Sentiment Flow – A General Model of Web Review Argumentation

Henning Wachsmuth and **Johannes Kiesel** and **Benno Stein**
Faculty of Media, Bauhaus-Universität Weimar, Germany
{henning.wachsmuth,johannes.kiesel,benno.stein}@uni-weimar.de

Abstract

Web reviews have been intensively studied in argumentation-related tasks such as sentiment analysis. However, due to their focus on content-based features, many sentiment analysis approaches are effective only for reviews from those domains they have been specifically modeled for. This paper puts its focus on domain independence and asks whether a general model can be found for how people argue in web reviews. Our hypothesis is that people express their global sentiment on a topic with similar sequences of local sentiment independent of the domain. We model such *sentiment flow* robustly under uncertainty through abstraction. To test our hypothesis, we predict global sentiment based on sentiment flow. In systematic experiments, we improve over the domain independence of strong baselines. Our findings suggest that sentiment flow qualifies as a general model of web review argumentation.

1 Introduction

The web is full of user-generated reviews on products, services, and works of art, like those from Amazon, TripAdvisor, and Rotten Tomatoes. Such web reviews provide facts, positive opinions, and negative opinions on different aspects. By that, the reviews express, implicitly or explicitly, an overall opinion on the topic in question. From an abstract viewpoint, the argumentation of a web review can thus be seen as a composition of local sentiments used to justify some global sentiment.

Both local and global sentiment of reviews are in the focus of numerous sentiment analysis approaches (cf. Section 2 for details). Many of these approaches model reviews primarily with content-based features, derived from the words in the reviews. The use of words, however, varies strongly across domains, as illustrated in Figure 1 for a product, a hotel, and a movie review. As a consequence, sentiment analysis suffers from domain dependence (Wu et al., 2010), i.e., high effectiveness is often achieved only in the domain an approach has been specifically modeled for. To adapt to other domains, prior knowledge about these domains or about domain-independent features is needed (Prettenhofer and Stein, 2010).

This paper considers the question as to whether the overall argumentation of web reviews can be modeled in a general way in order to increase domain independence in sentiment analysis. We observe that people structure web reviews largely sequentially—in contrast to the complex structures of many other argumentative texts. While the reviewed aspects differ between domains, our assumption is that the overall argumentation of a web review is generally represented by a sequence of local sentiments, called the review's *sentiment flow* (Mao and Lebanon, 2007). In particular, we hypothesize that, under an adequate model, similar sentiment flows express similar global sentiments, also across domains. All reviews in Figure 1, for instance, express neutral global sentiment starting with positive, continuing with negative, and ending with positive local sentiment.

Unlike in our previous approach (Wachsmuth et al., 2014a), we analyze the major abstraction steps when modeling sentiment flow to represent global sentiment. A general model should abstract from both content and other domain differences, such as a review's length or the density of local sentiment in it. Based on web review corpora with known sentiment flows, we empirically analyze several model variants across three domains. Our results offer clear evidence for the truth of our hypothesis, indicating the generality of sentiment flow as a model of web review argumentation.

The abstract nature of sentiment flow, however, does not directly achieve domain independence, as

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 601–611,
Lisbon, Portugal, 17-21 September 2015.

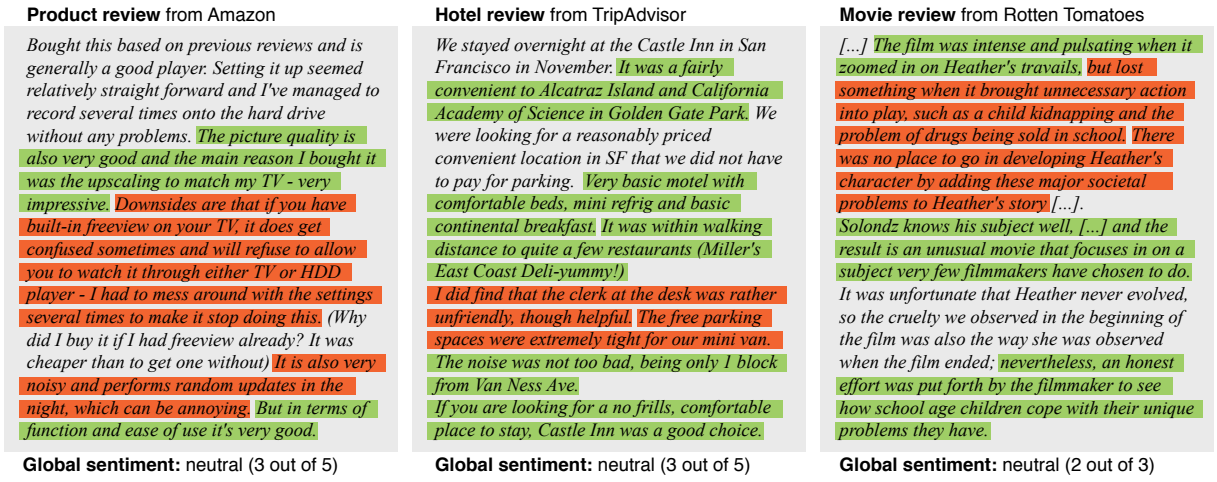

Figure 1. Example web reviews with neutral global sentiment from three domains, taken from the corpora described in Section 5. Corpus annotations of positive and negative local sentiment are marked in light green and medium red, respectively.

the recognition of local sentiment in unknown reviews may still be domain-dependent. We therefore also present a novel edit distance approach to robustly compare flows, when local sentiment is obtained using state-of-the-art techniques (Socher et al., 2013). In systematic cross-domain experiments with the given corpora, we classify global sentiment based on sentiment flow without any domain adaptation. While not being perfectly effective, our approach improves over the domain robustness of strong baselines.

