A Review Corpus for Argumentation Analysis

Henning Wachsmuth[1], Martin Trenkmann[2], Benno Stein[2],
Gregor Engels[1], Tsvetomira Palarkarska[2]

[1] Universität Paderborn, s-lab – Software Quality Lab, Paderborn, Germany
{hwachsmuth,engels}@s-lab.upb.de
[2] Bauhaus-Universität Weimar, Weimar, Germany
{benno.stein,martin.trenkmann,tsvetomira.palakarska}@uni-weimar.de

Abstract. The analysis of user reviews has become critical in research and industry, as user reviews increasingly impact the reputation of products and services. Many review texts comprise an involved argumentation with facts and opinions on different product features or aspects. Therefore, classifying sentiment polarity does not suffice to capture a review's impact. We claim that an argumentation analysis is needed, including opinion summarization, sentiment score prediction, and others. Since existing language resources to drive such research are missing, we have designed the *ArguAna TripAdvisor corpus*, which compiles 2,100 manually annotated hotel reviews balanced with respect to the reviews' sentiment scores. Each review text is segmented into facts, positive, and negative opinions, while all hotel aspects and amenities are marked. In this paper, we present the design and a first study of the corpus. We reveal patterns of local sentiment that correlate with sentiment scores, thereby defining a promising starting point for an effective argumentation analysis.

1 Introduction

Argumentation is a key aspect of human communication and cognition, consisting in a regulated sequence of speech or text with the goal of providing persuasive arguments for an intended conclusion or decision. It involves the identification of relevant facts about the topic or situation being discussed as well as the structured presentation of pros and cons [3]. In terms of text, one of the most obvious forms of argumentation can be found in reviews. Reviews provide facts and opinions about a product, service, or the like in order to justify a particular overall rating or sentiment, as in the following example: *"This was truly a lovely hotel to stay in. The staff were all friendly and very helpful. The location was excellent. The atmosphere is great and the decor is beautiful."*

In the last decade, the vast amount of user reviews in the web has become a primary influence factor of the reputation of products and services. As a consequence, research and industry put much effort into approaches and resources for the automatic analysis of reviews. Most approaches classify sentiment polarity at the text-level [12]. However, the facts, pros, and cons in review texts have proven beneficial for more complex tasks, such as summarizing opinions on different product features [7], interpreting local sentiment flows [9], or predicting sentiment scores [19]. Still, there has been no publicly available linguistic resource until now that makes it possible to jointly analyze the different types of information involved in the argumentation of reviews (cf. Section 2 for details).

In this paper, we present our design of the annotated *ArguAna TripAdvisor corpus* for analyzing the argumentations of web user reviews. The corpus consists of 2,100 English hotel reviews from an existing *TripAdvisor* dataset [17, 19], evenly distributed across seven hotel locations. Such a review comprises a text, a set of ratings, and some metadata. In each text, we let experts manually annotate all hotel aspects and amenities as product features. In addition, we segmented the texts into subsentence-level statements. Then, we used crowdsourcing to classify every statement as a fact, a positive, or a negative opinion. In total, the corpus comprises 24.5k product features and 31k statements, while it is balanced with respect to the reviews' overall ratings, i.e., sentiment scores from 1 to 5.

The corpus is freely available at http://www.arguana.com for scientific use. It serves as a linguistic resource for the development and evaluation of approaches to sentiment-related tasks. Some example tasks have been named above [7, 9, 19], but the corpus also enables research on novel tasks. For large-scale evaluations and semi-supervised learning [13], nearly 200k further reviews from [19] are given without manual annotations. In general, we think that an *argumentation analysis* of texts will provide new insights into the use of language and can improve effectiveness in several natural language processing tasks.

To show the benefit of our corpus, here we investigate how the argumentation of a review text relates to the review's global sentiment. We offer evidence for the importance of the distribution of local sentiment in a review text, both in general and regarding specific product features. Moreover, we reveal common patterns of changes in the flow of local sentiment and their correlations with global sentiment scores. Altogether, our main contributions are the following:

1. We present the design of a freely available text corpus for analyzing the argumentation of web user reviews in terms of sentiment (Section 3).
2. We analyze the corpus to obtain new findings on correlations between a hotel review's sentiment score and the local sentiment in the review's text, giving insights into the ways web users argue in reviews (Section 4).

