Modeling Review Argumentation for Robust Sentiment Analysis

Henning Wachsmuth
Universität Paderborn
s-lab – Software Quality Lab
Paderborn, Germany
henningw@upb.de

Martin Trenkmann, Benno Stein
Bauhaus-Universität Weimar
Webis Group
Weimar, Germany
<1st>.<last>@uni-weimar.de

Gregor Engels
Universität Paderborn
s-lab – Software Quality Lab
Paderborn, Germany
engels@upb.de

Abstract

Most text classification approaches model text at the lexical and syntactic level only, lacking domain robustness and explainability. In tasks like sentiment analysis, such approaches can result in limited effectiveness if the texts to be classified consist of a series of arguments. In this paper, we claim that even a shallow model of the argumentation of a text allows for an effective and more robust classification, while providing intuitive explanations of the classification results. Here, we apply this idea to the supervised prediction of sentiment scores for reviews. We combine existing approaches from sentiment analysis with novel features that compare the overall argumentation structure of the given review text to a learned set of common *sentiment flow patterns*. Our evaluation in two domains demonstrates the benefit of modeling argumentation for text classification in terms of effectiveness and robustness.

1 Introduction

Text classification is a key technique in natural language processing and information retrieval that is applied for several tasks. Standard classification approaches map a text to a vector of lexical and shallow syntactic surface-level features, from which class information is inferred using supervised learning (Manning et al., 2008). Even though the results of such approaches can hardly be explained, they have proven effective for narrow-domain texts with explicit class information (Joachims, 2001; Pang et al., 2002).

However, surface-level features often do not help to classify out-of-domain texts correctly, because they tend to model the domain of the texts and not the classes to be inferred, as we observe in (Wachsmuth and Bujna, 2011) among others. Moreover, they are likely to fail on texts where the class information is implicitly represented by the argumentation of the writer. Such texts are in the focus of popular tasks like authorship attribution, automatic essay grading, and, above all, sentiment analysis. As an example, consider the short hotel review at the top and bottom of Figure 1. It contains more positive than negative statements. Hence, a surface-level analysis would probably classify the review to have a positive overall sentiment polarity. In fact, the argumentation of the review text reveals a clear negative sentiment.

The analysis of argumentation is recently getting more attention (cf. Section 2 for details). With respect to sentiment, related approaches analyze discourse relations (Mukherjee and Bhattacharyya, 2012), identify the different aspects mentioned in a text (Lazaridou et al., 2013), or the like. While these approaches can infer implicit class information from argumentative texts like reviews, they do not address the domain dependency problem of sentiment analysis (Wu et al., 2010). In addition, they still lack explainability, which limits end user acceptance in case of wrong results (Lim and Dey, 2009).

In this paper, we consider the question of how to capture the argumentation of reviews for a domain-robust and explainable text classification. As Figure 1 illustrates, we rely on a shallow model of review argumentation, which represents a text as a sequence of statements that express local sentiment on domain concepts and that are connected by discourse relations. We claim that, by focusing on features that model the abstract argumentation structure of a text, a more robust sentiment analysis can be achieved. At the same time, such an analysis can explain its results based on the underlying model.

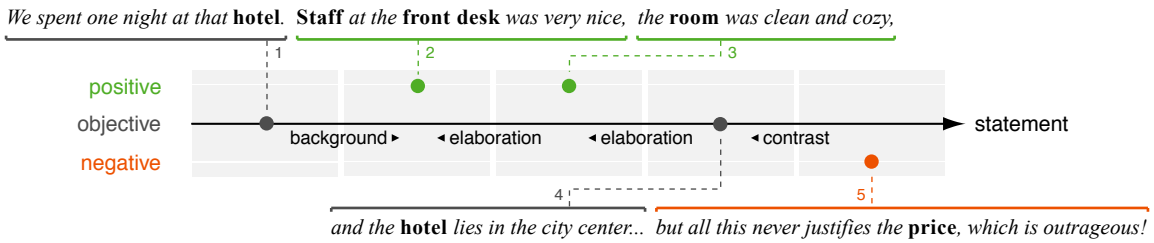

Figure 1: Illustration of our shallow model of review argumentation for a sample review text from the hotel domain. Domain concepts, such as "front desk", are marked in bold. Each circle denotes a statement with local sentiment. The statements are connected by directed discourse relations like "elaboration".

