Creating an Annotated Corpus for Sentiment Analysis of German Product Reviews

Boland, Katarina; Wira-Alam, Andias; Messerschmidt, Reinhard

gesis
Leibniz-Institut
für Sozialwissenschaften

TECHNICAL *Reports* 2013|05

Creating an Annotated Corpus for Sentiment Analysis of German Product Reviews

*Katarina Boland, Andias Wira-Alam und
Reinhard Messerschmidt*

GESIS-Technical Reports 2013|05

Creating an Annotated Corpus for Sentiment Analysis of German Product Reviews

Katarina Boland, Andias Wira-Alam und Reinhard Messerschmidt

GESIS – Leibniz-Institut für Sozialwissenschaften 2013

GESIS-Technical Reports

GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln
50667 Köln
Telefon: (0221) 476 94 - 0
Telefax: (0221) 476 94 - 199
E-Mail: katarina.boland@gesis.org

ISSN: 1868-9051 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

Abstract

The availability of annotated data is an important prerequisite for the development of machine learning algorithms for sentiment analysis. However, as manually labeling large datasets is time-consuming and expensive, few datasets are available and most of them represent a small sample of a very narrow domain, e.g. movie reviews or reviews of a certain product type. Additionally, many annotated datasets are available for English texts only. However, the influence of different characteristics of the input dataset on the performance of algorithms for sentiment analysis remains unclear if only training data from one specific domain is available or if specific domains are mixed in the test corpus. We therefore introduce a new dataset for German product reviews of various product types and investigate whether even small variances in this specific domain (different product types) already exhibit different characteristics, e.g. with regard to the difficulty of sentiment annotation. The annotation of this corpus lays the basis for future enhanced annotations of similar corpora and for the extension of our annotations to corpora of inherently different domains. These will then serve to investigate the influence of different corpus characteristics on different algorithms for sentiment analysis and as a basis to apply machine learning methods for sentence-wise sentiment analysis for German texts.

1 Introduction

Sentiment Analysis, also referred to as Opinion Mining, refers to the task of detecting sentiments or opinions in texts. In this context, several different definitions of sentiments and opinions can be found in the literature (see Pang&Lee 2008 for a detailed overview). In this report, we are engaged in the classification of sentences with regard to the opinions they express towards a reference object of any kind (negative, positive and mixed sentiments or neutral sentences).

Availability of labeled data facilitates large-scale evaluation of sentiment analysis approaches and enables the development of supervised learning algorithms. Sentiment analysis is domain-, topic-, and temporally-dependent (see Pang&Lee 2008). Therefore, the characteristics of the labeled data used for training and evaluation must be explicit in order to fully interpret the performance of sentiment analysis systems. Dave et al. 2003 already expressed the idea that classifiers trained on reviews of one product type might not perform as well when applied on reviews of other product types. However, they did not investigate these effects. To allow the analysis of domain-effects on sentiment analysis of product reviews, we aim to provide a well-documented product review corpus consisting of subcorpora for different product types.

In order to provide a corpus suitable for training and evaluation of sentence-based sentiment analysis algorithms, we focus on strict sentence-wise classification and instruct our annotators not to take the context into account. All sentences are labeled as belonging to one of 4 categories with respect to the sentiments they express: positive, negative, positive and negative (mixed) or neutral. Furthermore, all sentences are categorized by two more labels: Ideally, we would like to have information on the target of a sentiment. However, whether it is the product reviewed itself or another object cannot be determined reliably without knowing the context. Therefore, we cannot expect our annotators to give information about the target of a sentiment in our strictly context-free classification task. Instead, we only ask them to indicate when multiple objects are judged in one sentence. This way, the resulting annotations can be used to train and evaluate sentence-based sentiment analysis algorithms but they are not labeled specifically with regard to sentiments expressed towards the respective reviewed products. As a last category, we ask the annotators to indicate whether background knowledge is required to determine the sentiment expressed. This often is the case when technical details about products are listed which can only be comprehended by domain experts.

