B2W-Reviews01
An open product reviews corpus

Livy Real[1], Marcio Oshiro[1], Alexandre Mafra[1]

[1]B2W Digital – Tech Labs
São Paulo – BR – Brazil

{livy.coelho,marcio.oshiro,alexandre.mafra}@b2wdigital.com

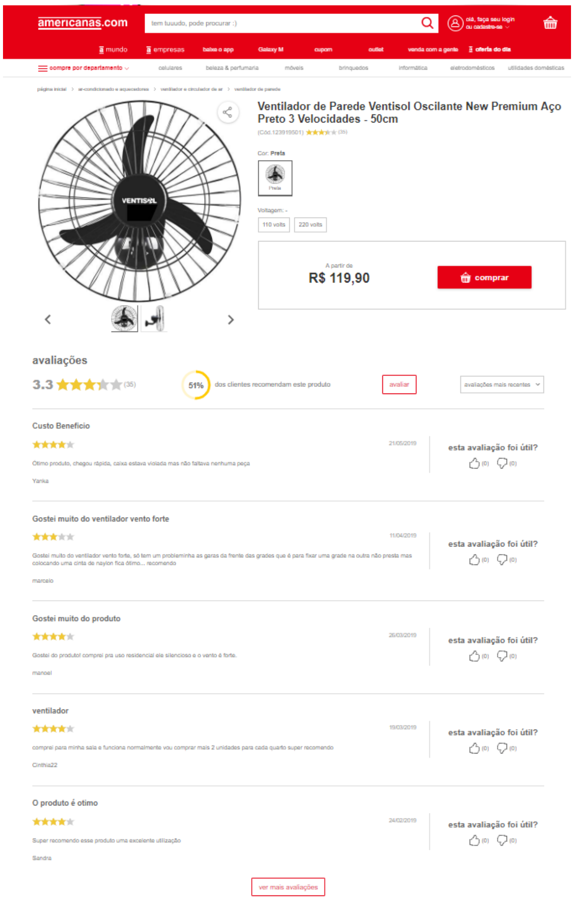
Abstract. *This paper describes* B2W-Reviews01*, an open corpus of product reviews.* B2W-Reviews01 *contains more than 130k e-commerce customer reviews, collected from the* Americanas.com *website between January and May, 2018.* B2W-Reviews01 *offers rich information about the reviewer profile, such as gender, age, and geographical location. The corpus also has two different review rates: the usual 5 point scale rate, represented by stars in most e-commerce websites, and also a 'recommend to a friend' label, a 'yes or no' question representing the willingness of the customer to recommend the product to someone else. By comparing these two rates, we found that the common approach of conducting sentiment analysis, based on a simplification over the 5 point scale, does not always reflect users' sentiments about the product. It suggests that, for production applications, the approach of analyzing the 5 point scale rate as a three level scale can lead to wrong conclusions.*

1. Introduction

In the era of machine learning and big data, one of the biggest bottlenecks of Natural Language Processing (NLP) and Computational Linguistics (CL) is having open corpora available. While there is a lot of information available on the internet, it is still difficult to have structured, high quality, curated data. This work aims to help fill this gap, presenting a new open corpus of product reviews in Brazilian Portuguese, the *B2W-Reviews01*.

In particular, customer product reviews represent a difficult information source for web crawlers, since there is no standardization on how to represent them on e-commerce websites. For example, on the *Americanas.com* website, reviews are displayed in product pages, but only the latest 5 left by customers automatically appear on that page. Another click is required to display 5 more reviews, as shown in figure 1.

Although customer generated data is often seen as a valuable by-product of online companies, few of them actually notice the value of making this information generally available. *B2W Digital* is a major e-commerce platform in Latin America. Its first e-commerce brand was released in 1999: *Americanas.com*. Today, the *B2W Digital* marketplace platform has three major brands in Brazilian e-commerce: *Americanas.com*, *Submarino*, and *Shoptime*. These brands support more than 25,500 sellers trading on its digital platform. For a company such as *B2W Digital*, being able to analyze and extract information from user reviews became a critical task. User reviews not only have a high impact on the reputation of products, sellers and services, but can also be seen as the first and most direct way to obtain feedback from customers. Although digital companies count on several techniques to track customer activity and assess customer satisfaction ratings, analyzing product reviews often represents the easiest way to get customer feedback. Review analysis becomes especially relevant when the customer journey works as expected and, therefore, the customer does not need to contact Customer Service.