Altogether, the paper's main contributions are:

1. Evidence that sentiment flow qualifies as a general model of the overall argumentation of web reviews across domains.
2. A domain-robust approach for the classification of the global sentiment of web reviews.

2 Related Work

As surveyed by Pang and Lee (2008) and by Liu (2012), numerous sentiment analysis approaches have been proposed for different text types, levels of granularity, sentiment scales, and domains. We target at global text-level sentiment of web reviews. While we distinguish three sentiment classes here, our approach can be adapted to other scales. Our goal is not to optimize sentiment analysis in a specific domain, but to find a model that supports sentiment analysis across domains.

As common in text classification (Manning et al., 2008), sentiment analysis often relies on words and other content features, which tends to be prone to domain dependence (Wu et al., 2010). Existing domain adaptation techniques for sentiment analysis require a few training texts from each target domain or a few domain-independent pivot features to align domain-specific features (Prettenhofer and Stein, 2010). Our model complements these techniques and could be leveraged for pivot features. In tasks like authorship attribution and argumentative zoning, non-topical words benefit domain independence (Menon and Choi, 2011; Ó Séaghdha and Teufel, 2014). Instead, we focus on the local sentiment on different aspects in a review here.

Aspect-based sentiment analysis extracts fine-grained opinions from a review (Popescu and Etzioni, 2005). These aspects in turn impact the review's global sentiment (Wang et al., 2010). However, relevant aspects naturally tend to be domain-specific, like the picture quality of HDD players or the beds of hotels (cf. Figure 1). While weakly-supervised approaches to extract aspects and local sentiment exist (Brody and Elhadad, 2010; Lazaridou et al., 2013), it is not clear how to align aspects from different domains. We ignore aspects here, only preserving the local sentiment itself.

State-of-the-art approaches for classifying local sentiment within a domain model the composition of words, e.g., relying on deep learning (Socher et al., 2013). We do not compete with such an approach, but we use it to then predict global sentiment. Täckström and McDonald (2011) observe that local and global sentiment correlate, aiming for the opposite direction, though. In (Wachsmuth et al., 2014b), we already compute frequent flows of local sentiment, but we neither analyze their generality, nor do we use them for prediction.

The idea of modeling sentiment flow was introduced by Mao and Lebanon (2007) who classify local sentiment based on neighboring local sentiment in a review. When inferring global sentiment from a flow, however, the authors model only

single flow positions, not their ordering. In contrast, we capture the overall structure of reviews in (Wachsmuth et al., 2014a) by measuring the similarity of a given flow to known sentiment flow patterns. We point out the domain robustness of sentiment flow there, but we still use domain-specific local sentiment classifiers and we do not handle some major domain differences of web reviews. Both limitations are addressed in this paper, where we align flows similar to how Persing et al. (2010) align essay organizations.

We claim that sentiment flow models a review's argumentation, such that local sentiments ressemble rhetorical moves. Comparable simplifications are common for scientific argumentation (Teufel, 2014). Usually, argumentative texts are studied more deeply, considering different types of argument components and their relations (Mochales and Moens, 2011). Mining such structure is getting increasing attention recently (Habernal et al., 2014), also in the analysis of reviews (Villalba and Saint-Dizier, 2012). To express global sentiment, however, web reviews argue in simpler ways.

3 Web Review Argumentation

Argumentation refers to the exchange of opinions, to defending positions, and to convincing others of certain stances (van Eemeren et al., 2014). A review is a written form of monological argumentation, where an author structures a selection of arguments in order to justify his or her conclusion on a topic of discussion (Besnard and Hunter, 2008). Reviews, in particular, discuss products, services, works of art, or similar. The arguments in a review correspond to objective facts, positive and negative opinions, and mixtures of these on the topic as a whole or on specific aspects of the topic.

In this paper, we are interested in the *overall argumentation* of reviews. Our assumption is that the conclusion of a review's overall argumentation consists in its global sentiment. Global sentiment is often explicitly reflected by an assigned overall rating, at least for *web reviews*.

Many web reviews are written by people in an ad-hoc fashion to quickly share opinions. As a result, unlike other argumentative texts, web reviews often remain with a sequential structure (Villalba and Saint-Dizier, 2012) and miss explicit relations between the shared opinions. E.g., while the product review and the hotel review in Figure 1 cover opinions on several aspects, no deliberate structure is found in their argumentation. However, the excerpt of the more professional movie review shows that this is not always the case.

3.1 Domain Differences

In Figure 1, we categorize domains by source and topical theme (e.g., Amazon products). Other granularities would be possible (e.g., consumer electronics) or other categorization schemes (e.g., user vs. pro reviews). That being said, we speak of *domains* only to roughly distinguish web reviews that vary in how they argue for a conclusion.[1] We observe major differences in three broad respects:

Content Especially topical web review domains differ widely regarding the terms and phrases that play a role in their argumentation. This includes the aspects being discussed (e.g., "beds" of hotels vs. the "subject" of movies) as well as the words used to express sentiment and their explicitness (e.g., "yummy" vs. "an unusual movie").

Form As sketched above, further differences refer to the structure and style of web reviews. Some are rather subtle, like a careful use of paragraph breaks, whereas others are obvious, like a review's length. The movie review in Figure 1, for instance, is actually over twice as long as the shown excerpt, starting with an objective synopsis of the plot and including "sub-reviews" of different aspects.[2]

Subjectivity Finally, the use of subjectivity varies across web review domains: First, the density of sentiment tends to be high in some cases, like hotel reviews (cf. Figure 1), but low in others, like movie reviews, where objective plot descriptions and subjective opinions often alternate (the shortened excerpt in Figure 1 hides this to some extent). Second, sentiment is sometimes very intense (as in the product review in Figure 1), sometimes subtle. And third, in some domains even single sentences often contain mixed sentiment, whereas in others opinions tend to be laid out across sentences.