2 Related Work

In his pioneer study of arguments, Toulmin [16] models the basic argumentation structure with facts and warrants justified by a backing, leading to a qualified claim unless a rebuttal counters the facts. An approach to infer similar structures from scientific articles is given by *argumentative zoning* [15]. Recently, research has started to generally address *argumentation mining*, which analyzes natural language texts to detect the different types of arguments that justify a claim as well as their interactions [10]. In the reviews we consider, however, the actual claim is often not explicit, but it is quantified in terms of a sentiment score.

The argumentation of reviews is related to the concept of discourse, but it differs from *conversational* discourse, where the participants present arguments to persuade each other [4]. A review comprises a *monological* and *positional* argumentation, where a single presenter collates and structures a choice of facts and opinions in order to inform the intended recipient about his or her beliefs [3]. Accordingly, the aim of our corpus is not to check whether a claim is well argued, but to analyze what information is chosen and how arguments are structured to justify the claim, assuming the claim holds.

Following [10], an *argumentation analysis* enables a better understanding of discourse, intentions, and beliefs. This helps analyzing the sentiment of reviews, which in turn benefits the reputation management of products and services [12]. Different recent approaches exploit discourse structure on the subsentence-level to improve sentiment polarity classification, e.g. [11, 21]. Others extract and summarize opinions [7] or they infer scores for several aspects from reviews [19]. All these approaches capture review argumentation to some extent. However, while sentiment corpora exist for several tasks and domains (cf. [12] for a selection), to our knowledge our corpus is the first that enables a combination of the approaches. The *MPQA corpus* [20] contains phrase-level annotations of opinions and other private states, but it is not meant for analyzing argumentations.

Below, we analyze review texts with respect to the flow of local sentiment. Our work resembles [9] where a sequential model first classifies the sentiment of each sentence in a text. The resulting flow is then used to predict the global sentiment of the text. In contrast, we focus on the identification of abstract argumentation patterns and we provide a corpus for related research.

3 Design of a Corpus for Argumentation Analysis

We now present our main design decisions in the compilation, annotation, and formatting of the *ArguAna TripAdvisor corpus* for the argumentation analysis of web user reviews. The corpus serves the scientific development and evaluation of approaches to tasks like sentiment score prediction [19] and opinion summarization [7]. It can be freely accessed at http://www.arguana.com.

3.1 Balanced Sampling of Web User Reviews

The ArguAna TripAdvisor corpus is based on a carefully chosen subset of a dataset originally used for aspect-level rating prediction [19]. The original dataset contains nearly 250k crawled English hotel reviews from *TripAdvisor* [17] that cover 1,850 hotels from over 60 locations. Each review comprises a text and a set of numerical ratings. The text quality is not perfect in all cases, certainly due to crawling errors: Some line breaks have been lost, which hides a number of sentence boundaries and, sporadically, word boundaries. In our experience, however, such problems are typical for web contexts. We rely on this dataset because its size, the quite diverse hotel domain, and the restriction to English serve as a suitable starting point for analyzing argumentations. We computed the distributions of locations and sentiment scores in the dataset, as shown in Figure 1. The latter should be representative for TripAdvisor in general.

Our sampled subset consists of 2,100 texts balanced with respect to both location and sentiment score. In particular, we selected 300 texts of seven of the 15 most-represented locations in the original dataset, 60 for each sentiment score between 1 (worst) and 5 (best). This supports an optimal training for learning

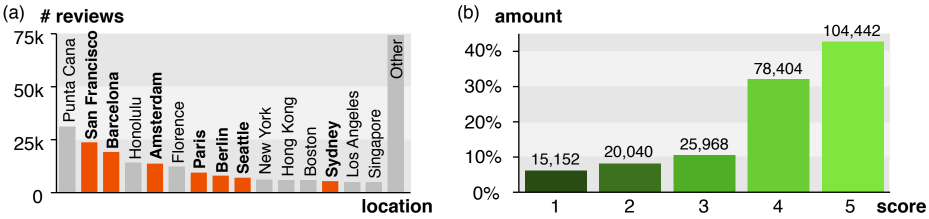


Fig. 1. (a) Distribution of the locations of the reviewed hotels in the original dataset from [19]. The ArguAna TripAdvisor corpus contains 300 annotated texts of each of the seven marked locations. (b) Distribution of sentiment scores in the original dataset.

Table 1. The number of reviewed hotels of each location in the ArguAna TripAdvisor corpus as well as the number of texts for each sentiment score from 1 to 5 and in total.