Ventilador de Parede Ventisol Oscilante New Premium Aço Preto 3 Velocidades - 50cm

(Cód.123919501) ★★★★☆ (35)

Cor: Preta

Voltagem: -

A partir de
R$ 119,90

avaliações

3.3 ★★★★☆ (35)

51% dos clientes recomendam este produto

Custo Beneficio
★★★★☆ 21/05/2019
Otimo produto, chegou rápida, caixa estava violada mas não faltava nenhuma peça
Yanka
esta avaliação foi útil? (0) (0)

Gostei muito do ventilador vento forte
★★★☆☆ 11/04/2019
Gostei muito do ventilador vento forte, só tem um probleminha as garas da frente das grades que é para fixar uma grade na outra não presta mas colocando uma cinta de naylon fica ótimo... recomendo
marcelo
esta avaliação foi útil? (0) (0)

Gostei muito do produto
★★★★☆ 26/03/2019
Gostei do produto! comprei pra uso residencial ele silencioso e o vento é forte.
manoel
esta avaliação foi útil? (0) (0)

ventilador
★★★★☆ 19/03/2019
comprei para minha sala e funciona normalmente vou comprar mais 2 unidades para cada quarto super recomendo
Cinthia22
esta avaliação foi útil? (0) (0)

O produto é otimo
★★★★☆ 24/02/2019
Super recomendo esse produto uma excelente utilização
Sandra
esta avaliação foi útil? (0) (0)

Figure 1. Example of product and product reviews display on *Americanas.com*

This corpus is a set of product reviews submitted to *Americanas.com* from January through May, 2018. *B2W-Reviews01* has both the review text and the meta-data related to each review: dates and times, ratings, geographical locations and ages of reviewers. *B2W-Reviews01* is available at `https://github.com/b2wdigital/b2w-reviews01/` under the license CC BY-NC-SA 4.0[1], which means that licensees may only copy, distribute, display, work on and make derivative works and remixes based on it if they give credit to *B2W Digital* in the manner specified in this work. Also, licensees may only distribute derivative works under a license identical ("not more restrictive") to the license that governs the original work. Finally, licensees may use it for non-commercial purposes.

1.1. Aims and Usages

B2W-Reviews01 is a corpus which contains such varied information, that it can be useful for several NLP/CL tasks. The first that comes to mind is probably sentiment analysis. Sentiment analysis is the task of assigning a sentiment (or a position) to the content of a

[1] `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

given text. Thereunto, *B2W-Reviews01* offers two different evaluation ratings, described in section 3. Product reviews often have complex information, related not only to the product that was purchased, but also to the online shopping experience, payment methods, or even the product delivery process. Therefore, different facts and opinions can be extracted from such a corpus, and classifying sentiment may not be enough to capture the content of reviews [Wachsmuth et al. 2014]. For real world applications, dealing with topic modeling, user intent identification and feature extraction also become necessary. It is relevant to know not only the reviewer's sentiment, but also the object of this feeling.

Since *B2W-Reviews01* offers the exact text written by users, this corpus also offers rich material for those interested on out-of-vocabulary words, slang identification, or spell-checker tasks. For those interested on socio-linguistics analysis, the present corpus offers a rich possibility of crossing reviewer information considering gender, age and geographical location. One can, for example, find easily how negative or positive reviews are distributed among age groups or which product categories receive more reviews from women or men. It is also possible to conduct a study on bias in reviews by joining and aggregating data. Although *B2W-Reviews01* is mainly a product review dataset, we believe that important insights about the current language in use in the web register can be made, since *Americanas.com* customers are spread throughout Brazil and have different social backgrounds.

2. Previous Works

To the best of our knowledge, the biggest product review data available is the *Amazon Customer Reviews Dataset*[2], which was made available by the *Amazon.com*[3] website, containing more than 130M reviews in four different languages. A particularly interesting subset of the *Amazon Customer Reviews Dataset* corpus is the work of [Filatova 2012], who produced the *Sarcasm Amazon Reviews Corpus*[4], a crowd-sourced annotated corpus that contains both sarcastic and non-sarcastic reviews of the same product. Another dataset made available by a marketplace is the *Rakuten*[5] marketplace dataset, the major Japanese digital marketplace. *Rakuten* offers some 64M reviews in Japanese available upon request[6].