2 Characteristics of Product Reviews

2.1 Characteristics of Different Product Types

Reviews for different products exhibit different characteristics. For example, reviews for cameras often include many facts about technical specifications and are written in a rather unemotional style while book reviews tend to include sarcasm, irony and other less prosaic stylistic means. Providing annotated reviews for different kinds of products allows the systematic comparison of sentiment expressions in different domains both qualitatively and quantitatively (i.e. with regard to the difficulty of sentiment detection for each different domain). As different product types, we chose tablets, books, washing machines, cameras, mobiles and smartphones.

2.2 Positive vs. Negative Reviews

With each negative review, the probability that a product will be bought by a new costumer decreases. Thus, products with a high proportion of negative reviews tend to have lower total numbers of reviews than products with a more balanced or positive assessment (this observation holds true especially for products whose qualities can at least in part be measured on an objective scale such as for cameras. For products whose assessment depends primarily on taste (e.g. for books), more diverse reviews can be found.) Therefore, Amazon product reviews contain in many cases more positive than negative sentences.

However, strongly biased corpora may pose problems for machine learning approaches. In order to obtain a corpus that is roughly balanced with regard to the number of positive and negative sentences, we used information on the ratings of products: We assumed that reviews assigning a high score (4-5 stars) contain more positive than negative sentences while low rating reviews (1-2 stars) contain mostly negative sentences. For ambivalent ratings (3 stars), we assumed that both positive and negative statements would be contained. Therefore, we balanced the number of sentences contained in low and high rating reviews and added all sentences contained in ambivalent reviews. For camera reviews, however, only a small number of negative reviews was available. Since our first analyses of annotated data furthermore revealed that our corpus contained more negative than positive sentences, we decided to select more sentences from positive than from negative reviews for the camera sub-corpus.

3 Experimental Setup

3.1 Extraction of Amazon Product Reviews

In order to extract Amazon reviews, we used the Amazon reviews parser provided by Andrea Esuli[1]. We had to make some language specific modification, as the original tool only supported English reviews. For each product category, the modified tool was run for several days, obtaining thousands of reviews. For the annotation of the dataset, we split the reviews into sentences. This was done by the ASV Segmentizer, provided by the University of Leipzig[2].

The total number of extracted sentences (after balanced sampling) from each domain is as follows:

Table 1: Resulting test set after balanced sampling

Domain	#Sentences	1 star	2 stars	3 stars	4 stars	5 stars
Washing Machine	4337	1318	388	925	352	1354
Camera	3531*	300	504	1180 *	385	1162
Book	14191	3265	2261	3139	1557	3969
Tablet	9097	1454	1919	2351	1024	2349
Mobile	4685	1067	637	1277	504	1200
Smartphone	27226	5523	4460	7260	3146	6837
Total	63067 *	12927	10169	16132	6968	16871

* Due to an error in the export script, 21 test sentences of the camera reviews were exported as empty strings. They are counted as ambivalent 3-star reviews in this table; annotators rated them as “neutral”.

3.2 Annotation Process and Guidelines

3.2.1 Annotators

In total, nine annotators worked on the corpus for three weeks. Five of them are female and four of them male, all aged between 25 and 40 years. All annotators are German native-speakers and share a comparable educational background (university students or graduates). We split the corpus into chunks of 500 sentences. 21 of them were labeled by three or more annotators consisting of 4 chunks for books, 3 chunks for mobiles, 8 chunks for washing machines and 2 chunks for tablets, cameras and smartphones respectively. Of those, 2 chunks were labeled by four annotators (1 chunk for cameras and 1 chunk for washing machines). Additional 2 chunks each for books and mobiles were annotated by two annotators. This allows the measurement of inter-rater-agreement values for all different product types. Also, for the 10500 sentences with more than two annotations, a majority voting of annotations can be performed which increases the data quality. Furthermore, when using the 15500 sentences with more than one annotations, controversial sentences may be filtered out. We distributed the rest of the data only to single annotators in order to gain as many annotated sentences as possi-

[1] http://www.esuli.it/software/amazon-reviews-downloader-and-parser/

[2] http://asv.informatik.uni-leipzig.de/

ble. Consequently, we achieved a total number of 63067 sentences annotated by at least one annotator. The distribution of sentences to annotators is summarized in table 2.