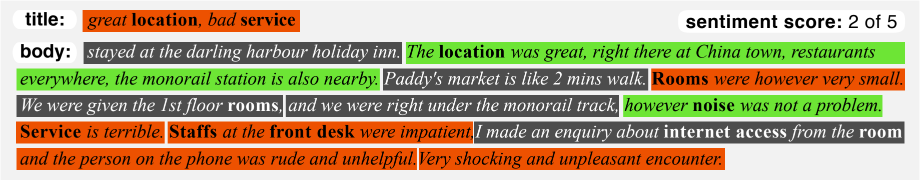
Set	Location	Hotels	1	2	3	4	5	Σ
training	Amsterdam	10	60	60	60	60	60	300
	Seattle	10	60	60	60	60	60	300
	Sydney	10	60	60	60	60	60	300
validation	Berlin	44	60	60	60	60	60	300
	San Francisco	10	60	60	60	60	60	300
test	Barcelona	10	60	60	60	60	60	300
	Paris	26	60	60	60	60	60	300
complete	**all seven**	**120**	**420**	**420**	**420**	**420**	**420**	**2100**

approaches to sentiment score prediction. For opinion summarization, we ensured that the reviews of each location cover at least 10 but as few as possible hotels. To counter location-specific bias, we propose a corpus split with a training set containing the reviews of three locations, and both a validation set and a test set with two of the other locations. Table 1 lists details about the compilation.

3.2 Tailored Annotation Scheme for Argumentations

The reviews in the original dataset from [19] include optional ratings for seven aspects of hotels, namely, *value*, *room*, *location*, *cleanliness*, *front desk*, *service*, and *business service*, as well as a mandatory overall rating. We interpret the latter as the review's *sentiment score*. Besides, there is metadata about each review text (the username of the *author* and the creation *date*) and the reviewed hotel (*ID* and *location*). We maintain this data as text-level annotations in our corpus. In addition, we have enriched the corpus with annotations of local sentiment and product features to allow for an analysis of review argumentation.

Researchers have observed that reviews often contain *local sentiment* on the subsentence-level [21]. A common approach to handle this level is to divide a text into discourse units according to the *rhetorical structure theory* [8]. However, parsing discourse tends to be error-prone on noisy text [11] while being computationally expensive, which can be critical in web contexts. Also, not all discourse units are meaningful on their own, as in the following example, where the first unit depends on the context of the second one:



title: *great* **location**, *bad* **service** sentiment score: 2 of 5

body: *stayed at the darling harbour holiday inn. The* **location** *was great, right there at China town, restaurants everywhere, the monorail station is also nearby. Paddy's market is like 2 mins walk.* **Rooms** *were however very small. We were given the 1st floor* **rooms**, *and we were right under the monorail track, however* **noise** *was not a problem.* **Service** *is terrible.* **Staffs** *at the* **front desk** *were impatient. I made an enquiry about* **internet access** *from the* **room** *and the person on the phone was rude and unhelpful. Very shocking and unpleasant encounter.*

Fig. 2. Illustration of a text from the ArguAna TripAdvisor corpus. Each text is segmented into *positive opinions* (light green background), *negative opinions* (medium red), and objective *facts* (dark gray). All annotated aspects and amenities are marked in bold.

Statement 1. [*Although we had the suite,*]$_{\text{unit1}}$ [*our room was small,*]$_{\text{unit2}}$
Statement 2. [*but everything in the room was great.*]$_{\text{unit3}}$

Therefore, we have segmented each text into single *statements* instead, where we define a statement to be at least a clause and at most a sentence that is meaningful on its own. We assume each statement to have only one sentiment, even though this might be wrong in some cases. For reproducibility, the segmentation was done automatically using a rule-based algorithm provided with the corpus. The algorithm relies on lexical and syntactic clues derived from tokens, sentences, and part-of-speech tags. To classify the sentiment of all statements, we used crowdsourcing (see below). Our classification scheme follows approaches like [5], which see sentiment as a combination of subjectivity and polarity: We distinguish objective *facts* from subjective *opinions*. The latter are either *positive* or *negative*.

With the term *product features*, on the one hand we refer to *aspects*, such as those given above or others like "atmosphere". On the other hand, a product feature can be anything that is called an *amenity* in the hotel domain. Examples are facilities, e.g. "coffee maker" or "wifi", and services like "laundry". All mentions of such product features have been manually annotated in the corpus.