We will empirically underpin most observations in Section 5, where we analyze the domain independence of the model described next.

4 Sentiment Flow as a General Model

In the following, we introduce our model of web review argumentation. We discuss how to abstract for generality and how to deal with uncertainty.

[1] In the end, this paper seeks for findings that generalize from domains, making an exact distinction unnecessary.

[2] Also, many web reviews have explicit structure elements like a title. To obtain a common ground, however, we consider only the plain text of a web review's body in this paper.

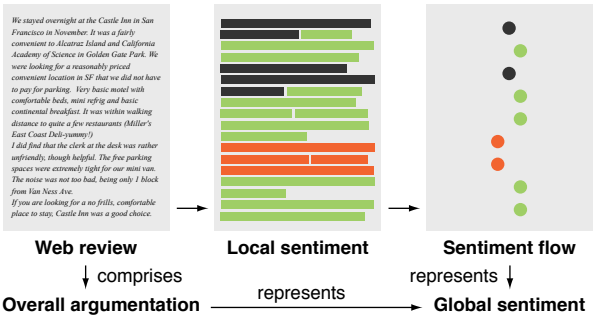

Figure 2. Modeling the overall argumentation of a web review as a flow of positive (light green), neutral (dark gray), and negative (medium red) local sentiments.

Figure 3. The original sentiment flow from Figure 2 and the resulting flow for each of the three proposed transformations.

4.1 Modeling a Review by its Sentiment Flow

We propose a fairly simple argumentation model based on the observation that many web reviews are organized sequentially (cf. Section 3). As we assume that the overall argumentation of a web review represents global sentiment in the first place, we fully abstract from the content of the facts and opinions that serve as arguments. In particular, we model the argumentation of a web review solely by its *sentiment flow*, i.e., the sequence of local sentiments comprised in the review's text. We do not presume the granularity of local sentiment, but we propose to distinguish positive, neutral, and negative local sentiment, which is the common ground of related works (cf. Section 2).

Figure 2 illustrates how we model web review argumentation. Our hypothesis is that similar sentiment flows are used across domains of web reviews to express the same global sentiment. However, because of the domain differences described in Section 3, we do not expect that the *original* sentiment flows of web reviews generalize well.

4.2 Abstracting Flows for Generality

By concept, sentiment flow avoids to capture content and some facets of form like paragraph usage. To abstract from the length of reviews, Mao and Lebanon (2007) and our approach in (Wachsmuth et al., 2014a) length-normalize sentiment flow via interpolation. While this may preserve all information, it does not account for sub-reviews and the density of subjectivity. Here, we investigate more informed ways of abstracting flows. In particular, we consider three transformations of flows:

Change Deletion of repeating local sentiments. The rationale is to reduce subjectivity differences by focusing on changes of local sentiment.

NoLoops Deletion of repeating sequences of two or more local sentiments. The rationale is to reduce length differences by merging similar sub-reviews.

2Class Deletion of neutral local sentiments. The rationale is to reduce length and subjectivity differences emanating from objective descriptions.

Figure 3 exemplifies the three transformations. They are not commutative, as can partly be seen for the example. In Section 5, we test what combinations of transformations lead to an adequate sentiment flow model. While more transformations will benefit generality, the lost specificity may decrease the correlation with global sentiment.

4.3 Analyzing Flows under Uncertainty

Given an adequate sentiment flow model, we seek to find out to what extent it enables domain-robust sentiment analysis. This brings up two challenges related to uncertainty: (1) The classification of local sentiment in unknown reviews will not be free of errors, and, (2) reviews may comprise flows for which the global sentiment is unknown.

Classification errors are naturally problematic for modeling sentiment flow. At least, some errors are bypassed by the three transformations. E.g., if one negative local sentiment in the original flow in Figure 3 is misclassified as positive, the *Change* transformation fixes this. If it is classified as neutral, *Change* and *2Class* together eliminate the effect. Moreover, errors can be countered by limiting the impact of single positions in a flow.

In (Wachsmuth et al., 2014a), we learn to infer global sentiment from the Manhattan distances between a sentiment flow and a set of common flows, thereby analyzing the flow as a whole. The common flows are found in a preceding clustering step. While we adopt the learning approach here, the Manhattan distances imply that flows are similar only if their changes are at similar positions.

Instead, we compare sentiment flows (modified with zero to three transformations) based on their normalized minimum *edit distance* (Cormen et al., 2009). Analog to Persing et al. (2010), we incrementally compute the edit distance using sequence alignment. To this end, we specify costs for pos-

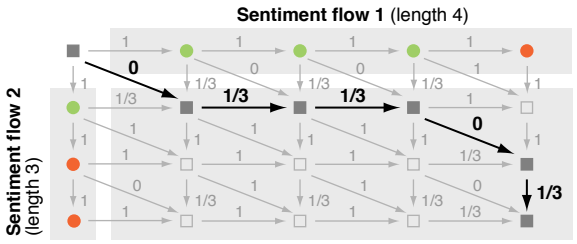

Figure 4. Computation of the normalized edit distance of two sentiment flows, resulting in $(2 \cdot 0 + 3 \cdot 1/3) \,/\, 4 = 1/4$.

sible edit operations, i.e., substitutions, insertions, and deletions of single local sentiments. We map positive local sentiment to the value 1.0, neutral to 0.5, and negative to 0.0. The cost is then provided by a function d for any two values s and s':

$$d(s, s') = \begin{cases} |s - s'| & \text{If } s' \text{ substitutes } s. \\ \alpha + (1-\alpha) \cdot |s - s'| & \text{If } s' \text{ is inserted or deleted after } s. \end{cases}$$

Here, $\alpha \in [0, 1]$ specifies some fixed cost (we set α to $1/3$ in Section 6). The intuition behind d is to have a higher cost the more s and s' differ. Still, insertions and deletions are never free, as they affect differences that remain after applying transformations to abstract from irrelevant differences.

