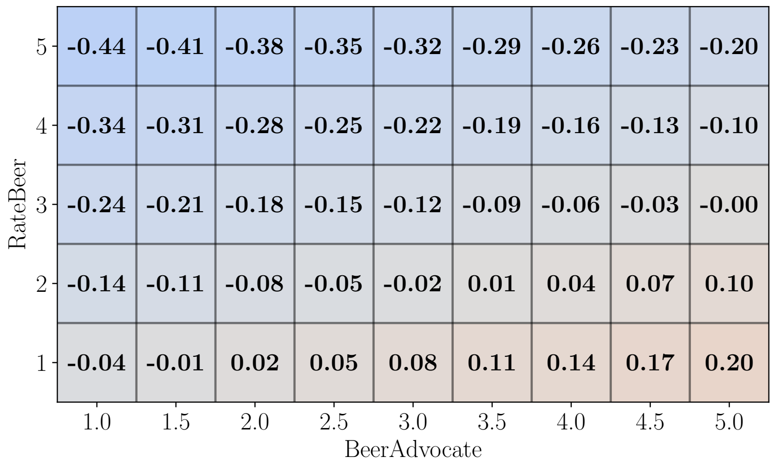
Camille 97

Lvivske Dubova Dizhka 90

People Like Us - Skippinge Brew Stereotypical Lager 99

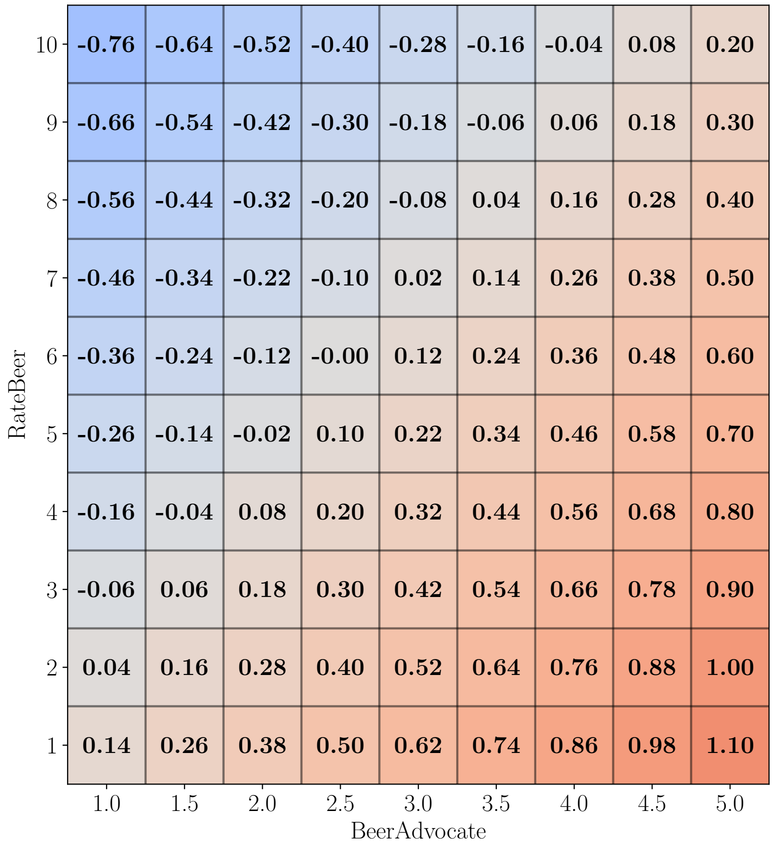
Figure 54: Webpage of the Budweiser beer on RateBeer. Screenshot taken on the 23rd of May 2017 at 3:40 pm. Link to the webpage.

Effects of the Aspects on the Final Rating



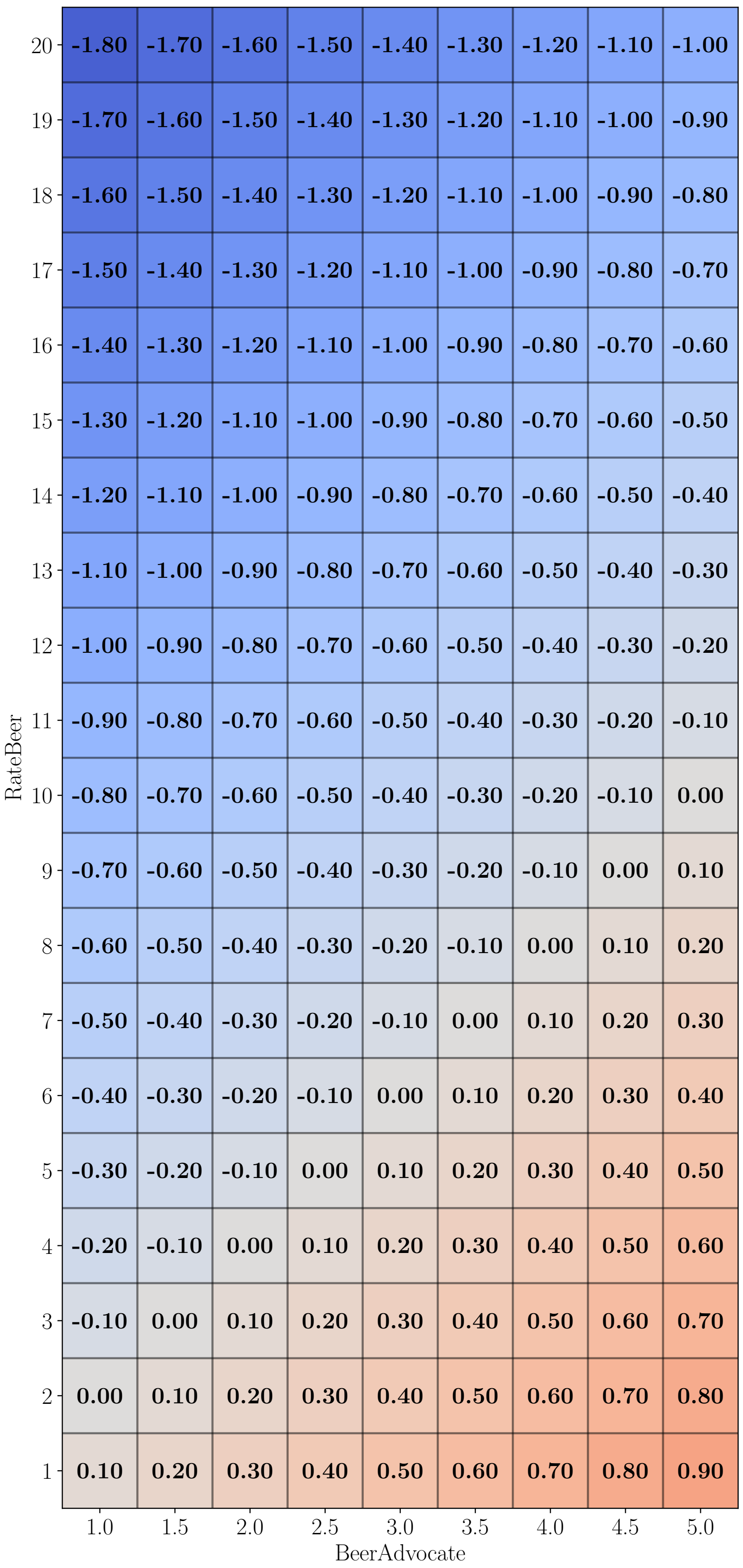
RateBeer	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
5	-0.44	-0.41	-0.38	-0.35	-0.32	-0.29	-0.26	-0.23	-0.20
4	-0.34	-0.31	-0.28	-0.25	-0.22	-0.19	-0.16	-0.13	-0.10
3	-0.24	-0.21	-0.18	-0.15	-0.12	-0.09	-0.06	-0.03	-0.00
2	-0.14	-0.11	-0.08	-0.05	-0.02	0.01	0.04	0.07	0.10
1	-0.04	-0.01	0.02	0.05	0.08	0.11	0.14	0.17	0.20

Figure 55: Effects of the aspect **appearance** on the final rating. The ratings are weighted with the weights given in Table 4. If the value is negative, it means that the RateBeer rating will be higher than the BeerAdvocate rating. If the value is positive, it is the inverse.



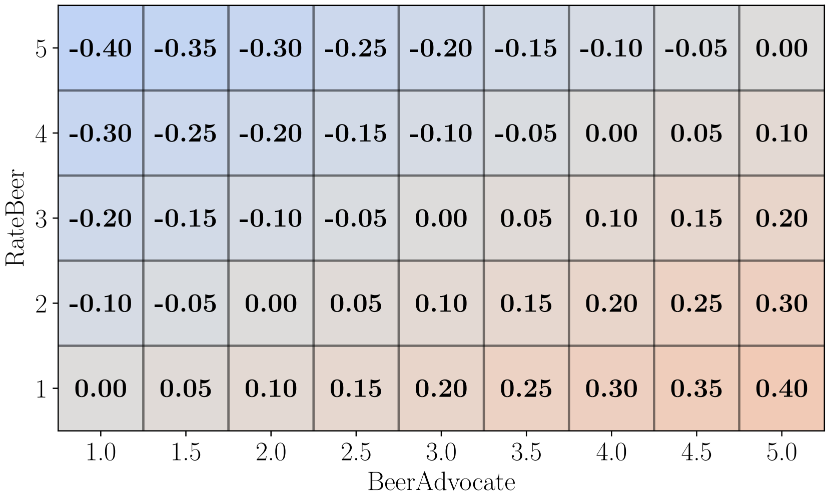
RateBeer	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
10	-0.76	-0.64	-0.52	-0.40	-0.28	-0.16	-0.04	0.08	0.20
9	-0.66	-0.54	-0.42	-0.30	-0.18	-0.06	0.06	0.18	0.30
8	-0.56	-0.44	-0.32	-0.20	-0.08	0.04	0.16	0.28	0.40
7	-0.46	-0.34	-0.22	-0.10	0.02	0.14	0.26	0.38	0.50
6	-0.36	-0.24	-0.12	-0.00	0.12	0.24	0.36	0.48	0.60
5	-0.26	-0.14	-0.02	0.10	0.22	0.34	0.46	0.58	0.70
4	-0.16	-0.04	0.08	0.20	0.32	0.44	0.56	0.68	0.80
3	-0.06	0.06	0.18	0.30	0.42	0.54	0.66	0.78	0.90
2	0.04	0.16	0.28	0.40	0.52	0.64	0.76	0.88	1.00
1	0.14	0.26	0.38	0.50	0.62	0.74	0.86	0.98	1.10

Figure 56: Effects of the aspect **aroma** on the final rating. The ratings are weighted with the weights given in Table 4. If the value is negative, it means that the RateBeer rating will be higher than the BeerAdvocate rating. If the value is positive, it is the inverse.