Concretely, here we address the supervised prediction of sentiment scores. To this end, we combine a number of existing argumentation-related features with a novel approach that learns common patterns in sequences of local sentiment through a cluster analysis in order to capture a review's overall argumentation structure. Inspired by explicit semantic analysis (Gabrilovich and Markovitch, 2007), we then compute the similarity of a given review text to each of these *sentiment flow patterns* and we use these similarities as features for sentiment scoring. To explain a predicted score and, hence, to increase user acceptance, both the underlying model and the sentiment flow patterns can be visualized.

We evaluate our approach on reviews of the hotel domain and the movie domain. In comparison to standard baselines, we demonstrate the effectiveness and robustness of modeling argumentation. Our results suggest that especially the sentiment flow patterns learned in one domain generalize well to other domains. Altogether, the contributions of this paper are:

1. A shallow model of review argumentation for text classification that enables a more domain-robust and explainable sentiment analysis (Section 3).
2. A novel feature type named *sentiment flow patterns* that, for the first time, captures the abstract overall argumentation structure of review texts, irrespective of their domain (Section 4).
3. Experimental evidence for the existence of common patterns in the argumentation structure of review texts across domains (Section 5).

2 Related Work

Argumentation plays a key role in human communication and cognition. Its purpose is to provide persuasive information for or against a decision or claim. This involves the identification of facts and warrants justified by a backing or countered by a rebuttal (Toulmin, 1958). Argumentation is studied in various disciplines, such as logic, philosophy, and artificial intelligence. We consider the linguistics perspective, where it is pragmatically viewed as a regulated sequence of speech or text (Walton and Godden, 2006).

In particular, we analyze *monological argumentations* in written text as opposed to dialogical argumentations where participants persuade each other with arguments (Cabrio and Villata, 2012). In terms of text, one of the most obvious forms of monological argumentation can be found in reviews. A review comprises a *positional argumentation*, where an author collates and structures a choice of facts, pros, and cons in order to inform intended recipients about his or her beliefs (Besnard and Hunter, 2008).

According to Mochales and Moens (2011), an *argumentation analysis* targets at "the content of serial arguments, their linguistic structure, the relationship between the preceding and following arguments, recognizing the underlying conceptual beliefs, and understanding within the comprehensive coherence of the specific topic." The authors work on *argumentation mining*, i.e., the detection of different arguments for justifying a conclusion as well as their interactions. Our model of argumentation matches the quoted definition. Similar to the distinction between shallow and deep parsing (Jurafsky and Martin, 2009), our approach can be seen as a *shallow argumentation analysis* in that we consider only the sequence of arguments. This abstraction appears very promising to address text classification.

Unlike *argumentative zoning* (Teufel et al., 2009), which classifies segments of scientific articles according to argumentative functions, we predict the sentiment scores of reviews from a sequence of classified segments. Sentiment scoring is tackled in both computational linguistics (Pang and Lee, 2005) and

information retrieval (Wang et al., 2010). Such kind of sentiment analysis benefits from modeling argumentative discourse (Villalba and Saint-Dizier, 2012). Related works already employ discourse features to detect sentiment polarity. Some rely on complex discourse parsing (Heerschop et al., 2011), whereas others argue that a lightweight approach is more robust for noisy texts (Mukherjee and Bhattacharyya, 2012). We rather follow the latter, but we see discourse only as one part of review argumentation.

In accordance with Lazaridou et al. (2013) who address *aspect-based sentiment analysis*, we additionally analyze the connection of local sentiment to domain concepts and discourse relations. Even more important for us is the *local sentiment flow* in a text. This term was introduced by Mao and Lebanon (2007), who infer a text's global sentiment from its sequence of local (sentence) sentiments, classified with conditional random fields. Their approach converts each sentiment in the sequence to a single feature and learns a mapping from the features to global sentiment. By that, it actually disregards the ordering of local sentiment. In contrast, our sentiment flow patterns measure the similarity between complete sequences of local sentiment. This resembles *explicit semantic analysis* (Gabrilovich and Markovitch, 2007), which classifies texts based on their relatedness to concepts modeled by complete texts.