Figure 4 illustrates the alignment of two flows as a shortest-path search. We normalize the flows' minimum edit distance by their maximum length. Before we evaluate if the edit distance captures flow similarity more robustly than the Manhattan distance, we analyze what representation of sentiment flows proves most general. This will also reveal that the proposed abstractions reduce the need to perform clustering for finding common flows.

5 Analysis of the Generality of the Model

We now report on experiments on corpora from three domains that empirically analyze to what extent different sentiment flow variants qualify as general models of web review argumentation.[3]

5.1 Ground-Truth Data with Sentiment Flow

We process three existing corpora with local sentiment annotations of complete texts. While the first two are available online, we obtained the last from the authors. Each corpus comprises English web reviews from one broad topical domain. Table 1 lists some statistics of the three corpora, which indicate clear domain differences.

Product Domain The *Finegrained Sentiment Data Set, Release 1* (Täckström and McDonald, 2011) contains 294 Amazon reviews, nearly balanced among five categories: books (59 reviews), DVD (59), electronics (57), music (59), and videogames (60). We use the first three for training and the others for testing. Under the authors' mapping from Amazon star ratings to global sentiment, all categories subsume 19 to 20 positive, neutral, and negative reviews each. In each review, every sentence is classified as positive, negative, neutral, mixed, or irrelevant. To match the other corpora, we merge the three latter into one neutral class.

Corpus domain	Sentences per text	Tokens per sent.	Local sentiment positive	neutral	negative
Product	14.0	22.7	24.1%	41.5%	34.4%
Hotel	11.5	18.3	38.0%	20.3%	41.7%
Movie	28.8	30.3	17.6%	61.2%	21.2%

Table 1. Sentences, tokens, and annotated local sentiments for the domains represented by the given web review corpora.

Hotel Domain Our *ArguAna TripAdvisor corpus* (Wachsmuth et al., 2014b) consists of 2 100 TripAdvisor reviews, 300 for seven hotel locations each. Three locations belong to a predefined training set and two to a validation and a test set each. For all locations, the reviews are evenly distributed over the five TripAdvisor overall scores. In accordance with the product corpus, we see score 4–5 as positive global sentiment, 3 as neutral, and 1–2 as negative. In each review, all main clauses together with their subordinate clauses have been classified as being positive, negative, or neutral.

Movie Domain Finally, the third corpus (Mao and Lebanon, 2007) compiles 450 Rotten Tomatoes reviews from the *Cornell Movie Review Data scale dataset v1.0* (Pang and Lee, 2005) that refer to two authors. We use the 201 reviews of Scott Renshaw for training and the 249 of Dennis Schwartz for testing. The reviews lack punctuation, capitalization, and their overall ratings. We recovered the overall ratings from the original dataset based on the rating scale 0–2, resulting in 178 positive, 139 neutral, and 133 negative reviews. In each review, Mao and Lebanon (2007) classified all sentences to be very positive, positive, neutral, negative, or very negative, which we reduce to three classes.

5.2 Experimental Set-up

To find the most general model of web review argumentation across domains, we compare 16 sentiment flow variants using three measures:

Model Variants The original sentiment flow of all corpus reviews can be directly derived from the ground-truth data. In each model variant, the flow is modified by a combination of zero to three of the

[3] The source code that can be used to reproduce the experiments is provided at http://www.arguana.com/software.