Figure 2 shows a sample text from the corpus, exemplifying the typical writing style often found in web user reviews: A few grammatical inaccuracies (e.g. inconsistent capitalization) and colloquial phrases (e.g. "like 2 mins walk"), but easily readable. More importantly, Figure 2 illustrates the corpus annotations. Each text has a specified title and body. In this case, the body spans nine mentions of product features, such as "location" or "internet access". It is segmented into 12 facts and opinions, which reflect the review's rather negative sentiment score 2 while e.g. showing that the internet access was not seen as negative.

The general numbers of corpus annotations are listed in Table 2 together with some statistics. The corpus includes 31,006 classified statements and 24,596 product features. On average, a text comprises 14.76 statements and 11.71 product features. Figure 3(a) shows a histogram of the text length in the number of statements, grouped into intervals. As can be seen, over one third of all texts span less than 10 statements (intervals 0-4 and 5-9), whereas less than one fourth spans 20 or more. Figure 3(b) visualizes the distribution of sentiment scores for all intervals that cover at least 1% of the corpus. Most significantly, the fraction of reviews with sentiment score 3 increases under higher numbers of statements. This matches the intuition that long reviews may indicate so-so experiences.

Table 2. Statistics of the tokens, sentences, manually classified statements, and manually annotated product features in the ArguAna TripAdvisor corpus.

Type	**Total**	**Average**	**Std. dev.**	**Median**	**Min**	**Max**
tokens	442,615	210.77	171.66	172	3	1823
sentences	24,162	11.51	7.89	10	1	75
statements	**31,006**	**14.76**	**10.44**	**12**	**1**	**96**
facts	6,303	3.00	3.65	2	0	41
positive opinions	11,786	5.61	5.20	5	0	36
negative opinions	12,917	6.15	6.69	4	0	52
product features	**24,596**	**11.71**	**10.03**	**10**	**0**	**180**

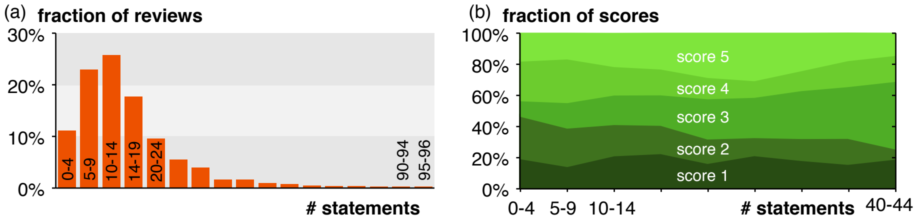


Fig. 3. (a) Histogram of the number of statements in the texts of the ArguAna TripAdvisor corpus. The numbers are grouped into intervals. (b) Interpolated curves of the fraction of sentiment scores in the corpus depending on the numbers of statements.

3.3 Annotation by Web Users and Review Experts

Most hotel reviews are written by regular travelers and hence reflect the argumentation of average web users rather than review experts. Consequently, the classification of a statement as being a fact, a positive, or a negative opinion is in general a straightforward task. For this reason, we let web users annotate the sentiment of all 31,006 statements in our corpus using crowdsourcing. In particular, we relied on *Amazon Mechanical Turk* [1] where so called *workers* can be requested to perform *Human Intelligence Tasks* (HITs) and are paid a small amount of money in case their results are approved by the requester.

The HIT that we assigned to the workers involved the classification of 12 statements. To make the task as simple as possible, we experimented with different task descriptions. The main question of the final description was the following:

> *"When visiting a hotel,*
> *are the following statements positive, negative, or neither?"*

Below, we added notes: (1) to pick "neither" only for facts, not for unclear cases, (2) to pay attention to subtle statements where sentiment is expressed implicitly or ironically, and (3) to pick the most appropriate answer in controversial cases. A carefully chosen set of examples was given to illustrate the different cases.

The workers were allowed to work on a HIT at most 10 minutes and were paid $0.05 for an approved HIT. To assure quality, we assigned the HITs only to workers with over 1,000 approved HITs and an average approval rate of at least 80%. Moreover, we always put two hidden check statements with known and unambiguous classification among the 12 statements in order to recognize faked or otherwise flawed answers. The workers were informed that HITs with incorrectly classified check statements are rejected. For a consistent annotation, we assigned each statement to three workers and then applied majority voting to obtain the final classifications. Rejected HITs were reassigned to other workers.

Altogether, we received 14,187 HITs from 328 workers with an approval rate of 72.8%. On average, a worker spent 75.8 seconds per HIT. We measured the inter-annotator agreement for all statements, resulting in the value 0.67 of *Fleiss' Kappa* [6], which is interpreted as "substantial agreement". 73.6% of the statements got the same classification from all workers and 24.7% had a 2:1 vote (4.8% with opposing opinion polarity). The remaining 1.7% mostly referred to controversial statements, e.g. *"nice hotel, overpriced"* or *"It might not be the Ritz"*. So, we classified them ourselves in the context of the associated review.