In (Wachsmuth et al., 2014), we reveal correlations between a review's sentiment score and its local sentiment flow. Similar to Socher et al. (2013), we therefore argue that global sentiment emanates from the composition of local sentiment. The authors model the semantic compositionality of words in given sentences, thus capturing the language of a given domain. Conversely, our sentiment flow patterns focus on the structure of complete texts in order to reduce domain dependency, which is a general problem in text classification (Wu et al., 2010). Among others, existing strategies to tackle this problem align features of the source and the target domain, as we do in (Prettenhofer and Stein, 2010).

Given a vector of features, text classification approaches typically output only a class label (Manning et al., 2008). This renders the understanding and debugging of classification results hard (Kulesza et al., 2011). Instead, our approach explains results by making the argumentation of texts visible. Thereby, we increase intelligibility and, thus, support user acceptance (Lim and Dey, 2009).

3 A Shallow Model of Review Argumentation

This section first sketches our general hypothesis. Then, we present our model of review argumentation.

3.1 Hypothesis behind Modeling Argumentation for Text Classification

Several text classification tasks relate to the argumentation of a text. As an obvious example, *automated essay scoring* explicitly rates argumentative texts, mostly targeting at structural aspects (Dikli, 2006). In *genre identification*, a central concept is the form of texts. Some genre-related tasks address argumentation, e.g. by classifying texts according to their function (Wachsmuth and Bujna, 2011). Criteria in *text quality assessment* often measure structure (Anderka et al., 2012), while *readability* is connected to discourse (Pitler and Nenkova, 2008). *Authorship attribution* profits from argumentation clues like unconsciously used function words (Stamatatos, 2009), and *plagiarism detection*, in the end, aims to check if the argumentation in a fragment of a text refers to the author of the text (Potthast et al., 2013).

We hypothesize that in these and further tasks the class of a text is often decided by the structure of its argumentation rather than by its content, while the content adapts the argumentation to the domain at hand. Following Besnard and Hunter (2008), an argumentation consists of a composition of arguments used to justify a decision or claim. Each argument can be seen as a statement with some evidence. Under our hypothesis, an explicit model of statements and their composition hence supports the identification of domain-independent patterns. Together with the content, the statements enable a fine-grained analysis, while serving as the basis for an explanation. Since the relevant types of statements vary among tasks, we argue that such a model should be task-specific. Below, we investigate reviews on products and services from a sentiment analysis perspective. Because of its positional nature (cf. Section 2), review argumentation makes its arguments explicit, i.e., facts and opinions on different product features and aspects.

3.2 Modeling Review Argumentation for Sentiment Analysis

We consider reviews that comprise a text about some product or service as well as a numerical overall rating. Any other metadata that might be given for reviews is ignored in the following. Our assumption

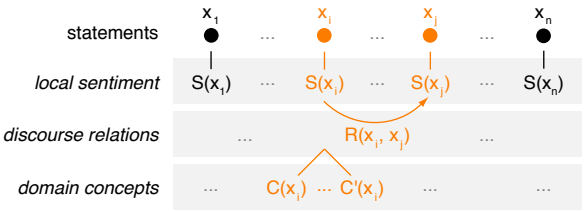

Figure 2: Our shallow model of review argumentation defined by a segmentation into statements and by three functions based on the statements.

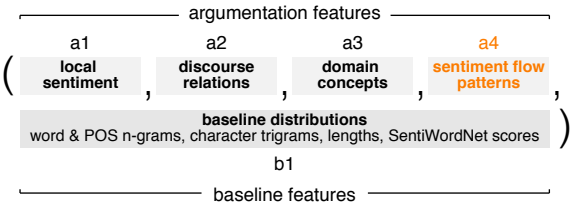

Figure 3: A vector with all five considered feature types including the novel sentiment flow patterns. Each found pattern becomes a single feature in a4.

is that the overall rating denotes a *sentiment score* y from a metric sentiment scale that quantifies the possibly implicit conclusion of the review text in terms of its global sentiment.