Training domain	Model variant of sentiment flows	# Flows		Aggregate recall			Weighted precision			W'd Hellinger distance		
		all	1%	Product	Hotel	Movie	Product	Hotel	Movie	Product	Hotel	Movie
Product	2class-change-noloops	7	7	100.0	100.0	100.0	75.6	69.9	62.4	0.17	0.21	0.23
175 reviews	2class-noloops-change	11	7	90.8	96.0	98.0	74.1	71.0	64.6	0.17	0.26	0.28
	change-noloops-2class	22	14	89.9	81.5	80.7	74.8	74.6	68.3	0.19	0.26	0.26
	noloops-2class-change	11	8	89.9	86.1	86.9	73.8	72.3	66.2	0.16	0.24	0.27
	2class-change	12	8	89.9	85.6	87.1	74.8	73.2	66.3	0.17	0.25	0.27
	change-2class-noloops	37	20	**85.7**	**85.8**	**76.9**	78.4	74.1	72.5	0.19	0.26	0.33
	noloops-change-2class	35	19	81.5	72.8	62.2	76.3	76.3	73.6	**0.17**	0.26	**0.28**
	2class-noloops	49	24	77.3	62.9	60.0	81.5	**80.8**	**76.7**	0.27	0.27	0.31
	change-2class	47	22	77.3	61.6	49.6	79.3	79.5	71.7	0.20	**0.24**	**0.28**
	change-noloops	55	29	70.6	64.6	59.1	**84.5**	77.2	72.2	0.25	0.30	0.32
	noloops-2class	67	24	62.2	48.2	36.7	85.1	82.7	78.8	0.18	0.25	0.22
	noloops-change	78	29	55.5	52.0	30.9	89.4	79.7	77.0	0.16	0.28	0.23
	change	93	30	49.6	43.5	27.3	89.8	82.6	75.6	0.18	0.25	0.22
	2class	91	27	45.4	36.8	28.7	83.3	85.1	79.8	0.21	0.22	0.22
	noloops	153	17	12.6	14.0	6.0	100.0	91.5	85.2	0.07	0.11	0.04
	original	173	2	0.8	3.6	0.0	100.0	94.7	0.0	0.02	0.05	0.00
Hotel	2class-change-noloops	7	7	100.0	100.0	100.0	71.1	68.5	62.4	0.19	0.06	0.13
900 reviews	2class-noloops-change	20	14	99.7	98.8	99.8	72.0	69.3	64.8	0.25	0.08	0.17
	change-noloops-2class	85	21	99.0	91.2	91.3	74.6	72.0	67.9	0.26	0.14	0.19
	noloops-2class-change	27	15	100.0	98.3	99.6	72.4	69.8	65.2	0.26	0.10	0.17
	2class-change	31	17	99.7	98.2	98.9	72.7	70.6	65.6	0.26	0.11	0.19
	change-2class-noloops	91	22	**92.9**	**91.7**	**85.3**	75.5	72.5	72.7	**0.23**	**0.14**	0.26
	noloops-change-2class	145	21	90.5	85.3	76.7	74.4	74.6	73.0	0.28	0.15	0.29
	2class-noloops	246	19	85.0	77.2	75.6	78.0	**81.0**	**75.6**	0.27	0.20	0.27
	change-2class	231	19	88.1	74.0	56.2	75.7	76.4	73.5	0.26	0.17	0.28
	change-noloops	212	24	69.7	77.7	66.7	**80.0**	75.5	73.7	0.20	0.18	0.**24**
	noloops-2class	398	14	64.6	55.3	44.0	81.6	82.5	77.8	0.19	0.15	0.26
	noloops-change	343	17	48.0	64.0	38.0	81.6	77.9	77.8	0.19	0.15	0.21
	change	426	14	36.1	52.3	24.7	83.0	77.7	77.5	0.18	0.14	0.16
	2class	549	17	54.4	32.5	29.8	83.1	86.2	77.6	0.14	0.08	0.17
	noloops	626	9	9.2	27.7	1.8	92.6	83.7	100.0	0.04	0.08	0.01
	original	743	4	1.4	16.5	0.0	75.0	78.8	0.0	0.01	0.04	0.00
Movie	2class-change-noloops	6	6	97.3	94.5	99.6	71.7	70.3	58.9	0.26	0.10	0.19
201 reviews	2class-noloops-change	14	10	96.9	93.7	99.2	72.6	70.7	59.9	0.32	0.15	0.24
	change-noloops-2class	44	17	96.6	85.4	91.6	75.4	73.9	62.7	0.34	0.22	0.32
	noloops-2class-change	16	14	96.9	92.2	98.0	73.0	71.0	59.8	0.33	0.16	0.26
	2class-change	19	15	96.9	90.8	98.0	73.3	72.1	61.1	0.33	0.18	0.28
	change-2class-noloops	57	19	**85.7**	**73.0**	**81.1**	76.2	73.4	68.8	0.36	0.25	0.30
	noloops-change-2class	84	19	82.0	63.0	72.7	77.6	72.3	71.8	0.34	0.25	0.27
	2class-noloops	103	20	78.2	58.0	62.7	78.7	**85.1**	**72.4**	0.34	0.24	0.27
	change-2class	107	20	57.5	34.4	40.6	72.8	76.2	72.3	**0.33**	0.22	**0.22**
	change-noloops	94	17	66.7	41.6	66.7	**81.6**	81.0	70.5	0.35	**0.21**	0.33
	noloops-2class	154	8	38.8	22.1	28.5	86.8	89.7	85.9	0.19	0.14	0.14
	noloops-change	146	9	30.6	19.7	30.5	87.8	83.3	77.6	0.16	0.14	0.18
	change	161	8	16.0	9.9	13.7	85.1	87.5	82.4	0.15	0.10	0.12
	2class	182	5	21.8	11.1	8.0	90.6	96.6	95.0	0.08	0.06	0.03
	noloops	200	5	0.3	0.6	0.4	100.0	91.7	100.0	0.01	0.01	0.00
	original	200	0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00

Table 2. Results on the generality of sentiment flow for all evaluated model variants on ground-truth data for each combination of training and test domain. The most general variants in terms of both aggregated recall and weighted precision are marked in bold. For illustration, *# Flows* lists the numbers of *all* flows in the training reviews and of those with a recall of at least *1%*.

transformations from Section 4. All variants are named according to the applied transformations.

Measures In (Wachsmuth et al., 2014b), we propose specific notions of the recall and precision of a sentiment flow f in a given collection of reviews: The recall R_f denotes the relative frequency of reviews with flow f, while the precision $P_f(\mathrm{s})$ with respect to some global sentiment s denotes the relative co-occurrence of f with s. Here, we extend these measures for complete models as follows.