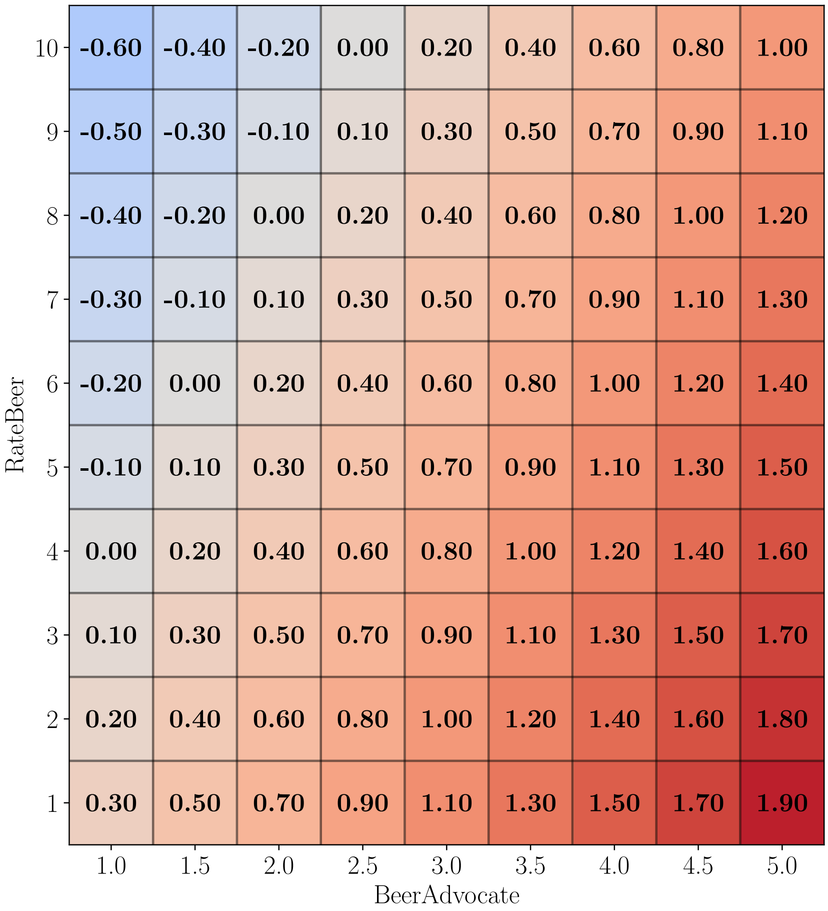
RateBeer \ BeerAdvocate	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
20	-1.80	-1.70	-1.60	-1.50	-1.40	-1.30	-1.20	-1.10	-1.00
19	-1.70	-1.60	-1.50	-1.40	-1.30	-1.20	-1.10	-1.00	-0.90
18	-1.60	-1.50	-1.40	-1.30	-1.20	-1.10	-1.00	-0.90	-0.80
17	-1.50	-1.40	-1.30	-1.20	-1.10	-1.00	-0.90	-0.80	-0.70
16	-1.40	-1.30	-1.20	-1.10	-1.00	-0.90	-0.80	-0.70	-0.60
15	-1.30	-1.20	-1.10	-1.00	-0.90	-0.80	-0.70	-0.60	-0.50
14	-1.20	-1.10	-1.00	-0.90	-0.80	-0.70	-0.60	-0.50	-0.40
13	-1.10	-1.00	-0.90	-0.80	-0.70	-0.60	-0.50	-0.40	-0.30
12	-1.00	-0.90	-0.80	-0.70	-0.60	-0.50	-0.40	-0.30	-0.20
11	-0.90	-0.80	-0.70	-0.60	-0.50	-0.40	-0.30	-0.20	-0.10
10	-0.80	-0.70	-0.60	-0.50	-0.40	-0.30	-0.20	-0.10	0.00
9	-0.70	-0.60	-0.50	-0.40	-0.30	-0.20	-0.10	0.00	0.10
8	-0.60	-0.50	-0.40	-0.30	-0.20	-0.10	0.00	0.10	0.20
7	-0.50	-0.40	-0.30	-0.20	-0.10	0.00	0.10	0.20	0.30
6	-0.40	-0.30	-0.20	-0.10	0.00	0.10	0.20	0.30	0.40
5	-0.30	-0.20	-0.10	0.00	0.10	0.20	0.30	0.40	0.50
4	-0.20	-0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60
3	-0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70
2	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80
1	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90

Figure 57: Effects of the aspect **overall** on the final rating. The ratings are weighted with the weights given in Table 4. If the value is negative, it means that the RateBeer rating will be higher than the BeerAdvocate rating. If the value is positive, it is the inverse.



RateBeer \ BeerAdvocate	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
5	-0.40	-0.35	-0.30	-0.25	-0.20	-0.15	-0.10	-0.05	0.00
4	-0.30	-0.25	-0.20	-0.15	-0.10	-0.05	0.00	0.05	0.10
3	-0.20	-0.15	-0.10	-0.05	0.00	0.05	0.10	0.15	0.20
2	-0.10	-0.05	0.00	0.05	0.10	0.15	0.20	0.25	0.30
1	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40

Figure 58: Effects of the aspect **palate** on the final rating. The ratings are weighted with the weights given in Table 4. If the value is negative, it means that the RateBeer rating will be higher than the BeerAdvocate rating. If the value is positive, it is the inverse.



RateBeer \ BeerAdvocate	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
10	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00
9	-0.50	-0.30	-0.10	0.10	0.30	0.50	0.70	0.90	1.10
8	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20
7	-0.30	-0.10	0.10	0.30	0.50	0.70	0.90	1.10	1.30
6	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40
5	-0.10	0.10	0.30	0.50	0.70	0.90	1.10	1.30	1.50
4	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60
3	0.10	0.30	0.50	0.70	0.90	1.10	1.30	1.50	1.70
2	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	1.80
1	0.30	0.50	0.70	0.90	1.10	1.30	1.50	1.70	1.90

Figure 59: Effects of the aspect **taste** on the final rating. The ratings are weighted with the weights given in Table 4. If the value is negative, it means that the RateBeer rating will be higher than the BeerAdvocate rating. If the value is positive, it is the inverse.

Influence of the Scores on the Ratings - Supplementary Figures

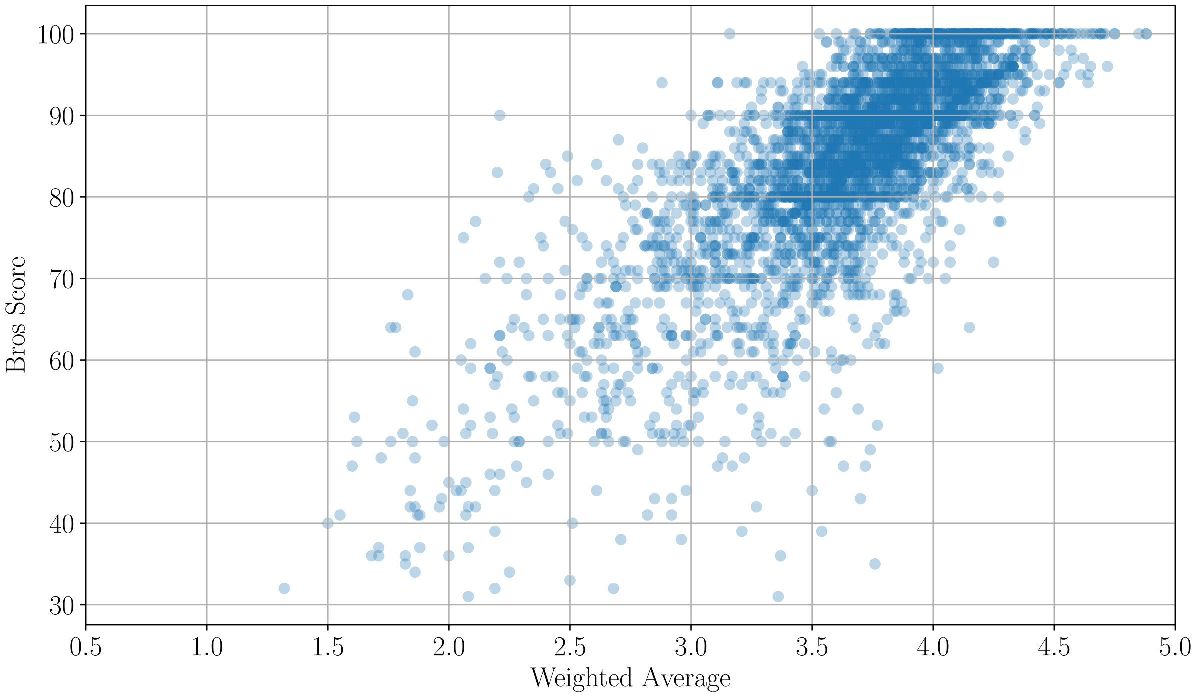


Figure 60: Bros score in function of the weighted average in BeerAdvocate.