For Portuguese, the situation of freely available data is not great. We have the Brazilian E-Commerce Public Dataset by *Olist*[7], which has 100k product reviews, collected from 2016 to 2018 by *Olist*, which is also one of the biggest sellers in the *B2W Digital* marketplace. Although the Brazilian E-Commerce Public Dataset by *Olist* is open, it is distributed in several files, so anyone who is interested in capturing very basic information, such as the amount of positive or negative reviews per geographical location information, or what is the average product rating for a given category, has to process, join, and aggregate data. Although *Olist* and *B2W-Reviews01* data can be seen as two datasets of the same nature, the *Olist* corpus has information related to payment methods, for example, which *B2W-Reviews01* does not have, while our corpus offers more infor-

[2] https://s3.amazonaws.com/amazon-reviews-pds/readme.html

[3] https://www.amazon.com

[4] https://github.com/ef2020/SarcasmAmazonReviewsCorpus/wiki

[5] https://www.rakuten.co.jp

[6] https://rit.rakuten.co.jp/data_release

[7] https://www.kaggle.com/olistbr/brazilian-ecommerce

mation about the reviewer, making her/his gender and birth year available, for example. Therefore, one interested in the reviews considering the user profile will probably find more useful information in *B2W-Reviews01*.

Another corpus of reviews, collected from *Mercado Livre*, is the one used in [Hartmann et al. 2014], made available by its creators upon request. Few different works can also be found about the use of product reviews in Portuguese, such as [Avanço 2015], a Master's thesis on normalization and classification of products that uses the data from [Hartmann et al. 2014]. [Siqueira and Barros 2010] uses reviews collected from the *Ebit* services web-page to extract features and polarity from users' feedbacks, but the used corpus is not available. [Ribeiro et al. 2012] is another work on feature extraction and opinion classification from reviews obtained at the *Carros na Web*[8] blog. However, collected data is also not available. [Nobre et al. 2016] also works on sentiment analysis using reviews collected from *Amazon.com* and *Buscapé*[9], but, yet again, the used corpora are not available.

Considering the amount of work done in information extraction, customer review analysis, and the fact that most of the works use different, not always available, corpora, we believe *B2W-Reviews01* can be really useful for the NLP Portuguese community. At least, this data can serve as a public and generic dataset, and further works interested on opinion mining and topic modeling of product reviews, among other subjects, could be compared and reproduced.

3. Corpus Description

The *B2W-Reviews01* corpus has 132,373 reviews, left by 112,993 different users regarding 48,001 unique products. The reviews were collected from January to May, 2018. All reviews submitted to the *Americanas.com* website are present in the corpus. It means that one can find offensive language, repeated reviews and reviews composed by only one word in the present data. These kinds of reviews are often not accepted to be displayed on the *Americanas.com* website. So, this means that this resource is richer than one could get crawling the *Americanas.com* website.

The reviews present in *B2W-Reviews01* have 3,160,781 tokens and 148,541 unique tokens[10]. The median number of tokens per review is 16 tokens, while the minimum found is 1 token and the maximum 795. The median number of tokens present in titles is 2, while the minimum found is 1 and the maximum is 30 tokens per title. The user names, identification codes, and nicknames were anonymized, but a unique `user_id` was kept to make identification of all reviews written by the same user possible. User attributes also include their gender, birth date and geographical location, making this corpus particularly interesting for social analysis of e-commerce customer behavior.

B2W-Reviews01 is distributed as a comma-separated values file (`.csv`), each line representing one customer review. Each review has with 14 fields, described in Table 1.