We define the *aggregate recall* of a model on a collection of reviews as the sum of the recall of the set F of all its known sentiment flows:

$$\textit{Aggregate Recall}(F) = \sum_{f \in F} R_f$$

With *weighted precision*, we denote the sum of the maximum precision of each such flow in F, weighted with the recall of the flow:

$$\textit{Weighted Precision}(F) = \sum_{f \in F} \max_{\mathrm{s}} \{P_f(\mathrm{s})\} \cdot R_f$$

In addition, we assess how much two domains differ under a given model variant. To this end, we measure the *Hellinger distance* H_f (in the range $[0, 1]$) between the global sentiment distributions of each flow f known for both domains:

Model variant	Domain	Most common flow	Rank	Recall	Positive	Neutral	Negative
change-2class-noloops	**Product**	(negative, negative)	**1.**	13.6	0.0 %	25.0 %	75.0 %
	Hotel		9.	4.2	0.0 %	8.9 %	91.1 %
	Movie		4.	6.7	0.0 %	0.0 %	100.0 %
	Product	(positive, negative, positive)	8.	3.4	34.1 %	65.9 %	0.0 %
	Hotel		**1.**	10.5	45.1 %	49.6 %	5.3 %
	Movie		14.	2.3	0.0 %	73.3 %	26.7 %
	Product	(negative, negative, positive, negative)	10.	3.4	0.0 %	33.3 %	66.7 %
	Hotel		15.	2.2	0.0 %	16.7 %	83.3 %
	Movie		**1.**	8.6	0.0 %	57.9 %	42.1 %
2class-noloops	**Product**	(negative, negative)	**1.**	15.3	7.6 %	29.6 %	62.8 %
	Hotel		5.	4.4	0.0 %	8.5 %	91.5 %
	Movie		**1.**	6.7	0.0 %	0.0 %	100.0 %
	Product	(positive, positive)	3.	12.8	86.8 %	8.8 %	4.4 %
	Hotel		**1.**	7.6	87.8 %	12.2 %	0.0 %
	Movie		10.	2.1	61.1 %	38.9 %	0.0 %
change-noloops	**Product**	(positive, neutral, positive)	**1.**	10.5	94.6 %	0.0 %	5.4 %
	Hotel		6.	3.3	88.9 %	11.1 %	0.0 %
	Movie		23.	1.3	100.0 %	0.0 %	0.0 %
	Product	(positive)	22.	1.1	50.9 %	49.1 %	0.0 %
	Hotel		**1.**	6.6	83.1 %	14.1 %	2.8 %
	Movie		–	–	–	–	–
	Product	(neutral, negative)	2.	9.1	6.5 %	24.9 %	68.6 %
	Hotel		5.	3.5	0.0 %	10.5 %	89.5 %
	Movie		**1.**	7.6	8.6 %	0.0 %	91.4 %

Table 3. The most common flow in the training set of each evaluated domain for three of the 16 evaluated model variants. For each flow, the recall rank, the recall, and the distribution over the three global sentiments within each domain are given.

$$H_f(\mathbf{p}_1, \mathbf{p}_2) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{\mathrm{s}} (\sqrt{\mathbf{p}_1(\mathrm{s})} - \sqrt{\mathbf{p}_2(\mathrm{s})})^2}$$

Here, $\mathbf{p}_1$ and $\mathbf{p}_2$ denote the global sentiment distributions of f. For *weighted Hellinger distances*, we multiply the distance of each flow in F with the sum of its recall in both domains.[4]

Experiments Given all 16 possible model variants for all reviews, we analyze the generality of each variant for every combination of domains. I.e., we first determine the known sentiment flows on the training set of one domain. Then, we compute the aggregate recall, weighted precision, and weighted Hellinger distance once for the in-domain test set and once for both full out-of-domain corpora.[5]

5.3 Results on the Generality across Domains

Table 2 contains the number of known flows and the experiment results for each domain combination. Model variants whose benefit seems limited are not marked in bold: The bottom six have a low aggregate recall in all domains, suggesting that they do not generalize well. Most significantly, the *original* flows from the movie training set are not found in any test domain. The top five achieve almost total recall, but much less precision than the others, indicating that they abstract too much.

Among the five robust model variants (marked in bold), *change-2class-noloops* has the highest aggregate recall throughout, ranging from 73.0 to 92.9. Consistently, global sentiment is represented best by *change-noloops* in the product domain and by *2class-noloops* in the other domains (with up to 85.1 weighted precision). Also, *2class-noloops* is third-best in terms of recall. While no clear "winner" exists, this variant seems most promising for modeling web review argumentation.

The weighted Hellinger distances show that the domain differences of many variants are small. On average, *change-2class* has the most stable global sentiment distribution. Most distances are only slightly higher out-of-domain than in-domain or even lower. Hence, sentiment flows hardly vary stronger across domains than within a domain.

5.4 The Most Common Sentiment Flows

To investigate what sentiment flows actually occur in web reviews, we determined the flow with highest recall for the training set of each corpus. For comparability, we balanced the flows in the training set before by weighting their occurrences according to the distribution of global sentiment.[6]

[4] We chose the Hellinger distance, as it applies to distributions with zero-probabilities (unlike alternatives like the KL-divergence). Also, it is a true metric (Lebret and Collobert, 2014), allowing for relative comparisons. On the flipside, the meaning of concrete distances is not clear by itself.

[5] Here, we use *all* occurring sentiment flows to evaluate a model variant in its overall manifestation. In Section 6, we consider only frequent flows in order to refrain from outliers.

[6] E.g., if 40% of all reviews are positive, 30% neutral, and 30% negative, then the occurrences of flows with positive global sentiment are weighted by 0.75 and the others by 1.0.

Table 3 shows the recall and the sentiment distribution of each such flow in all evaluated domains exemplarily for three of the model variants discussed above. While high-recall flows naturally tend to be simple, we also observe more complex flows, such as *(negative, negative, positive, negative)* in case of *change-2class-noloops*. Except for the *change-noloops* flow *(positive)*, which does not occur at all in the movie training set, all shown flows are common across domains, achieving a recall of over 2% in most cases. For *2class-noloops*, only two flows are listed, because *(negative, negative)* is the most common flow in both the product and the movie domain. Regarding the distribution of global sentiments, nearly all flows behave similar across domains. The only exception is *(positive, negative, positive)* in *change-2class-noloops*, which never turns out negative in product reviews but never positive in movie reviews.