Table 2: Distribution of sentences per domain to annotators

Domain / Blocks	1st Annotator	2nd Annotator	3rd Annotator	4th Annotator	5th Annotator	6th Annotator	7th Annotator	8th Annotator
wmachine 1-1000	–	–	–	–	–	√	√	√
wmachine 1001-2500	–	–	–	√	√	–	–	√
wmachine 2501-4000	–	–	√	–	√	–	–	√
wmachine 4001-4337	–	–	√	–	–	–	–	–
camera 1-1000	–	–	–	–	–	√	√	√
camera 1001-2500	–	–	–	√	–	–	–	–
camera 2501-3531	–	–	√	–	–	–	–	–
book 1-1000	–	–	–	–	–	√	√	√
book 1001-2500	–	–	–	–	√	–	√	–
book 2501-3500	–	–	√	–	√	–	√	–
book 3501-4500	–	–	–	–	√	√	√	–
book 4501-8000	–	√	–	–	–	–	–	–
book 8001-11500	√	–	–	–	–	–	–	–
book 11501-12000	–	–	–	√	–	–	–	–
book 12001-12500	–	–	–	–	–	√	–	–
book 12501-13000	–	–	–	–	–	–	√	–
book 13001-13500	–	–	–	–	–	–	–	√
book 13501-14191	–	–	√	–	–	–	–	–
smartphone 1-1000	–	–	–	–	–	√	√	√
smartphone 1001-4000	√	–	–	–	–	–	–	–
smartphone 4001-7000	–	√	–	–	–	–	–	–
smartphone 7001-10000	–	–	–	√	–	–	–	–
smartphone 10001-13000	–	–	–	–	√	–	–	–
smartphone 13001-16500	–	–	–	–	–	√	–	–
smartphone 16501-20000	–	–	–	–	–	–	√	–
smartphone 20001-23000	–	–	–	–	–	–	–	√
smartphone 23001-27226	–	–	√	–	–	–	–	–
tablet 1-1000	–	–	–	–	–	√	√	√

Domain / Blocks	1st Annotator	2nd Annotator	3rd Annotator	4th Annotator	5th Annotator	6th Annotator	7th Annotator	8th Annotator
tablet 1001-2000	√	-	-	-	-	-	-	-
tablet 2001-3000	-	√	-	-	-	-	-	-
tablet 3001-4000	-	-	-	√	-	-	-	-
tablet 4001-8000	-	-	-	-	√	-	-	-
tablet 8001-9000	-	-	-	-	-	√	-	-
tablet 9001-9097	-	-	√	-	-	-	-	-
mobile 1-1000	-	-	-	-	-	√	√	√
mobile 1001-1500	-	-	-	-	-	√	-	-
mobile 1501-2000	-	-	-	-	-	-	-	√
mobile 2001-4685	-	-	√	-	-	-	-	-

3.2.2 Guidelines

The guidelines for our sentiment annotation task were designed to be as specific as possible with regard to the definitions of the categories to be assigned. At the same time, we tried to be as brief as possible to minimize the cognitive load for our annotators. For each category, we provided example sentences and annotations. All of these were drawn from the domain of book reviews. We expected our annotators to translate these examples to other domains in order to avoid extensive guidelines consisting of multiple example sentences from each domain for every category. We particularly highlighted usages of categories that may appear counter-intuitive to some annotators. For example, consider the following sentence: “Und es ist spannend geschrieben, bis vielleicht ca. 4/5 des Buchs.” (“And it is written in an exciting style, until approx. ⅘ of the book”). This statement implicitly expresses a negative opinion on the remainder of the book. However, explicitly, a positive opinion on the first chapters of the book is expressed. This sentence can therefore be interpreted as being positive (explicit, literal meaning), negative (implicit criticism is initiated) or mixed (the book has both positive and negative aspects). In the absence of guidelines, different annotators would probably arrive at different conclusions depending on their preferred interpretation. Therefore, we explicitly defined our desired categorization in such cases:

Annotators are instructed to annotate the explicit meaning of a sentence if an elaboration of the implicitly indicated opposite aspect can be assumed to follow. In this example, the desired annotation would therefore be “positive”. Although the sentence implies a negative sentiment, it indicates that the negative aspects will be elaborated subsequently. As this elaboration would have to be labeled “negative”, the negative aspect would be weighted double if the annotations for both sentences were to incorporate the negative sentiment.

If both positive and negative sentiments are expressed in a sentence, they must not be weighted up. Consider this sentence: "Es ist nicht von der Hand zu weisen, dass die negativen Eindrücke überwiegen und doch habe ich es gerne gelesen.” (“Regardless of the fact that the negative impressions prevailed, I enjoyed reading it.”). This sentence could be interpreted as negative if the sentiments in the sentence were to be weighted up. However, we instructed the annotators to categorize sentences as “mixed” once both positive and negative sentiments are present regardless of their respective strengths. To emphasize this usage of the “mixed” category, we named the category “positiv und negativ” (“positive and negative") instead of "ambivalent", "mixed" or other labels that might suggest that only sentences with positive and negative sentiments of equal strengths should be assigned to this category.

Sentences that do not bear an opinion or subjective judgment are to be labeled "neutral". For example, the sentence "Habe darüber auch mit einem Autor gesprochen." ("I also talked about this with an author.") should be labeled as neutral. Also, the sentence "Es wurde ein einfacher Schreibstil gewählt, so dass die ca. 200 Seiten relativ schnell gelesen waren." ("The writing style was simple, therefore the approx. 200 pages could be read quickly.") belongs to the neutral category. This statement may be regarded as being positive or negative as a matter of taste while the sentence itself does not express any subjective evaluation but only an objective description.

The label "background knowledge required" is meant for sentences consisting of statements that contain evaluative components that may only be recognized with domain knowledge. For example, if the sentence "Die Robinie wird im Buch als essbar beschrieben." ("The robinia is described as edible in the book") is found in a book review, it should be annotated as "negative" because the robinia is a poisonous plant. However, knowledge about this plant is required to recognize this implicit sentiment. Therefore, the sentence should be labeled as "background knowledge required". If the annotator does not know how to categorize the sentiment because of lacking background knowledge, he or she should label the sentence as "neutral".

The category "multiple objects are judged" is to be applied to sentences expressing opinions about more than one object. For example, in the sentence "Die Klinischen Lexika von Roche und Springer bieten einfach mehr und sind deutlich ansprechender" ("The clinical lexica by Roche and Springer simply offer more and are considerably more appealing.") two different books, one lexicon by Roche and one by Springer, are evaluated. Thus, the label "multiple objects are judged" is to be assigned.

3.2.3 Annotation Procedure

For our annotation task, we used LimeSurvey 2.0[3], a free and open source survey tool. After having generated samples balanced with regard to the ratings assigned to the reviews (see section 2.2), we exported the sentences into the LimeSurvey format and loaded them into the tool. For each sentence, the assignment of exactly one sentiment category is mandatory. We employed radio buttons to model this condition. At the same time, the two additional labels ("background knowledge required" and "multiple objects are judged") both may or may not be assigned. For this, we employed checkboxes. Since we combined single choice answers (radio buttons) and multiple choice answers (checkboxes) and these are not the standard answer types in LimeSurvey, we wrote a simple Javascript template to fulfill this requirement. Radio buttons were presented in horizontal alignment in the order "negative", "positive and negative", "positive" and "neutral". Below, the optional checkboxes were arranged in vertical order to emphasize their disjointedness and independence. Each block of 500 sentences was presented as one survey to permit the annotators to work on the corpus in convenient chunks. To ensure that annotators are not influenced by the context, we provided all sentences of all reviews for a product type in random order.