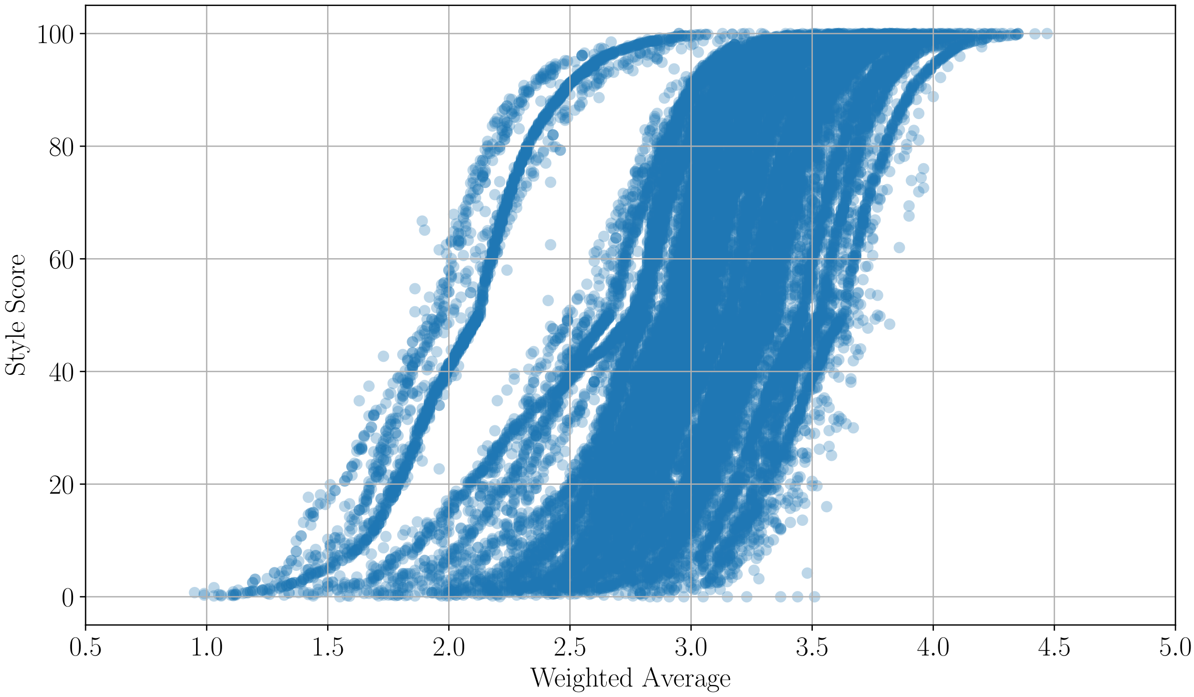


Figure 61: Style score in function of the weighted average in RateBeer.

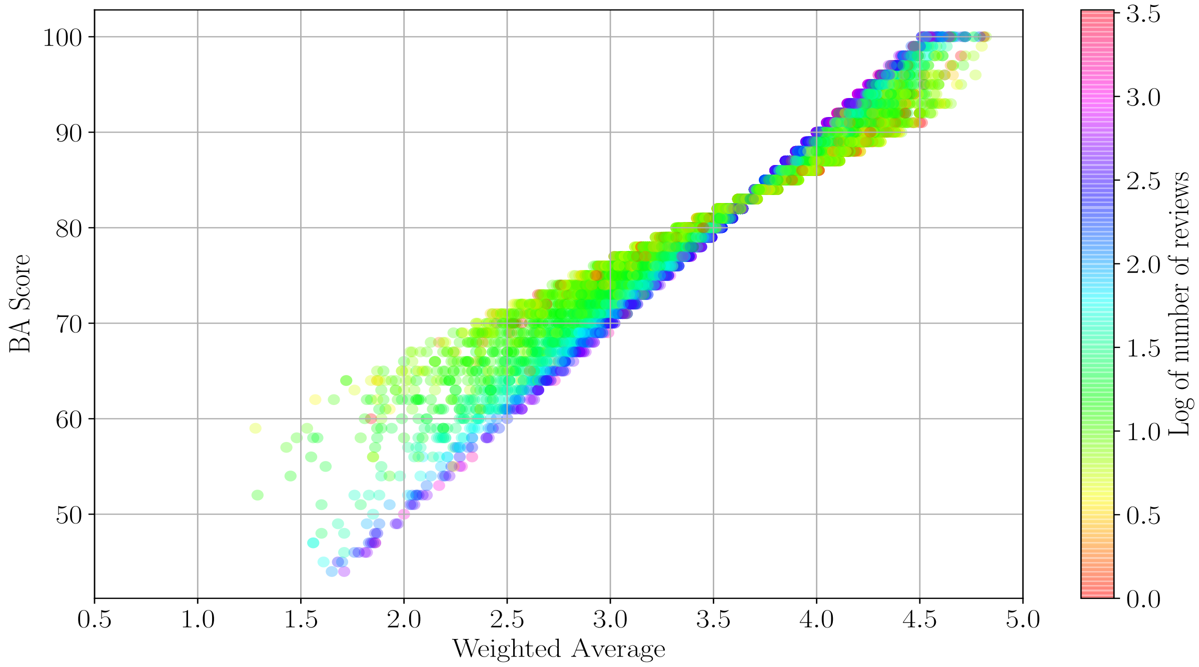


Figure 62: Score in function of the weighted average in BeerAdvocate. Colors are in function of the log of the number of reviews.

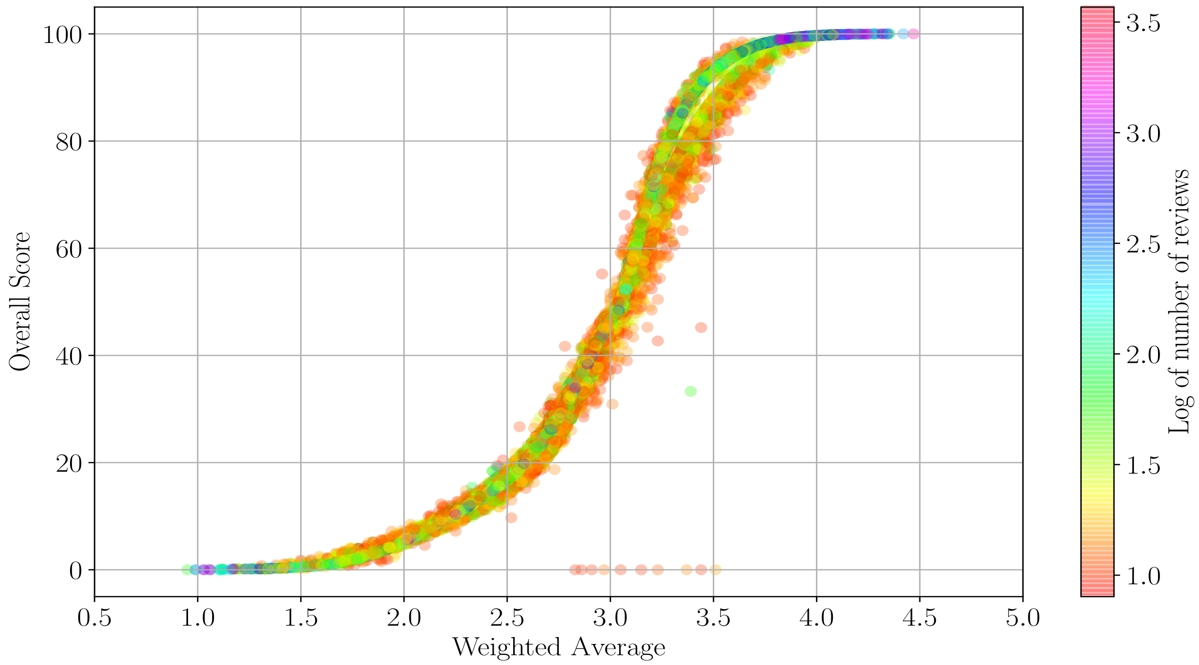


Figure 63: Score in function of the weighted average in RateBeer. Colors are in function of the log of the number of reviews.

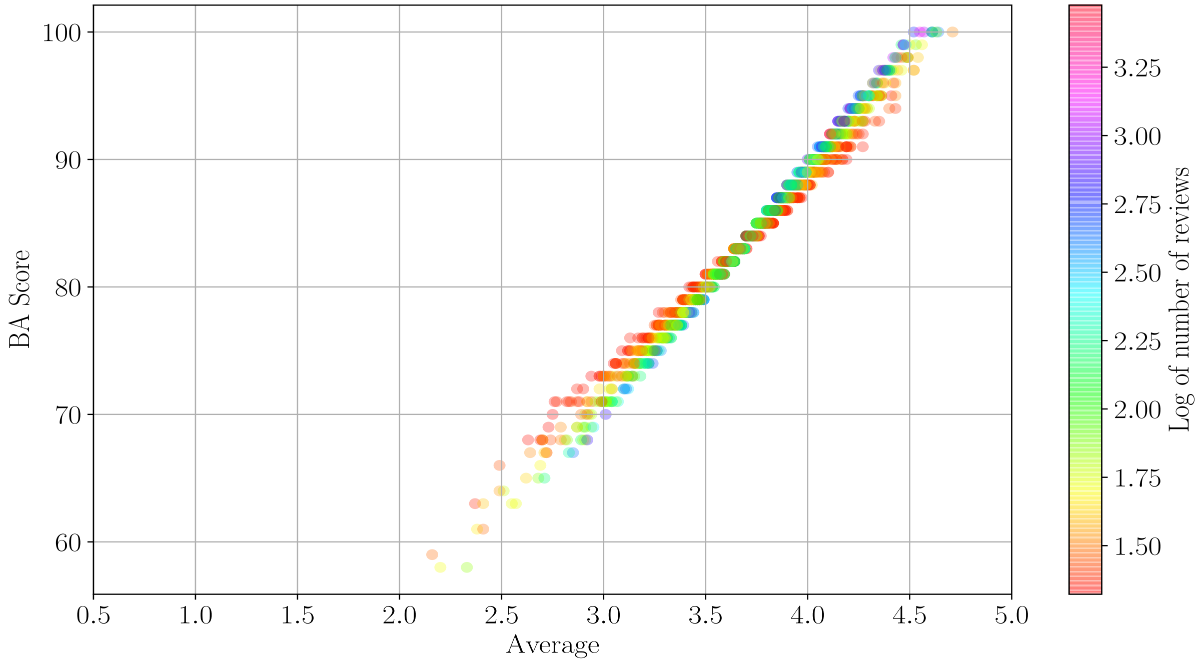


Figure 64: Score in function of the average in BeerAdvocate. A minimum of 20 reviews per beer is required. Colors are in function of the log of the number of reviews.