Altogether, we conclude that sentiment flow is not a fully precise model of web review argumentation, but it proves general with respect to global sentiment. What remains to be checked is the benefit of modeling sentiment flow under uncertainty.

6 Analysis of the Robustness of the Model

Finally, we evaluate how effectively and domain-robustly sentiment flow predicts global sentiment, when local sentiment is not given but classified. To analyze the domain independence of our model, no knowledge about target domains is used.

6.1 Sentiment Analysis Approaches

To classify sentiment flows, we build on the edit distance approach presented in Section 4:

Model Variants As in Section 5, we look at all 16 possible variants of the proposed model. For each variant, we determine all sentiment flows that represent at least 1% of all reviews in a given training set. We learn a mapping from the edit distance between a review's sentiment flow and each of these flows to global sentiment. For robustness, we also combine different model variants.

We compare the accuracy of the model variants to previous approaches evaluated on the given corpora. In addition, we analyze domain robustness based on three baselines, which relate to the three abstraction levels in Figure 2 (cf. Section 4):

Bag-of-Words (b1) The frequencies of all tokens that occur in at least 5% of all training reviews.

Local Sentiment (b2) The frequencies of positive, neutral, and negative local sentiment in a review as well as the first and last local sentiment (analog to Section 4, local sentiment is mapped to $[0, 1]$).

Sentiment Flow Patterns (b3) The Manhattan distances to those sentiment flows obtained by our clustering approach (Wachsmuth et al., 2014a).

All approaches are used as feature types in machine learning (with values normalized to $[0, 1]$).

6.2 Experimental Set-up

We tackle three-class sentiment analysis, which is supposed to be particularly hard due to the fuzzy nature of neutral sentiment (Täckström and McDonald, 2011). Given the three review corpora described in Section 5, we proceed as follows:[7]

Local Sentiment For feature computations, we split all reviews into tokens and sentences. Then, we classify local sentiment with the algorithm of Socher et al. (2013) from Stanford CoreNLP.[8] The algorithm was trained on subjective movie review sentences. We found that its accuracy is limited to around 50% on the given corpora, partly as it tends to misclassify objective sentences. Still, we use it to avoid any adaptation to the domains at hand.

Global Sentiment To determine global sentiment, we perform supervised learning based on the feature types outlined above. In particular, we use the default configuration of the random forest classifier from Weka (Breiman, 2001; Hall et al., 2009) without any parameter optimization.

Experiments Having classified local sentiment in all reviews, we learn a random forest classifier on each feature type and different feature sets for all combinations of training and test domain. To prevent class bias, the training sets are balanced with duplicate oversampling. Since the size of the corpora is limited, we evaluate in-domain accuracy on the whole corpora using 10-fold cross-validation, averaged over five runs. Afterwards, we test the out-of-domain accuracy by applying the learned classifier to the other complete corpora.

6.3 Results on the Domain Robustness

Table 4 lists accuracy results for all domain combinations. In the movie domain, we obtain an overall accuracy of 71.8. On average, we thus succeed over Pang and Lee (2005) who report about 75 on the reviews of Scott Renshaw and 63 on those of Dennis Schwartz. Similarly, we beat all our three-class sentiment analysis results from (Wachsmuth,

[7] Again, see http://www.arguana.com/software for code.
[8] Stanford CoreNLP, http://nlp.stanford.edu/software

Training	Feature types	Product	Hotel	Movie
Product	b1 Bag-of-words	49.0	45.9	32.4
	b2 Local sentiment	51.7	50.4	39.3
	b3 Sentiment flow patterns	46.8	57.5	47.8
	All baseline features b1-3	51.9	58.8	49.8
	v1 change-2class	46.0	46.6	41.3
	v2 change-2class-noloops	48.7	46.9	38.4
	v3 noloops-2class	48.7	50.1	43.6
	v4 2class	52.0	53.8	44.9
	v5 2class-noloops	48.0	50.4	42.4
	All model variants v1-5	50.5	51.3	42.4
	All flows (v1-5 + b3)	50.9	58.2	**51.1**
	All sentiment (v1-5 + b2-3)	50.8	59.7	50.2
	All features (v1-5 + b1-3)	**54.2**	**60.0**	48.7
Hotel	b1 Bag-of-words	37.8	**79.6**	39.8
	b2 Local sentiment	51.4	64.2	51.1
	b3 Sentiment flow patterns	50.7	74.2	51.1
	All baseline features b1-3	54.8	78.9	48.7
	v1 change-2class	43.2	54.3	43.3
	v2 change-2class-noloops	46.6	49.4	45.3
	v3 noloops-2class	49.3	57.4	46.7
	v4 2class	52.4	58.6	51.6
	v5 2class-noloops	46.9	54.0	48.2
	All model variants v1-5	53.4	69.0	**54.7**
	All flows (v1-5 + b3)	53.8	75.5	53.6
	All sentiment (v1-5 + b2-3)	**57.1**	75.6	51.8
	All features (v1-5 + b1-3)	56.4	79.0	53.3
Movie	b1 Bag-of-words	35.0	41.2	64.8
	b2 Local sentiment	43.2	44.2	59.0
	b3 Sentiment flow patterns	42.2	39.5	67.2
	All baseline features b1-3	48.0	50.4	70.5
	v1 change-2class	42.9	44.0	44.8
	v2 change-2class-noloops	40.8	48.1	44.2
	v3 noloops-2class	44.9	46.5	50.7
	v4 2class	–	–	–
	v5 2class-noloops	44.2	44.7	55.9
	All model variants v1-5	44.6	49.7	60.9
	All flows (v1-5 + b3)	47.6	51.9	65.2
	All sentiment (v1-5 + b2-3)	**49.7**	**54.1**	65.9
	All features (v1-5 + b1-3)	48.0	52.3	**71.8**

Table 4. Accuracy of predicting 3-class global sentiment for each combination of training and test domain using the baselines and/or a selection of the 16 evaluated model variants.