[3] http://www.limesurvey.org/

4 Statistics for the Resulting Sentiment Annotated Corpus

Table 3 shows the distribution of positive, negative, mixed, neutral and controversial sentences. For sentences annotated by multiple annotators, the majority vote of annotations was used as the final annotation. If no majority vote could be determined (i.e. if all voters assigned a different category or if there was a draw for annotations done by two or four voters), the sentence is counted as "controversial". The balancing of sentences with regard to their ratings succeeded in creating a roughly balanced sample of positive and negative sentences. However, for most product types, negative sentences prevail. This indicates that the number of negative sentences in positive and ambivalent reviews is higher than the number of positive sentences in negative and ambivalent reviews.

We calculated the inter-rater agreement for all different product types using chunks with three annotations using Fleiss' kappa (Fleiss 1971). For 500 sentences each of camera and washing machine reviews, four annotations were available. In order to compute a single agreement score for both product types using all 10500 sentences labeled by three or more annotators, we ignored the annotations of the fourth annotator for calculation of the aggregated agreement score. The scores are presented in table 4. Agreement scores for all individual chunks with at least two annotators can be found in table 5.

Table 3: Distribution of sentiment categories

Subcorpus	#positive sentences	#negative sentences	#mixed sentences	#neutral sentences	#controversial sentences
Washing Machine	1272	1455	314	1190	106
Camera	3765	4525	915	4425	561
Book	1085	903	196	1278	69
Tablet	2594	3337	827	2318	21
Mobile	1296	1282	334	994	779
Smartphone	7841	10379	2193	6793	20
Total	17853	21881	4779	16998	1556

Table 4: Aggregated agreement values for sentiment and additional annotations

Subcorpus	Fleiss' kappa (sentiment)	Fleiss' kappa (background-knowledge)	Fleiss' kappa (multiple-objects)
tablet (1000 sentences)	0.669023568608	0.0759408602151	0.497546752078
book (2000 sentences)	0.639490540088	0.0338983050847	0.239569063004
camera (1000 sentences)	0.618826040675	0.178621092146	0.495464327269
washing machine (4000 sentences)	0.731039456711	0.113085401537	0.374045995322
mobile (1500 sentences)	0.737001237941	0.0199372759856	0.576914098973
smartphone (1000 sentences)	0.701363957328	0.0574731903485	0.44204811262
total (10500 sentences, 3 annotators (4th annotations ignored))	0.696574300709	0.126468059085	0.430911237181
total (9500 sentences, 3 annotators)	0.69986560963	0.118139214863	0.43041180694
total (1000 sentences, 4 annotators)	0.676820023687	0.122821050162	0.441077163421
total (5000 sentences, 2 annotators)	0.640039275526	0.00995075647378	0.206423084315

Table 5: Agreement values for sentiment and additional annotations of individual chunks

Sentences	Number of annotators	Fleiss' kappa (sentiment)	Fleiss' kappa (background-knowledge)	Fleiss' kappa (multiple-objects)
books, 1-500	3	0.650506351999	0.0626802537973	0.250806823421
books, 501-1000	3	0.674712950124	-0.00603621730383	0.226423497972
books, 1001-1500	2	0.683692629751	-0.0384215991693	0.341919714205
books, 1501-2000	2	0.749643452895	0.020285846012	0.427083333333
books, 2001-2500	2	0.684983447312	0.00152149106124	0.418904958678
books, 2501-3000	2	0.791267654102	0.0879339020334	0.599483779093
books, 3001-3500	2	0.632081364208	-0.039501039501	0.0654621512171
books, 3501-4000	3	0.61344694336	0.0638712823013	0.0825580164434
books, 4001-4500	3	0.616044640208	-0.00431118949639	0.293090909091
cameras, 1-500	3 (ignoring last annotation)	0.609913601656	0.242407606667	0.511875341923
cameras, 1-500	4	0.641056013209	0.0969943997084	0.514583597147
cameras, 501-1000	3	0.624198413364	0.0523921832884	0.479858983991
mobiles, 1-500	3	0.744914620336	-0.0046885465506	0.46205975465
mobiles, 501-1000	3	0.71868071814	-0.00267379679143	0.604628818733
mobiles, 1501-2000	3	0.745888411759	0.0237288135593	0.630376650756
mobiles, 2001-2500	2	0.495446260775	-0.00502512562815	0.104390811282