Compared to the statement classifications, the annotation of product features is more complex since it requires to mark zero or more appropriate spans within a given text fragment. Moreover, the concept of a product feature is not clear by itself in the hotel domain. This renders crowdsourcing problematic, as it opens the door to ambiguities. In fact, a preliminary study produced very unsatisfying answers with a rejection rate of 43.3%. Thus, we decided to let two experts with linguistic background annotate the corpus, one from a university and one from our partner *Resolto Informatik GmbH.* We gave them the following guideline:

> *"Read through each review text. Mark all product features of the reviewed hotel in the sense of hotel aspects, amenities, services, and facilities."*

For clarity, we specified (1) to omit attributes of product features, e.g. to mark "location" instead of "central location" and "coffee maker" instead of "in-room coffee maker", (2) to omit guest belongings, and (3) not to mark the word "hotel" or brands like "Bellagio" or "Starbucks". Again, we gave a set of examples.

Based on 30 initial texts, we discussed and revised the annotations produced so far with each expert. Afterwards, the experts annotated all other texts from the corpus, taking about 5 minutes per text on average. To measure agreement, 633 statements were annotated twice. In 546 cases, the experts marked the same set of product features, which results in the value $\kappa = 0.73$ for *Cohen's Kappa* [6], assuming a chance agreement probability of 0.5.

3.4 Standard Corpus Format and Tool Support

The ArguAna TripAdvisor corpus comes as an 8 MB packed zip archive (28 MB uncompressed), which contains XMI files preformatted for the *Apache UIMA* framework, the industry standard for natural language processing applications [2]. Such an XMI file stores a text followed by its annotations, while the possible types of annotations are specified in a global *type system descriptor file.*

In addition, we converted all 196,865 remaining reviews of the original dataset with a correct text and a correct sentiment score between 1 and 5 into the same format without manual annotations but with all TripAdvisor ratings and metadata. This unannotated dataset (265 MB; 861 MB uncompressed) can be used both for semi-supervised learning techniques similar to [13] and for a large-scale

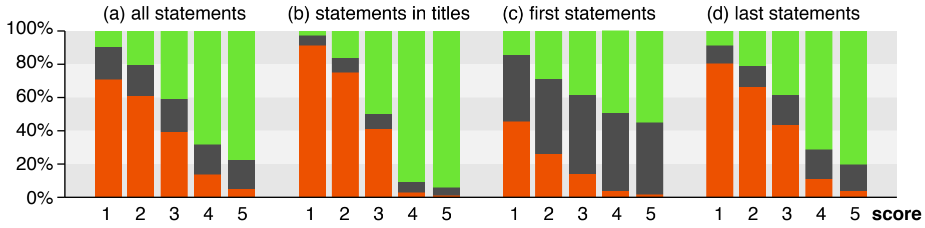


Fig. 4. (a) The fractions of positive opinions (light green), negative opinions (medium red), and objective facts (dark gray) in the texts of the ArguAna TripAdvisor corpus, separated by sentiment score. (b–d) The fractions for specific positions of statements.

evaluation of sentiment score prediction and the like. Also, we attached some software tools and UIMA-compliant text analysis algorithms with associated *UIMA analysis engine descriptor files* to the corpus. They can be executed to conduct the following analyses, thereby demonstrating how to process the corpus.

4 Analysis of Review Argumentation on the Corpus

In this section, we report on statistical analyses of the ArguAna TripAdvisor corpus. In particular, we focus on the questions how and to which extent the local sentiment in a review text determines the review's global sentiment.

4.1 The Impact of the Local Sentiment Distribution

First, we investigate how the distribution of local sentiment in a review text affects the review's global sentiment score. Intuitively, the larger the fraction of positive opinions, the better the sentiment score, and vice versa. More precisely:

> **Hypothesis 1.** *The global sentiment score of a hotel review correlates with the ratio of positive and negative opinions in the review's text.*

As can be seen in Figure 4(a), Hypothesis 1 turns out to be true statistically for our corpus. On average, a review with sentiment score 1 contains 71% negative and 9.4% positive opinions. This ratio decreases strictly monotonously under increasing sentiment scores down to 5.1% negative and 77.5% positive opinions for sentiment score 5. Interestingly, the fraction of facts remains quite stable close to 20% in all cases. To further analyze the connection of local and global sentiment, we computed the distributions of opinions and facts in the review titles as well as in the first and last statements of the review's bodies. Based on the results shown in Figure 4(b–d), we checked for evidence for or against Hypothesis 2:

> **Hypothesis 2.** *The global sentiment score of a hotel review correlates with the polarity of opinions at certain positions of the review's text.*

Compared to Figure 4(a), the distributions for titles in Figure 4(b) entail much stronger gaps in the above-mentioned ratio with a rare appearance of facts, suggesting that the sentiment polarity of the title often reflects the polarity of the whole review. Conversely, over 40% of all first statements denote facts, irrespective of the sentiment score (cf. Figure 4(c)). This number may originate in the

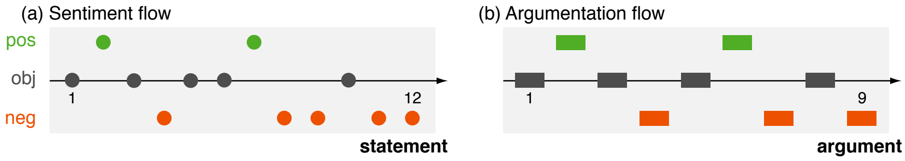


Fig. 5. Illustrations of the local sentiment in the sample text from Figure 2: (a) The *sentiment flow*, i.e., the sequence of all statement sentiments. (b) The *argumentation flow*, where consecutive statements with the same sentiment belong to the same argument.

introductory nature of first statements. It implies a limited average impact of the first statement on a review's sentiment score. So, both the titles and first statements support Hypothesis 2. In contrast, the distributions in Figure 4(d) do not differ clearly from those in Figure 4(a). A possible explanation is that last statements often serve as summaries, but they may also simply reflect the average.

4.2 The Impact of the Local Sentiment Flow

Knowing that both the distribution and the positions of local sentiment have an impact, we next look at the importance of the structure of review texts. For generality, we do not consider the title of a review text as part of its structure, since unlike in our corpus many review texts do not have a title.

To quantify the impact of the structure, we analyze the flow of local sentiment in review texts. In accordance with [9], we define the *sentiment flow* of a text as the sequence of all statement sentiments in the body of the text, where by *sentiment* we either mean the *positive* or *negative* polarity of an opinion or the *objective* nature of a fact. As an example, we visualize the sentiment flow of the text from Figure 2 in Figure 5(a). Our hypothesis is the following:

> **Hypothesis 3.** *The global sentiment score of a hotel review depends on the flow of local sentiment in the review's text.*

Our method to test Hypothesis 3 is to first determine common *flow patterns* in the corpus, i.e., flows of local sentiment that occur in a significant fraction of all texts in the corpus. Then, we check how much these patterns correlate with certain sentiment scores. From an analysis perspective, the two quantifications underlying these steps can be viewed as measuring recall and precision: We define the *recall* **R** of a flow pattern in a given corpus as the fraction of all texts in the corpus where the flow pattern occurs. The *precision* **P(s)** of a flow pattern with respect to a sentiment score **s** is the fraction of texts with sentiment score **s** under all texts in the given corpus where the flow pattern occurs.

However, the only five sentiment flow patterns with a recall of at least 1% in our corpus (i.e., more than 20 texts) are trivial without any change in local sentiment. In [9], improvements are obtained by ignoring the objective facts. Our according experiments did not yield new insights except for a higher recall of the trivial patterns. We thus omit to present their results here, but the results can be easily reproduced using our software tools. The problem lies in the high variance of the reviews' lengths (cf. Figure 3(a)). While a solution is to length-normalize

Table 3. The 13 argumentation flow patterns with the highest recall **R** in the ArguAna TripAdvisor corpus and their precision **P(s)** with respect to each sentiment score **s**.