2015) in the hotel domain. In the product domain, our approach fails to compete with Täckström and McDonald (2011) who classify the global sentiment of 66.6% of all reviews correctly after training on large-scale product corpora. The small size of the given corpus explains the limited in-domain accuracy in Table 4; even some out-of-domain classifiers perform better on the product reviews. Still, the value 54.2 significantly improves over all baselines under a paired t-test ($p < 5\%$).

As expected, bag-of-words (b1) proves strong in some in-domain tasks—achieving even the best overall accuracy in the hotel domain (79.6)—but it consistently fails out-of-domain. Although less clear, similar observations can be made for b2. This shows that a restriction to the distribution of local sentiment is insufficient to tackle domain dependence. The sentiment flow patterns are comparably effective out-of-domain, but still suffer from the domain change on the evaluated corpora.

For space reasons, we compare the baselines b1 to b3 only to a selection of five of the most effective model variants, v1 to v5. Alone, these variants only occasionally do better than the sentiment flow patterns (b3). However, their combination (v1–5) clearly outperforms b3 in 4 out of 6 out-of-domain experiments. A strong variant is *2class-noloops*, which already proved general in the results from Section 5. In contrast, *2class* (v4) appears controversial. While it turns out being both effective and domain-robust when training in the product and hotel domain, no *2class* sentiment flow represents at least 1% of the movie corpus, emphasizing that more abstraction is required for robustness.

Altogether, the bottom lines of each domain in Table 4 provide clear evidence that our approach improves domain robustness in sentiment analysis: In all cases, the out-of-domain accuracy is best when using our sentiment flow features v1–5. At the same time, our results suggest that very high effectiveness might require more adaptation to the target domain. In this regard, modeling sentiment flow serves as a promising basis to align more effective but domain-dependent features.

7 Conclusion

This paper puts the goal of domain independence in the sentiment analysis of web reviews into the focus. In particular, we hypothesize that an abstract model of the local sentiment flow in a review generally captures the review's overall argumentation regarding global sentiment. In ground-truth data from three domains, we have found clear evidence for our hypothesis, indicating that people write reviews in similar ways across domains.

On this basis, we have presented a novel learning approach, which predicts the global sentiment of a review from the edit distance between the review's sentiment flow and a set of common flows. While we determined common flows with clustering in previous work (Wachsmuth et al., 2014a), instead here we rely on different flow abstractions at the same time. Systematic experiments emphasize that, in this manner, our approach achieves domain robustness without any domain adaptation even when the accuracy of the local sentiment in the flows is limited.

However, our experiments also show that sentiment flow alone does not always suffice to predict global sentiment. In future sentiment analysis approaches, sentiment flows may therefore rather serve as pivot features for domain adaptation.

References

Philippe Besnard and Anthony Hunter. 2008. *Elements of Argumentation*. The MIT Press.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Samuel Brody and Noemie Elhadad. 2010. An Unsupervised Aspect-Sentiment Model for Online Reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms*. MIT Press, third edition.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining on the Web from Information Seeking Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 26–39.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1630–1639.

Rémi Lebret and Ronan Collobert. 2014. Word Embeddings through Hellinger PCA. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–490.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Yi Mao and Guy Lebanon. 2007. Isotonic Conditional Random Fields and Local Sentiment Flow. *Advances in Neural Information Processing Systems*, 19:961–968.

Rohith Menon and Yejin Choi. 2011. Domain Independent Authorship Attribution without Domain Adaptation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing 2011*, pages 309–315.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation Mining. *Artificial Intelligence and Law*, 19(1):1–22.

Diarmuid Ó Séaghdha and Simone Teufel. 2014. Unsupervised Learning of Rhetorical Structure with Un-topic Models. In *Proceedings of he 25th International Conference on Computational Linguistics: Technical Papers*, pages 2–13.

Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Informal Retrieval*, 2(1–2):1–135.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling Essay Organization in Student Essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting Product Features and Opinions from Reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346.

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics*, pages 1118–1127.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Oscar Täckström and Ryan McDonald. 2011. Discovering Fine-grained Sentiment with Latent Variable Structured Prediction Models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, pages 368–374.

Simone Teufel. 2014. Scientific Argumentation Detection as Limited-Domain Intention Recognition. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 101–109.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Handbook of Argumentation Theory*. Springer.

Maria Paz Garcia Villalba and Patrick Saint-Dizier. 2012. Some Facets of Argument Mining for Opinion Analysis. In *Proceedings of the 2012 Conference on Computational Models of Argument*, pages 23–34.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014a. Modeling Review Argumentation for Robust Sentiment Analysis. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 553–564.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014b. A Review Corpus for Argumentation Analysis. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 115–127.

Henning Wachsmuth. 2015. *Pipelines for Ad-hoc Large-Scale Text Mining*. To appear in Lecture Notes in Computer Science. Springer, available at http://is.upb.de/?id=wachsmuth.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 783–792.

Qiong Wu, Songbo Tan, Miyi Duan, and Xueqi Cheng. 2010. A Two-Stage Algorithm for Domain Adaptation with Application to Sentiment Transfer Problems. In *Information Retrieval Technology*, volume 6458 of *Lecture Notes in Computer Science*, pages 443–453. Springer.