[8] `https://www.carrosnaweb.com.br`

[9] `http://www.buscape.com.br`

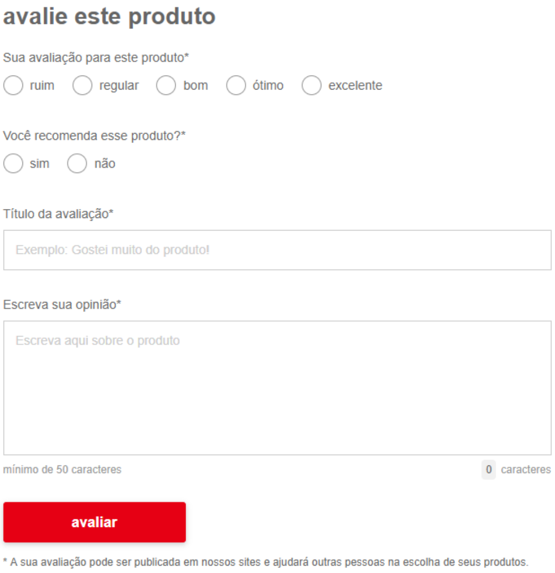
[10] Here, we consider token as the content between two white spaces. We did not pre-process the corpus to correct typos, for example.

#	Field	Data type	Description
1	submission_date	date/time	review submission date (format YYYY-MM-DD hh:mm:ss)
2	reviewer_id	string	unique reviewer id
3	product_id	integer	unique product id
4	product_name	string	product name
5	product_brand	string	product brand
6	site_category_lv1	string	product category - first level
7	site_category_lv2	string	product category - second level
8	overall_rating	integer	overall customer rating, from 1 to 5
9	recommend_to_a_friend	string	answer to “would you recommend this product to a friend?” (“Yes”/“No”)
10	review_title	text	review title, introduces or summarizes the review content
11	review_text	text	main text content of the review
12	reviewer_birth_year	integer	reviewer’s birth year
13	reviewer_gender	string	reviewer’s gender (”F” for female; ”M” for male)
14	reviewer_location	string	reviewer’s Brazilian State, according to the delivery address

Table 1. Fields, data types, and descriptions.

3.1. Reviews collection in *Americanas.com*

After a product is successfully delivered to the customer, the company sends an e-mail with a link to the product review form (Figure 2). This form can also be reached from the product page, so that any customer can write a review anytime, without needing to have bought the product on *Americanas.com*. The review form consists of an overall rating of the product ranging from 1 (bad) to 5 (excellent), a question on whether the customer recommends the product, a review title and the review text. All fields are required and the review text must have at least 50 characters.



avalie este produto

Sua avaliação para este produto*

○ ruim ○ regular ○ bom ○ ótimo ○ excelente

Você recomenda esse produto?*

○ sim ○ não

Título da avaliação*

Exemplo: Gostei muito do produto!

Escreva sua opinião*

Escreva aqui sobre o produto

mínimo de 50 caracteres 0 caracteres

avaliar

* A sua avaliação pode ser publicada em nossos sites e ajudará outras pessoas na escolha de seus produtos.

Figure 2. Product review form.

3.2. Examples and Discussion

One of the main goals of this work is to provide a data collection of opinions left by customers, with which one can get perceptions of evaluations across different user profiles and product features. Since we release the reviews exactly as they were written in the

B2W-Reviews01 corpus, one can find all kinds of noisy user-generated texts: simple typos, abbreviations, internet register and a vast amount of constructions hugely influenced by orality. One can also find offensive language and sarcasm. Here we present a few examples from the corpus.

Example 1.[11] In this example of a review scored as 5, one can see the language register present in many reviews. Even a review that can be considered well written lacks diacritics. In this example, we see a typical case of difficulty when processing the lack of diacritics in Portuguese: we miss the distinction of two of the most used words in Portuguese: ‘e’ (and) and ‘é’ (is). Of course, this characteristic imposes several challenges for one interested in processing the corpus. A task as simple as lemmatizing the data becomes a complex task when several kinds of mistakes, registers and typos appear indistinctly in the corpus.

Field	**Value**
`submission_date`	“2018-01-01 02:02:13”
`reviewer_id`	“a0fd1ad35b08d3b764ad6f884ef7183bf29fc7eb(...)”
`product_id`	122776350
`product_name`	“Ventilador de Teto Ventisol Fenix Premium Branco 3 velocidades com Controle Remoto”
`product_brand`	“ventisol”
`site_category_lv1`	“Casa e Construção”
`site_category_lv2`	“Climatização”
`overall_rating`	5
`recommend_to_a_friend`	“Yes”
`review_title`	“Gostei do produto”
`review_text`	“O barulho e minimo e o vento é bem forte na velocidade 2”
`reviewer_birth_year`	1987
`reviewer_gender`	“M”
`reviewer_location`	“SP”

Example 2.[12] In this example, the field `product_brand` has its value set to `null` as it is not present in the database.