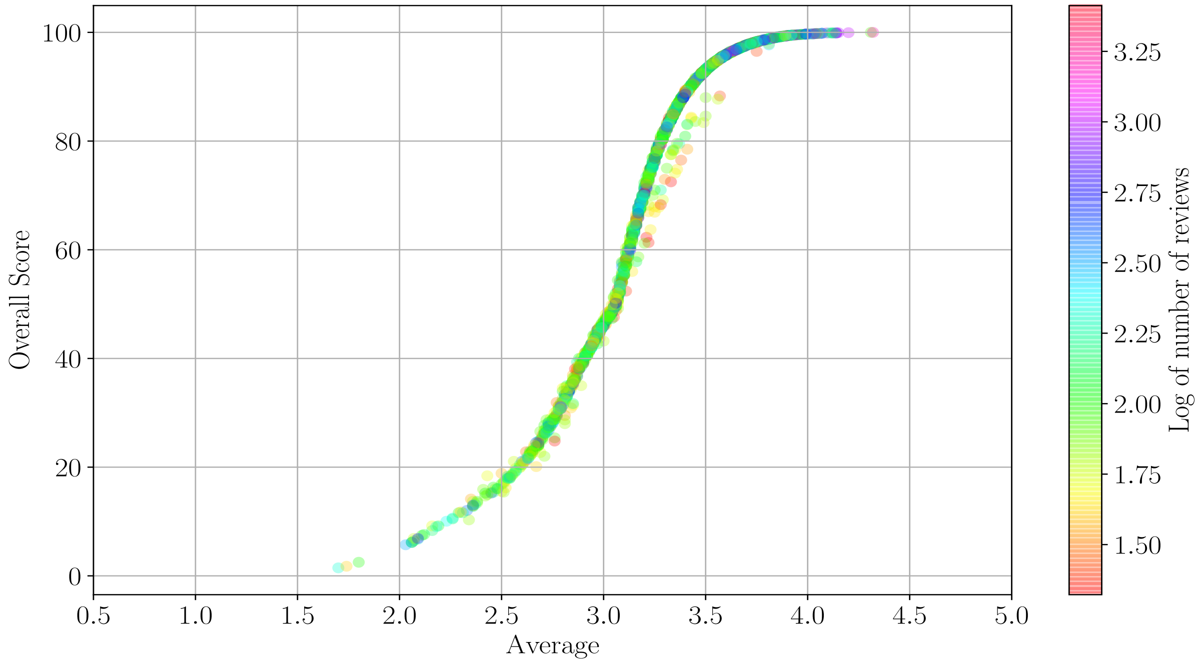


Figure 65: Score in function of the average in RateBeer. A minimum of 20 reviews per beer is required. Colors are in function of the log of the number of reviews.

Standard Deviation of First Ten Ratings by Year

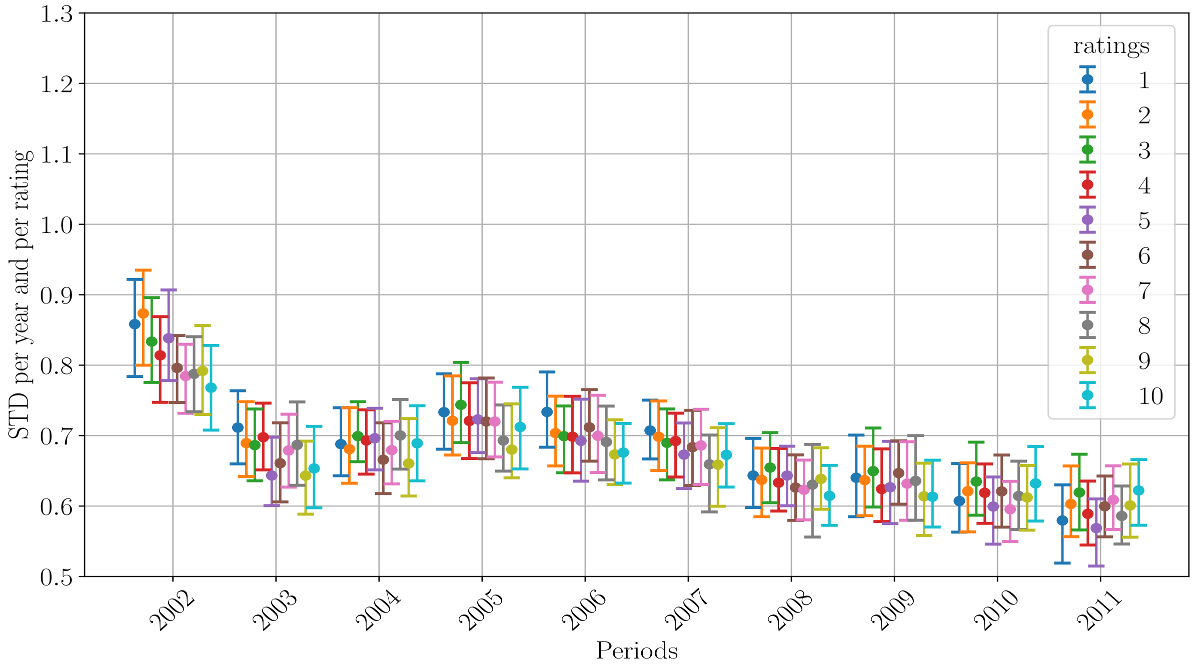


Figure 66: Standard deviation by year and by rating for BeerAdvocate.

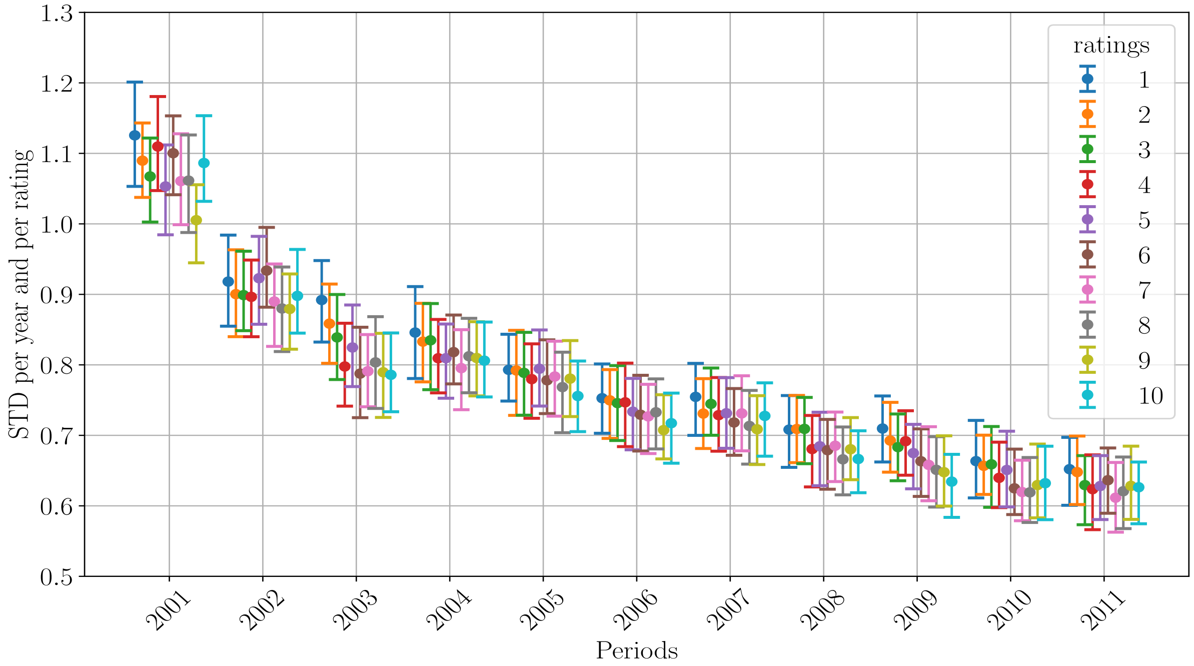
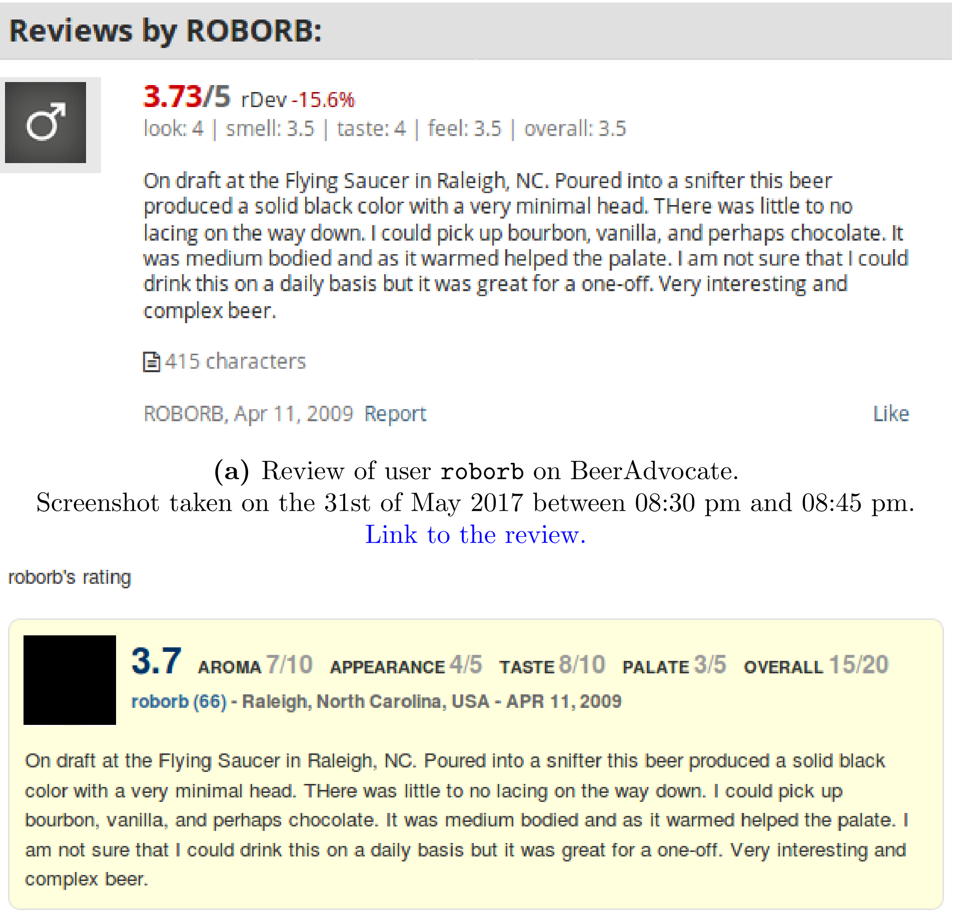