#	**Argumentation flow**	**R**	**P(1)**	**P(2)**	**P(3)**	**P(4)**	**P(5)**
1	(pos)	7.7%	1.9%	3.1%	7.5%	31.1%	**56.5%**
2	(obj)	5.3%	3.6%	13.6%	20.0%	**33.6%**	29.1%
3	(neg)	3.5%	**58.9%**	30.1%	9.6%	1.4%	–
4	(pos, obj, pos)	3.0%	–	–	6.5%	35.5%	**58.1%**
5	(obj, pos)	2.7%	–	1.8%	7.0%	31.6%	**59.6%**
6	(pos, neg, pos)	2.1%	–	15.9%	11.4%	**56.8%**	15.9%
7	(obj, pos, obj, pos)	1.9%	–	–	5.1%	35.8%	**59.0%**
8	(pos, neg)	1.7%	11.1%	**36.1%**	33.3%	19.4%	–
9	(neg, obj, neg)	1.7%	**88.9%**	8.3%	2.8%	–	–
10	(obj, pos, neg, pos)	1.5%	–	3.2%	**32.3%**	**32.3%**	**32.3%**
11	(neg, pos, neg)	1.5%	35.5%	**51.6%**	12.9%	–	–
12	(obj, neg, obj, neg)	1.1%	**77.3%**	18.2%	4.5%	–	–
13	(obj, neg)	1.1%	**83.3%**	16.7%	–	–	–

sentiment flows, a reasonable normalization is not straightforward. Instead, here we propose to move from statements to *arguments*, where we take the very simplyfing view that a single argument is a sequence of consecutive statements with the same sentiment. The following example shows the rationale behind:

Argument 1. [*I love that hotel!*]$_{\text{stmt1}}$ [*Huge rooms, great location...*]$_{\text{stmt2}}$
Argument 2. [*but it's so expensive!!!*]$_{\text{stmt3}}$

Though the first two statements discuss different topics, the second can be seen as an *elaboration* of the first one in the discourse sense [8]. The third statement contrasts the others, thus denoting a different argument. Based on the notion of arguments, we define the *argumentation flow* of a text as the sequence of all argument sentiments in the body of the text, as illustrated in Figure 5(b).

In total, 826 different argumentation flows exist in our corpus. Table 3 lists the flow patterns with a recall of at least 1%. They cover 34.8% of the corpus texts. The highest-recall pattern *(pos)* represents all 161 fully positive texts (7.7%). Patterns with a high precision **P(5)** are made up only of objective and positive arguments (table line 4, 5, and 7). Quite intuitively, typical patterns of reviews with sentiment score 2 and 4 are *(neg, pos, neg)* and *(pos, neg, pos)*, respectively, whereas none of the listed patterns clearly indicates sentiment score 3. The highest correlation is observed for *(neg, obj, neg)*, which results in sentiment score 1 in 88.9% of the cases. While such correlations offer strong evidence for Hypothesis 3, all 13 patterns cooccur with more than one sentiment score. Consequently, the structure of a review text does not decide the global sentiment alone.

4.3 The Impact of the Local Sentiment regarding Product Features

Finally, we quantify the impact of the content of a hotel review, which is represented by the product features discussed within the review's text:

Hypothesis 4. *The global sentiment score of a hotel review correlates with the polarity of opinions on certain product features in the review's text.*

Table 4. A selection of the 25 product features with highest recall **R** in the ArguAna TripAdvisor corpus, the fractions of their positive (**pos**) and negative (**neg**) mentions, and the precision with respect to sentiment score 1 and 5 depending on these polarities.

#	**Feature**	**R**	**pos**	$\mathbf{P_{pos}(1)}$	$\mathbf{P_{pos}(5)}$	**neg**	$\mathbf{P_{neg}(1)}$	$\mathbf{P_{neg}(5)}$
1	room	**80.3%**	36.9%	7.4%	31.1%	47.8%	38.4%	3.5%
2	staff	43.4%	62.9%	4.3%	38.0%	34.1%	**50.3%**	1.5%
3	location	42.2%	**84.7%**	5.7%	35.9%	**11.8%**	32.5%	1.6%
8	service	18.4%	38.9%	7.4%	44.1%	55.0%	45.1%	–
17	food	7.6%	52.3%	**9.9%**	34.7%	37.3%	45.8%	1.4%
20	towels	5.3%	**27.1%**	7.9%	**21.1%**	**67.1%**	35.1%	3.2%
24	parking	**5.1%**	30.6%	–	**46.3%**	56.0%	**25.3%**	**12.0%**

To investigate the hypothesis, we consider the 25 product features with the highest recall **R** in the corpus. Similar to above, here *recall* means the fraction of all texts where the product feature occurs. First, we compute the fractions of positive and negative mentions of each product feature. For simplicity, we assume that an opinion always refers to the product features it contains. Then, we quantify the correlation between the polarity of a mention and the sentiment score of the respective review by reusing the concept of *precision* from Section 4.2 accordingly. In Table 4, we present a selection of the 25 product features.