Field	**Value**
`submission_date`	“2018-03-18 06:10:10”
`reviewer_id`	“ecd8648fee87789e041522b6d2e0ee5e22bcacb7(...)”
`product_id`	132326651
`product_name`	“Smart TV LED 48” Sony KDL-48W655D com Conversor Digital 2 HDMI 2 USB Wi-Fi Foto Sharing Plus Miracast Preta”
`product_brand`	null
`site_category_lv1`	“TV e Home Theater”
`site_category_lv2`	“TV”
`overall_rating`	5
`recommend_to_a_friend`	“Yes”
`review_title`	“Gostei muito da minha TV Smart.”
`review_text`	“Vocês estão de parabéns fez uma excelente entrega o produto chegou em perfeito estado gostei muito.obrigado.”
`reviewer_birth_year`	1986
`reviewer_gender`	“F”
`reviewer_location`	“MG”

Example 3.[13] This example shows a typical case of offensive language and sarcasm being used, while the review is still positive. Considering the two rates we have

[11] I liked the product: The noise is minimal and the wind is very strong at speed 2.

[12] I liked very much my Smart TV: Congratulations you made an excellent delivery the product arrived in perfect condition I enjoyed it very much.thank you.

[13] good: it sucks why cannot only give stars need to write something to evaluate whatafuck americanas change it.

related to the user's opinion, the user liked the product. However, the user's text is quite negative, but it is about the process of reviewing offered by *Americanas.com*, and not the product itself. The user pointed out that he preferred to only rate the product with stars and not be forced to write something about it.

Field	**Value**
`submission_date`	"2018-05-05 22:05:54"
`reviewer_id`	"a4297137bb957850899982a232218(...)"
`product_id`	31053501
`product_name`	"Smartphone Multilaser MS80 4G 32GB 5,7 HD 3GB RAM Android 7.1 Dual Camera 20MP+8MP dourado- P9065"
`product_brand`	null
`site_category_lv1`	"Celulares e Smartphones"
`site_category_lv2`	"Smartphone"
`overall_rating`	5
`recommend_to_a_friend`	"Yes"
`review_title`	"bom"
`review_text`	"que merdapq n pode so da estrela tem que escrever alguma coisa para avaliar koé americanas muda isso"
`reviewer_birth_year`	1999
`reviewer_gender`	"M"
`reviewer_location`	"SP"

Considering these three examples, one can have a perception about the challenges this dataset imposes. One can naively think that classifying opinions using user generated text content can be enough to tackle the problem of analyzing users' perceptions of products, but many times the text content does not match the rate given by the user. That is mostly because the review is frequently not about the product itself, but about a specific aspect of the customer purchase journey. In Example 1, the review is about the product, in Example 2, the review is about the delivery and finally, in Example 3, the review is about the process of writing a review. In all these cases, the product is rated as a `5` star product.

4. Polarity Classification Challenges

Polarity classification is a subtask of sentiment analysis often done with corpora such as *B2W-Reviews01*, since such data offers not only the user text content but also a score that is supposed to express the user's sentiment in relation to what s/he is writing about. However, as the examples show, the polarity of the text doesn't always express the reviewer's sentiment: Example 3 shows how the text content can be negative while the customer is seemingly satisfied with the purchased product. Therefore, polarity text classification, which is a task on its own, is not a forward way to get the sentiment of the customer and, if available, other information about the customer and the review can be used together to better understand customer sentiment about the product or her/his experience.

As in the *Amazon Customer Reviews Dataset* the main rate associated with a review in *B2W-Reviews01* is a 5 points scale rate, here called an **overall_rating**. [Rain 2013, Avanço and Nunes 2014] and [Nobre et al. 2016], following many other works looking at corpora reviews also rated in a 5 point scale, argue that 'it is better to only consider reviews rated from 0 to 2 as negative and only those rated 5 as positive. Rates 3 and 4 could be positive, negative or neutral within the same universe (ambiguous) and should be discarded for polarity classification' [Nobre et al. 2016]. Table 2 shows the distribution of polarity reviews in *B2W-Reviews01*, considering the literature.