Figure 67: Standard deviation by year and by rating for RateBeer.

Examples of Reviews



Reviews by ROBORB:

3.73/5 rDev -15.6%
look: 4 | smell: 3.5 | taste: 4 | feel: 3.5 | overall: 3.5

On draft at the Flying Saucer in Raleigh, NC. Poured into a snifter this beer produced a solid black color with a very minimal head. THere was little to no lacing on the way down. I could pick up bourbon, vanilla, and perhaps chocolate. It was medium bodied and as it warmed helped the palate. I am not sure that I could drink this on a daily basis but it was great for a one-off. Very interesting and complex beer.

415 characters

ROBORB, Apr 11, 2009 Report

(a) Review of user `roborb` on BeerAdvocate.
Screenshot taken on the 31st of May 2017 between 08:30 pm and 08:45 pm.
Link to the review.

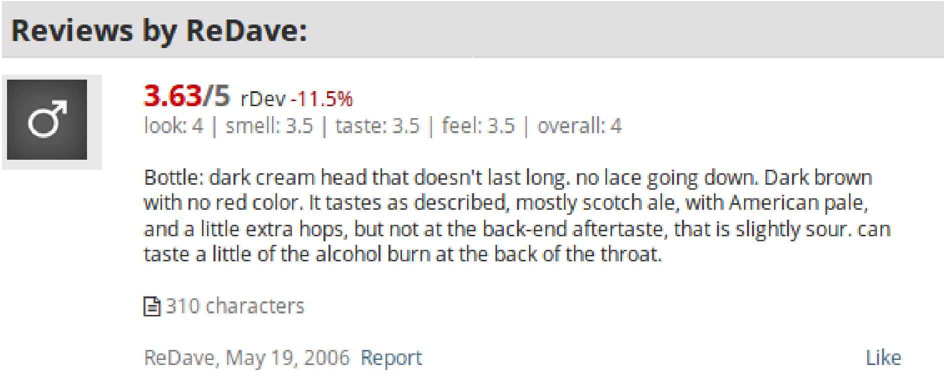
roborb's rating

3.7 AROMA 7/10 APPEARANCE 4/5 TASTE 8/10 PALATE 3/5 OVERALL 15/20
roborb (66) - Raleigh, North Carolina, USA - APR 11, 2009

On draft at the Flying Saucer in Raleigh, NC. Poured into a snifter this beer produced a solid black color with a very minimal head. THere was little to no lacing on the way down. I could pick up bourbon, vanilla, and perhaps chocolate. It was medium bodied and as it warmed helped the palate. I am not sure that I could drink this on a daily basis but it was great for a one-off. Very interesting and complex beer.

(b) Review of user `roborb` on RateBeer.
Screenshot taken on the 31st of May 2017 between 08:30 pm and 08:45 pm.
Link to the review.

Figure 68: Reviews of user `roborb` on the beer *Duck-Rabbit Paul's Day Off* from the brewery *The Duck-Rabbit Craft Brewery.*



Reviews by ReDave:

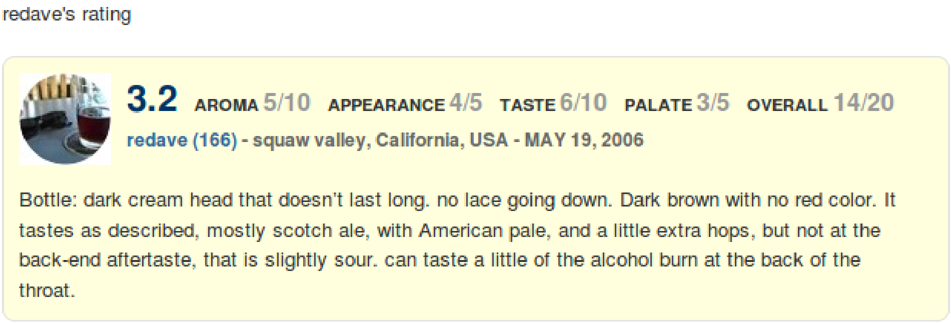
3.63/5 rDev -11.5%
look: 4 | smell: 3.5 | taste: 3.5 | feel: 3.5 | overall: 4

Bottle: dark cream head that doesn't last long. no lace going down. Dark brown with no red color. It tastes as described, mostly scotch ale, with American pale, and a little extra hops, but not at the back-end aftertaste, that is slightly sour. can taste a little of the alcohol burn at the back of the throat.

310 characters

ReDave, May 19, 2006 Report Like

(a) Review of user `redave` on BeerAdvocate.
Screenshot taken on the 31st of May 2017 between 08:30 pm and 08:45 pm.
Link to the review.



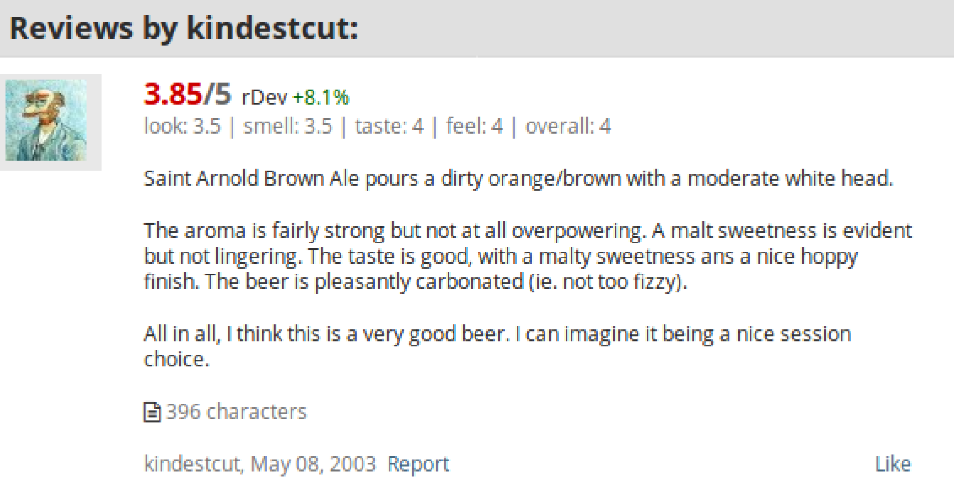
redave's rating

3.2 AROMA 5/10 APPEARANCE 4/5 TASTE 6/10 PALATE 3/5 OVERALL 14/20
redave (166) - squaw valley, California, USA - MAY 19, 2006

Bottle: dark cream head that doesn't last long. no lace going down. Dark brown with no red color. It tastes as described, mostly scotch ale, with American pale, and a little extra hops, but not at the back-end aftertaste, that is slightly sour. can taste a little of the alcohol burn at the back of the throat.

(b) Review of user `redave` on RateBeer.
Screenshot taken on the 31st of May 2017 between 08:30 pm and 08:45 pm.
Link to the review.

Figure 69: Reviews of user `redave` on the beer *Indian Brown Ale* from the brewery *Dogfish Head Craft Brewery*.



Reviews by kindestcut:

3.85/5 rDev +8.1%

look: 3.5 | smell: 3.5 | taste: 4 | feel: 4 | overall: 4

Saint Arnold Brown Ale pours a dirty orange/brown with a moderate white head.

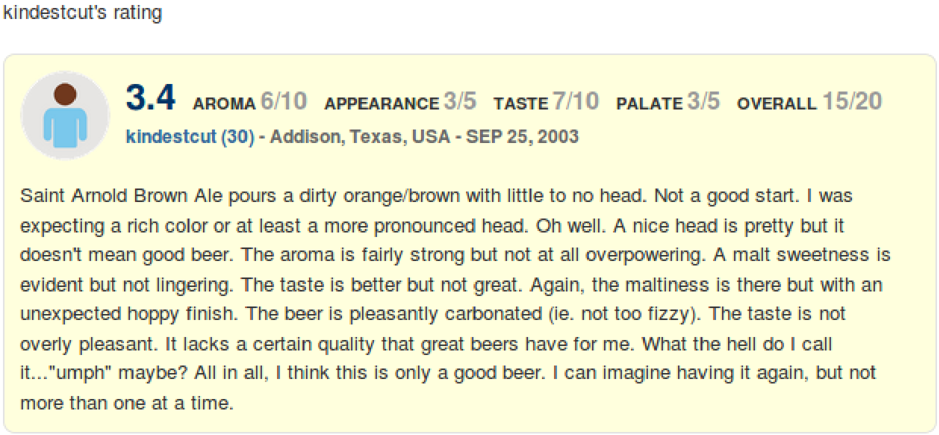
The aroma is fairly strong but not at all overpowering. A malt sweetness is evident but not lingering. The taste is good, with a malty sweetness ans a nice hoppy finish. The beer is pleasantly carbonated (ie. not too fizzy).