Sentences	Number of annotators	Fleiss' kappa (sentiment)	Fleiss' kappa (back-ground-knowledge)	Fleiss' kappa (multiple-objects)
mobiles, 2501-3000	2	0.676974211775	-0.00502512562815	0.172261565345
mobiles, 3001-3500	2	0.552466990966	-0.00502512562815	0.143747645306
mobiles, 3501-4000	2	0.5546867123	-0.00603621730383	0.0793931359552
mobiles, 4001-4500	2	0.572609656757	-0.00603621730383	0.152711323764
smartphones, 1-500	3	0.733021077283	-0.00536193029495	0.487174748669
smartphones, 501-1000	3	0.669620646766	0.120308310992	0.393843818685
tablets, 1-500	3	0.688260462546	0.075940860215	0.483859127951
tablets, 501-1000	3	0.648224183564	0.0759408602151	0.513334631108
washing machines, 1-500	3 (ignoring last annotation)	0.701543884602	0.106863970014	0.251730373511
washing machines, 1-500	4	0.711845101689	0.122134967758	0.283241371849
washing machines, 501-1000	3	0.681631505069	0.145762711864	0.2489749523
washing machines, 1001-1500	3	0.789193000959	0.0330193833177	0.594707262846
washing machines, 1501-2000	3	0.74118087254	0.0677131746704	0.357876712329
washing machines, 2001-2500	3	0.745809109159	0.0584527013602	0.601960539461
washing machines, 2501-3000	3	0.742056520352	0.180286521388	0.22520661157
washing machines, 3001-3500	3	0.724287504214	0.207733631707	0.309868875086
washing machines, 3501-4000	3	0.697703312916	0.0940547341931	0.242407606667

While the kappa values for the sentiment annotations show a good agreement, agreement scores for the additional categories generally are poor. For the first additional category, this may be due to the fact that the judgment if background knowledge is required itself may require a certain level of background knowledge and might therefore be hard to determine. Furthermore, there may have been confusion about the label itself. We asked annotators to assign this label if background knowledge is required to recognize the sentiment of a sentence. However, they might have generalized the condition and assigned the label whenever background knowledge was required to make any kind of judgment, including the other additional category.

For the second additional category, "multiple objects are judged", annotators might have found it hard to determine what the concept "object" should include. For example, sentences may include sentiments about abstract categories rather than concrete objects. The guidelines make no specific statement about what should be seen as an object which might have led to different interpretations by the annotators.

5 Conclusions

We introduced a corpus of German Amazon product reviews containing sub-corpora for six different product types: books, cameras, tablets, washing machines, mobiles and smartphones. The provision of corpora for multiple different product types allows the comparison of domain-specific characteristics and the evaluation of algorithms on reviews of different domains. The corpora are roughly balanced with regard to the number of positive and negative sentences contained and feature good inter-rater agreement scores for sentiment annotation. This makes the corpus suitable for training and evaluation of sentiment analysis approaches. Low scores for additional categories for sentences that require background knowledge for determining the sentiment and for sentences that express opinions on multiple objects reveal the need for refined guidelines to examine these properties.

6 References

Pang & Lee 2008:

Bo Pang, Lillian Lee (2008):
“Opinion Mining and Sentiment Analysis.”
In: Foundations and Trends in Information Retrieval, Vol. 2, No. 1-2, pages 1-135.

Dave et al. 2003:

Kushal Dave, Steve Lawrence, David M. Pennock (2003):
“Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews.”
In: Proceedings of WWW, pages 519-528.

Fleiss 1971:

Joseph L. Fleiss (1971): "Measuring nominal scale agreement among many raters."
In: Psychological Bulletin, Vol. 76, No. 5, pages 378–382