The general importance of the *room* is reflected by a recall of 80.3%. The *location* appears most often in positive opinions (84.7%) and *towels* in negative ones (67.1%). However, other aspects and amenities seem to have a larger impact on a review's global sentiment: When e.g. the *staff* is seen as negative, this results in sentiment score 1 in 50.3% of the cases. Even more obvious, a negative mention of *service* never cooccurs with sentiment score 5 (interestingly, *staff* is used more in positive and *service* more in negative contexts). Conversely, we see that a positive *food* experience alone does not make a good hotel ($\mathbf{P_{pos}(1)}$ = 9.9%), and 12% of all negative opinions on *parking* occur in reviews with the highest sentiment score. A good *parking* situation seems to be appreciated, though.

To summarize, our corpus reveals large differences in the impact of product features on a review's global sentiment, which supports Hypothesis 4. We hence conclude that an argumentation analysis of reviews should cover both structure and content. To this end, our results define a promising starting point.

5 Conclusion

The facts and opinions within the argumentation of a review text impact the reputation of products and services. To analyze argumentations, we have designed the freely available ArguAna TripAdvisor corpus based on a balanced collection of hotel reviews. Each review text is annotated with respect to local sentiment and the mentioned hotel aspects and amenities. We have explored the corpus to reveal argumentation patterns that correlate with the reviews' sentiment scores. While the corpus is restricted to hotel reviews, in future work we will investigate to what extent the patterns generalize to other domains. Generally, we believe that an argumentation analysis of texts allows for more effective approaches to sentiment-related tasks. At the same time, it implies new ways to explain obtained results, as it mimics the way humans interpret texts. Currently, we work on an approach that learns argumentation patterns in order to predict and explain sentiment scores. Apart from sentiment, our findings on argumentation may be transferrable to other natural language processing tasks, such as authorship attribution [14] or language function analysis [18]. For this purpose, we will need further resources that cover more domains and types of annotations.

Acknowledgments This work was funded by the German Federal Ministry of Education and Research (BMBF) under contract number 01IS11016A.

References

1. Amazon Mechanical Turk, http://www.mturk.com
2. Apache UIMA, http://uima.apache.org
3. Besnard, P., Hunter, A.: Elements of Argumentation. The MIT Press (2008)
4. Cabrio, E., Villata, S.: Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In: Proc. of the 50th ACL: Short Papers. pp. 208–212 (2012)
5. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: In Proceedings of the 5th LREC. pp. 417–422 (2006)
6. Fleiss, J.L.: Statistical Methods for Rates and Proportions. John Wiley & Sons, second edn. (1981)
7. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proc. of the Tenth SIGKDD. pp. 168–177 (2004)
8. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text 8(3), 243–281 (1988)
9. Mao, Y., Lebanon, G.: Isotonic Conditional Random Fields and Local Sentiment Flow. Advances in Neural Information Processing Systems 19, 961–968 (2007)
10. Mochales, R., Moens, M.F.: Argumentation Mining. AI and Law 19(1), 1–22 (2011)
11. Mukherjee, S., Bhattacharyya, P.: Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In: Proc. of the 24th COLING. pp. 1847–1864 (2012)
12. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1–2), 1–135 (2008)
13. Prettenhofer, P., Stein, B.: Cross-Language Text Classification using Structural Correspondence Learning. In: Proc. of the 48th ACL. pp. 1118–1127 (2010)
14. Sapkota, U., Solorio, T., Montes-y Gómez, M., Rosso, P.: The Use of Orthogonal Similarity Relations in the Prediction of Authorship. In: Proc. of the 14th CICLing. pp. 463–475 (2013)
15. Teufel, S.: Argumentative Zoning: Information Extraction from Scientific Text. Ph.D. thesis, University of Edinburgh (1999)
16. Toulmin, S.E.: The Uses of Argument. Cambridge University Press (1958)
17. TripAdvisor, http://www.tripadvisor.com
18. Wachsmuth, H., Bujna, K.: Back to the Roots of Genres: Text Classification by Language Function. In: Proc. of the 5th IJCNLP. pp. 632–640 (2011)
19. Wang, H., Lu, Y., Zhai, C.: Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. In: Proc. of the 16th SIGKDD. pp. 783–792 (2010)
20. Wiebe, J., Wilson, T., Cardie, C.: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation 1(2) (2005)
21. Zirn, C., Niepert, M., Stuckenschmidt, H., Strube, M.: Fine-Grained Sentiment Analysis with Structural Features. In: Proc. of the 5th IJCNLP. pp. 336–344 (2011)