All in all, I think this is a very good beer. I can imagine it being a nice session choice.

396 characters

kindestcut, May 08, 2003 Report

(a) Review of user `kindestcut` on BeerAdvocate.
Screenshot taken on the 31st of May 2017 between 08:30 pm and 08:45 pm.
Link to the review.



kindestcut's rating

3.4 AROMA 6/10 APPEARANCE 3/5 TASTE 7/10 PALATE 3/5 OVERALL 15/20

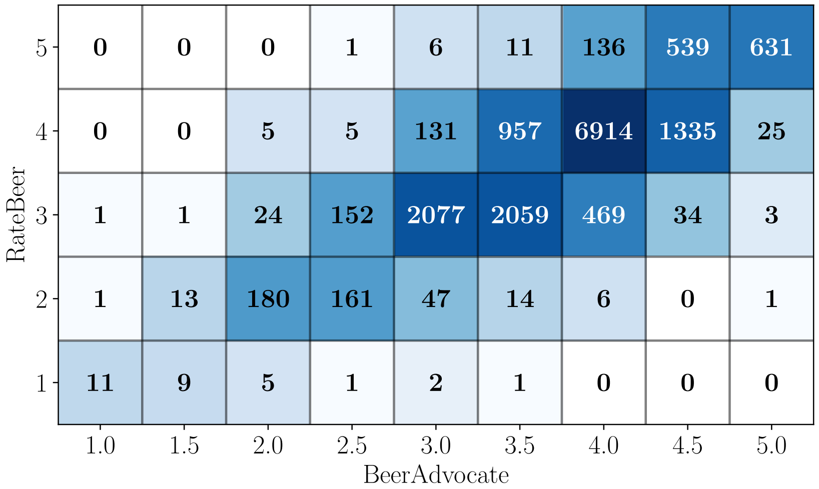
kindestcut (30) - Addison, Texas, USA - SEP 25, 2003

Saint Arnold Brown Ale pours a dirty orange/brown with little to no head. Not a good start. I was expecting a rich color or at least a more pronounced head. Oh well. A nice head is pretty but it doesn't mean good beer. The aroma is fairly strong but not at all overpowering. A malt sweetness is evident but not lingering. The taste is better but not great. Again, the maltiness is there but with an unexpected hoppy finish. The beer is pleasantly carbonated (ie. not too fizzy). The taste is not overly pleasant. It lacks a certain quality that great beers have for me. What the hell do I call it..."umph" maybe? All in all, I think this is only a good beer. I can imagine having it again, but not more than one at a time.

(b) Review of user `kindestcut` on RateBeer.
Screenshot taken on the 31st of May 2017 between 08:30 pm and 08:45 pm.
Link to the review.

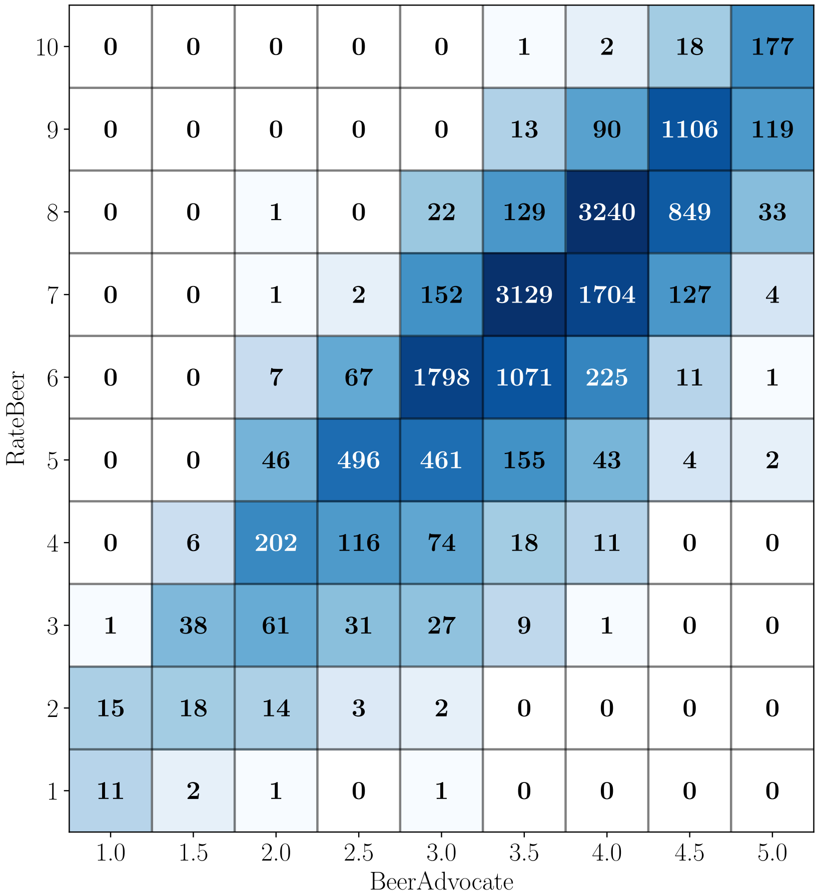
Figure 70: Reviews of user `kindestcut` on the beer *Saint Arnold Brown Ale* from the brewery *Saint Arnold Brewing Company*.

Correlation of the Aspects



RateBeer	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
5	0	0	0	1	6	11	136	539	631
4	0	0	5	5	131	957	6914	1335	25
3	1	1	24	152	2077	2059	469	34	3
2	1	13	180	161	47	14	6	0	1
1	11	9	5	1	2	1	0	0	0

Figure 71: RateBeer rating in function of the BeerAdvocate rating for the **appearance**. Pearson's correlation coefficient is 0.818.



RateBeer	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
10	0	0	0	0	0	1	2	18	177
9	0	0	0	0	0	13	90	1106	119
8	0	0	1	0	22	129	3240	849	33
7	0	0	1	2	152	3129	1704	127	4
6	0	0	7	67	1798	1071	225	11	1
5	0	0	46	496	461	155	43	4	2
4	0	6	202	116	74	18	11	0	0
3	1	38	61	31	27	9	1	0	0
2	15	18	14	3	2	0	0	0	0
1	11	2	1	0	1	0	0	0	0

Figure 72: RateBeer rating in function of the BeerAdvocate rating for the **aroma**. Pearson's correlation coefficient is 0.862.