However, when we consider the field **recommend_to_a_friend**, we see a unexpected, but complementary information: 72.8% of the customers actually recommend the

Negative (1-2)	35,758	27.01%
Neutral (3-4)	48,660	36.76%
Positive (5)	47,955	36.23%

Table 2. Reviews polarity classification following literature

product to a friend, most of them included in the 'neutral' portion of the reviews: 31,837 users that rated the reviews 4 stars and 14,434 users that rated the reviews 3 stars recommended the product to a friend.

This analysis suggests that the typical approach of distributing reviews scored 4 and 3 as neutral can lead one to wrongfully analyze user sentiment. As roughly pointed by [Liu 2012, Chap.03], the simplification of these scores is often related to the computational feasibility/precision of a given model. This is to say that this simplified analysis is often used because of technical issues and not based on what the data really shows. It is easier for an automated system to be correct when scoring a review among only three very different scores — and not among five slightly different classifications. Of course, future analyses of this case need to be carried out to really confirm that the cited approach of analyzing scored reviews is not the best. But it suggests that, for real case applications, this simplified analysis is not appropriate and that performing analyses based on what a model can do, and not based on data driven insights can lead to wrong perceptions of what your data actually says.

5. Conclusions and Future work

This work introduced the *B2W-Reviews01* corpus, a dataset composed by products reviews and the metadata related to them submitted to the *Americanas.com* marketplace between January and May, 2018. The main goal of this work is to describe an open and freely available dataset of product reviews that can be useful for further works interested on different NLP/CL tasks that can use such data.

We leave several tasks as future work. In particular, we are interested on topic modeling in reviews. Also seen as a feature extraction task, knowing the topic of the review — what the review is actually talking about: the delivery process, the whole user experience when buying in the marketplace, the product itself or a specific feature of it, for example — is a critical task for *B2W Digital*, since mining the opinions of users only becomes relevant when one can figure out what the opinion is about. Another experiment we leave as future work is the analysis of polarity reviews started in section 4. While computational models for classification work better when we have few and very distinct categories, if the final result of the carried out analysis does not represents a trustworthy conclusion that can be effectively applied to a business model, it would be better to invest more on augmenting the precision of said models before applying the output results to the business case we are interested in.

We also plan to have a smaller part of the data annotated for sarcasm, following [Filatova 2012], since detecting sarcasm in Portuguese is still an open task and *B2W-Reviews01* offers a rich content for it. Moreover, *B2W Digital* plans to periodically release open reviews corpora, considering the relevance of such data and the difficulty in having established open data in Portuguese.

References

Avanço, L. V. (2015). Sobre normalização e classificação de polaridade de textos opinativos na web. Master's thesis, ICMC/USP.

Avanço, L. V. and Nunes, M. G. V. (2014). Lexicon-based sentiment analysis for reviews of products in brazilian portuguese. *3th Brazilian Conference on Intelligent Systems*.

Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. *LREC*.

Hartmann, N. S., Avanço, L. V., Balage, P. P., Duran, M. S., Nunes, M. G., Pardo, T. A., and Aluísio, S. M. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. *LREC*.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Nobre, G., Justino, A., Tadao, F., Nunes, D., Takabayashi, D., and Küllian, R. (2016). Booviews: Aspect-based sentiment analysis on product reviews combining svm and crf in portuguese. *Student Research Workshop - PROPOR*.

Rain, C. (2013). Sentiment analysis in amazon reviews using probabilistic machine learning. Master's thesis, Swarthmore College.

Ribeiro, S. S., Junior, Z., Meira, W., and Pappa, G. L. (2012). Positive or negative? using blogs to assess vehicles features. *ENIA*.

Siqueira, H. and Barros, F. M. M. (2010). A feature extraction process for sentiment analysis of opinions on services. In *WTI 2010*.

Wachsmuth, H., Trenkmann, M., Stein, B., Engels, G., and Palakarska, T. (2014). A review corpus for argumentation analysis. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 115–127, Berlin, Heidelberg. Springer Berlin Heidelberg.