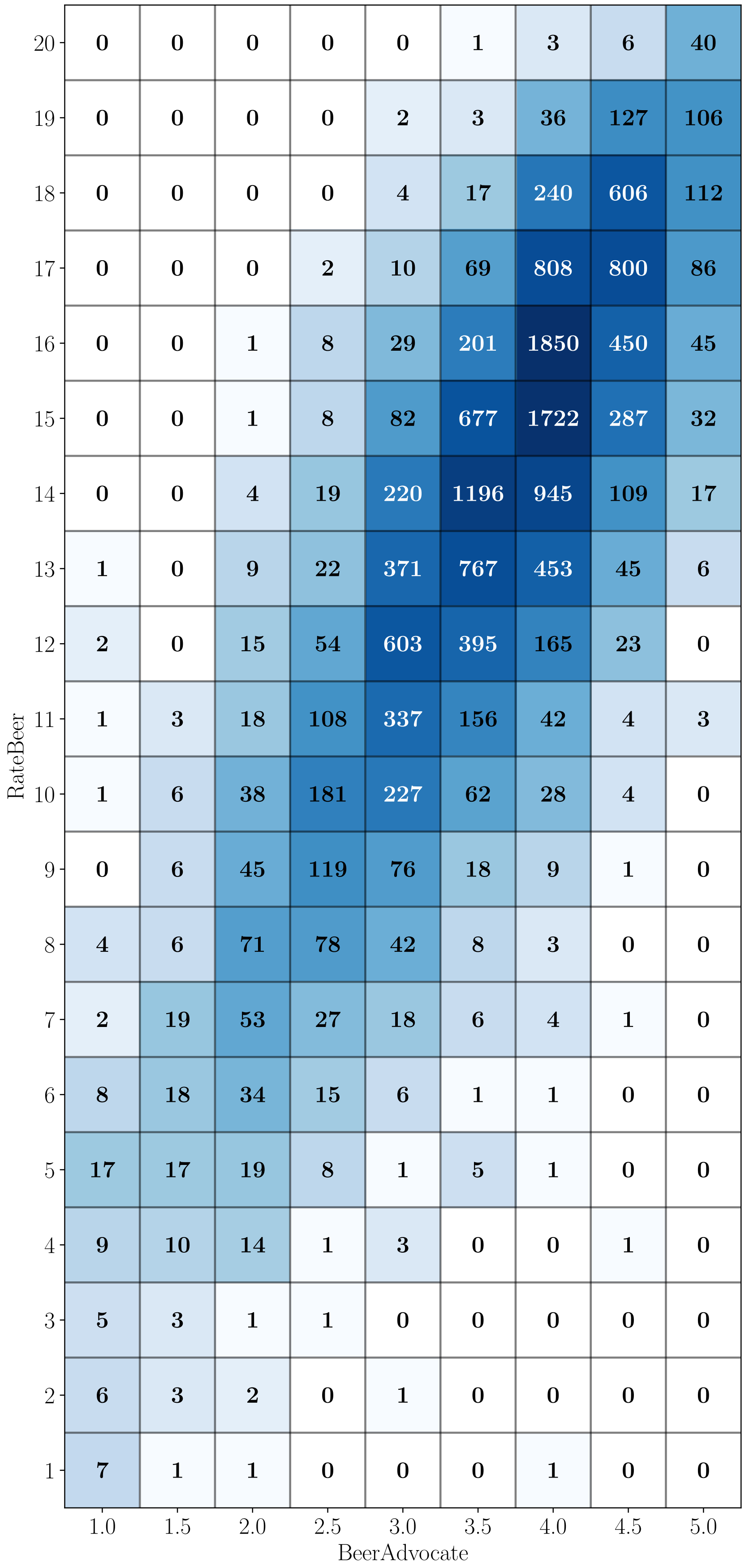
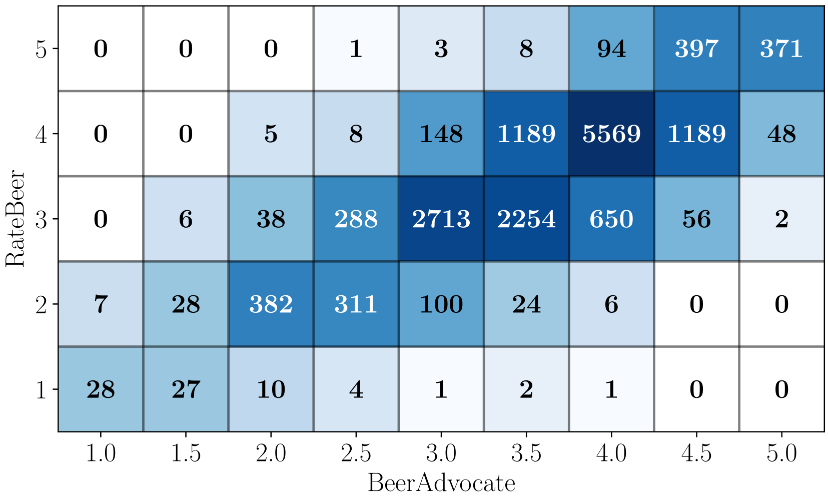
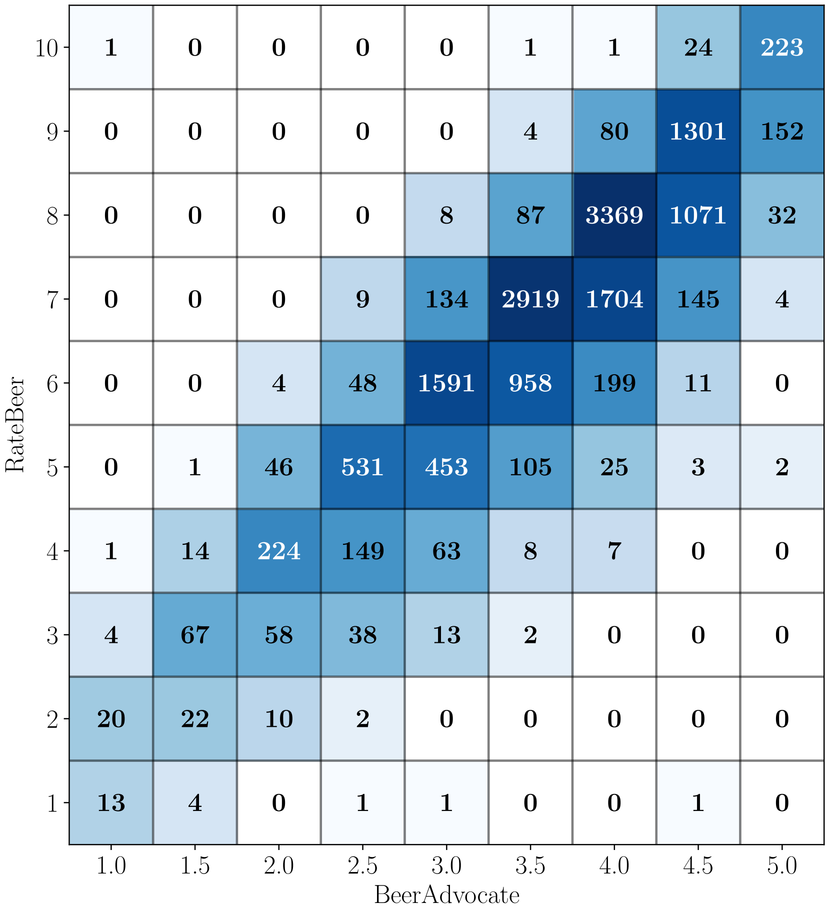


Figure 73: RateBeer rating in function of the BeerAdvocate rating for the **overall**. Pearson's correlation coefficient is 0.794.



RateBeer \ BeerAdvocate	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
5	0	0	0	1	3	8	94	397	371
4	0	0	5	8	148	1189	5569	1189	48
3	0	6	38	288	2713	2254	650	56	2
2	7	28	382	311	100	24	6	0	0
1	28	27	10	4	1	2	1	0	0

Figure 74: RateBeer rating in function of the BeerAdvocate rating for the **palate**. Pearson's correlation coefficient is 0.816.



RateBeer \ BeerAdvocate	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
10	1	0	0	0	0	1	1	24	223
9	0	0	0	0	0	4	80	1301	152
8	0	0	0	0	8	87	3369	1071	32
7	0	0	0	9	134	2919	1704	145	4
6	0	0	4	48	1591	958	199	11	0
5	0	1	46	531	453	105	25	3	2
4	1	14	224	149	63	8	7	0	0
3	4	67	58	38	13	2	0	0	0
2	20	22	10	2	0	0	0	0	0
1	13	4	0	1	1	0	0	1	0

Figure 75: RateBeer rating in function of the BeerAdvocate rating for the **taste**. Pearson's correlation coefficient is 0.886.

Multinomial Distributions of the Aspects' Differences

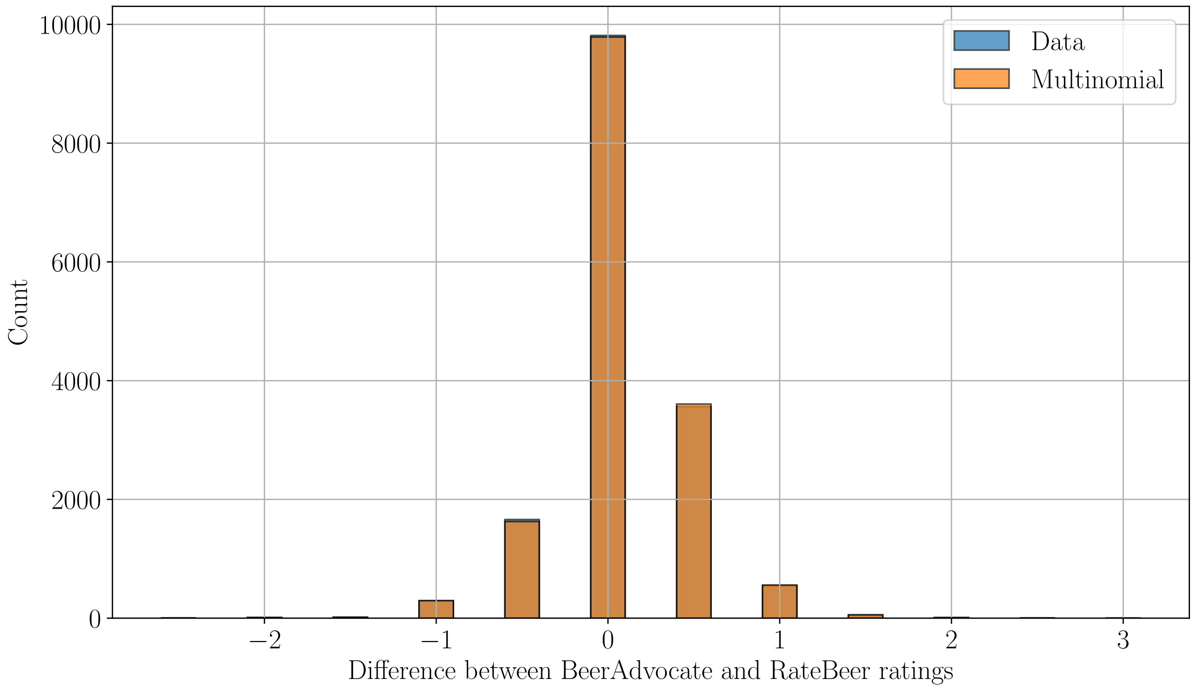


Figure 76: Multionomial distribution for the differences of ratings for the **appearance**. Differences are computed by subtracting the RateBeer rating to the BeerAdvocate rating.

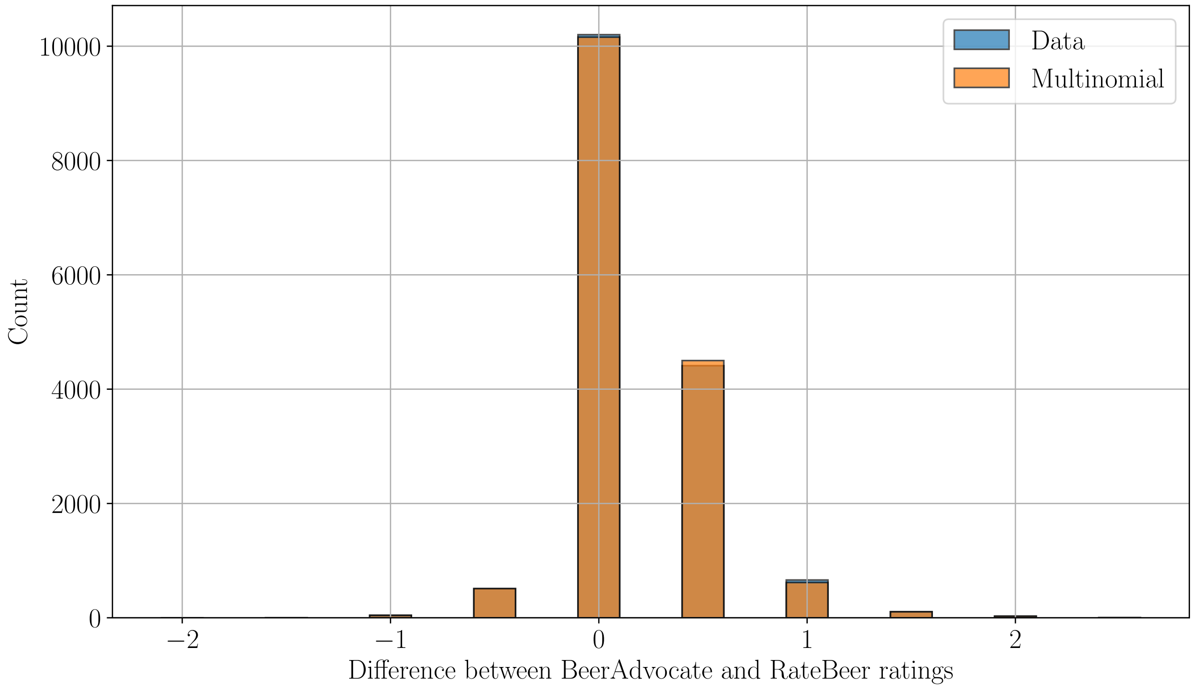


Figure 77: Multionomial distribution for the differences of ratings for the **aroma**. Differences are computed by subtracting the RateBeer rating to the BeerAdvocate rating.

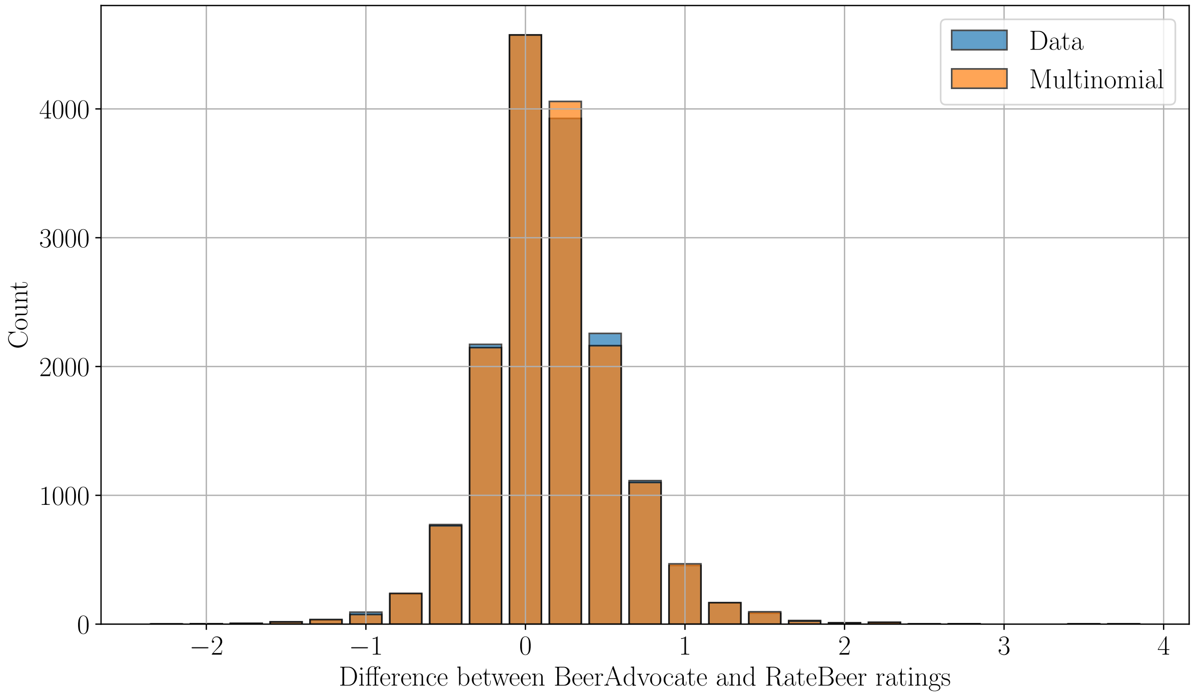


Figure 78: Multionomial distribution for the differences of ratings for the **overall**. Differences are computed by subtracting the RateBeer rating to the BeerAdvocate rating.

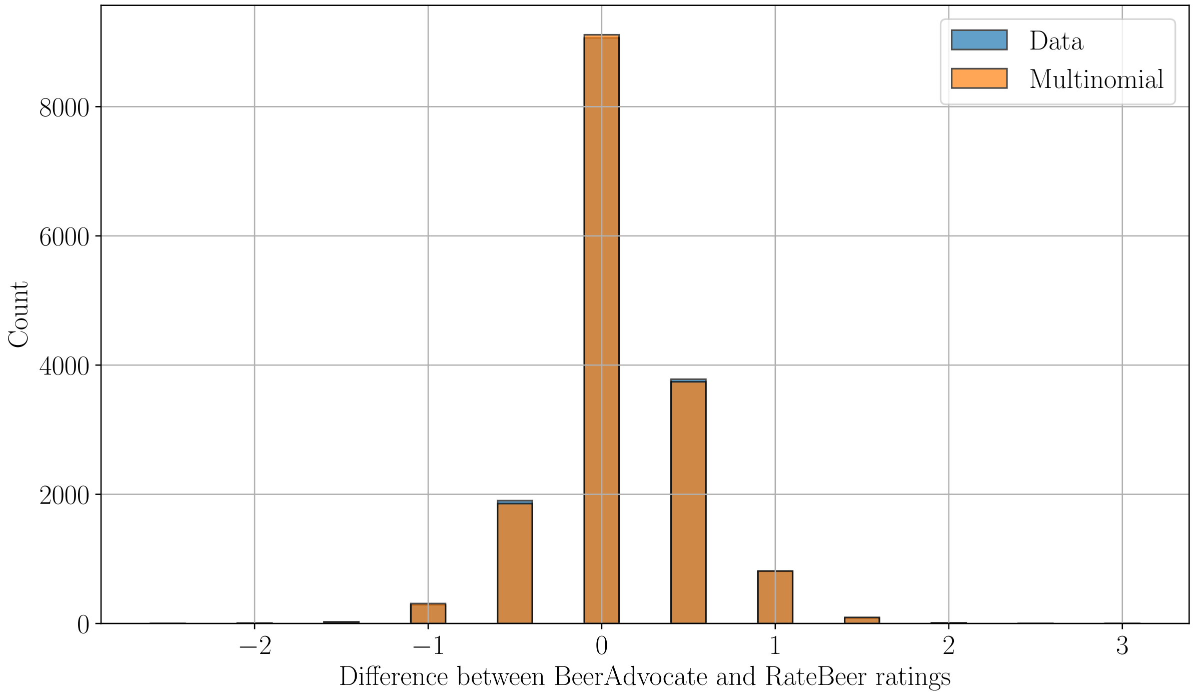


Figure 79: Multionomial distribution for the differences of ratings for the **palate**. Differences are computed by subtracting the RateBeer rating to the BeerAdvocate rating.

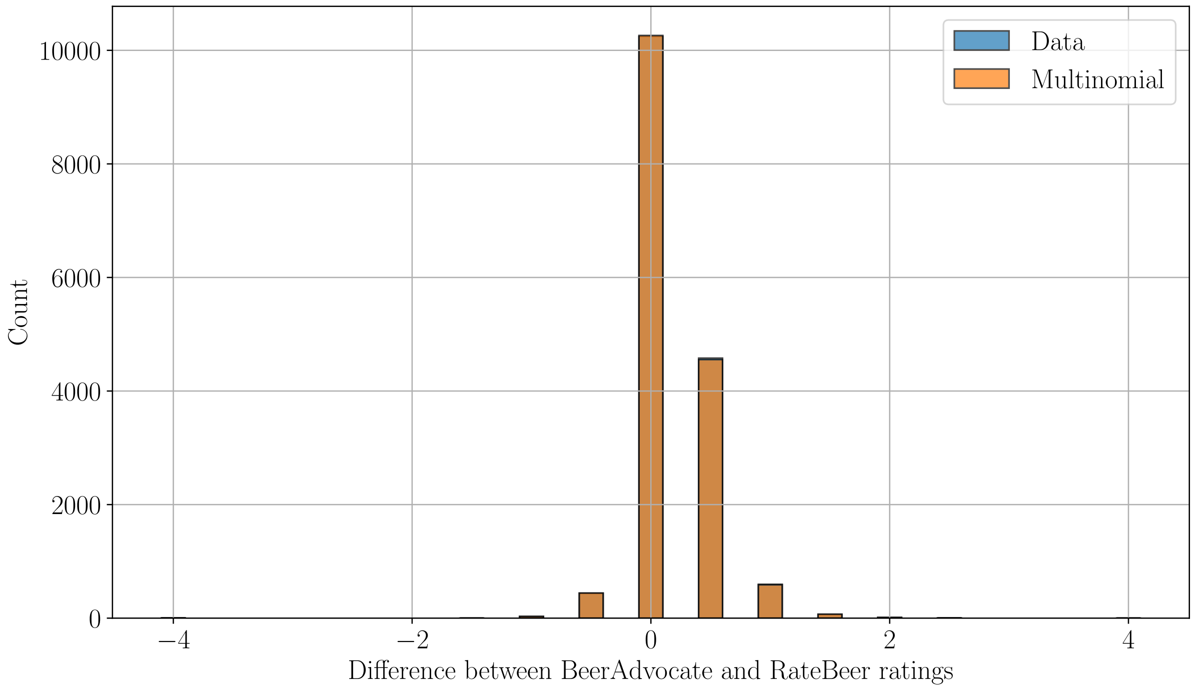


Figure 80: Multionomial distribution for the differences of ratings for the **taste**. Differences are computed by subtracting the RateBeer rating to the BeerAdvocate rating.