Statements To capture a review's argumentation, we model the review's text as a sequence of $n > 0$ statements $x_1, \ldots, x_n$. Here, we define a *statement* x syntactically to be a main clause together with all its subordinate clauses. The notion behind is that, in our experience, such a text segment is usually meaningful on its own while bearing at most one sentiment. Many sentences in reviews comprise series of statements. For instance, the following excerpt from Figure 1 consists of two statements, x_4 and x_5:

x_4: and the hotel lies in the city center... x_5: but all this never justifies the price, which is outrageous!

Based on the set of all statements $\mathbf{X}$, we capture the structure and content of review texts as follows:

Local Sentiment We assume each statement to represent either an objective fact obj, a positive opinion pos, or a negative opinion neg (for a wide applicability, we ignore sentiment intensity). So, there is an unknown function that maps each statement to a local sentiment, e.g. x_4 to obj and x_5 to neg:

$$\textit{local sentiment}: \quad \mathbf{X} \rightarrow \{S(x) \mid S \in \{pos, neg, obj\}\}$$

Discourse Relations As for x_4 and x_5, the composition of statements in a text is, in general, not coincidental. Rather, it implies a structure made up of an ordered choice of statements as well as of a number of directed discourse relations. We define a discourse relation to have some type R of a set of relation types $\mathbf{R}$ and to relate two (typically neighboring) statements, e.g. *contrast(x_5, x_4)* in the example above. The following function hence can be understood as a shallow version of the *rhetorical structure theory* (Mann and Thompson, 1988):

$$\textit{discourse relations}: \quad \mathbf{X} \rightarrow \{R(x_i, x_j) \mid 1 \leq i, j \leq n \; ; \; R \in \mathbf{R}\}$$

Domain Concepts The argumentation structure of a text is bound to the domain at hand through the text's content. In particular, a review text discusses a subset of the *domain concepts* $\mathbf{C}$ that are associated to a product or service, each being referred to in one or more statements. For instance, x_5 discusses the price of the hotel, i.e., *price(x_5)*. We capture the domain concepts in statements as follows:

$$\textit{domain concepts}: \quad \mathbf{X} \rightarrow \{C(x_i) \mid 1 \leq i \leq n \; ; \; C \in \mathbf{C}\}$$

Altogether, our model represents a review text as a sequence of interrelated statements of certain types and content. Figure 2 illustrates the defined functions. An instance of the model is visualized in Figure 1. The model is an abstraction of argumentation, covering some information only implicitly if at all (e.g. lexical or syntactic clues). However, it can be extended by further information, as we do below.

4 Features for Robust Sentiment Analysis and Explanation

We now present different types of features for supervised learning that capture both distributional and structural aspects of review argumentation based on our shallow model. Here, we assume that all information represented in the model is given, but Section 5 analyzes the effects of inferring the information from a text. Figure 3 gives an overview of the vector with all feature types that we consider, including a common set of baseline features (b1). The goal of all argumentation features (a1–a4) is twofold: (1) To enable an effective and robust sentiment analysis. (2) To provide means to explain analysis results.

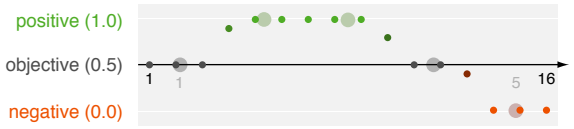

Figure 4: Illustration of a length-normalized version (small circles) of the sample local sentiment flow from Figure 1 (big circles) for length 16.

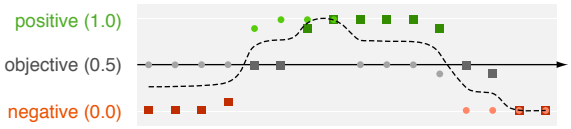

Figure 5: Sketch of the construction of a sentiment flow pattern (dashed curve), here from two sample local sentiment flows (circles and squares).

4.1 Quantification of Distributional Argumentation Aspects

In terms of the distributional aspects of the three functions introduced in Section 3, we combine a selection of ideas from existing sentiment analysis approaches that are related to review argumentation. Some features of the types described in the following are selected only if they occur frequently in a given set of training texts. Thus, the concrete numbers of features vary, as we see in the evaluation in Section 5.

Local Sentiment (a1) In (Wachsmuth et al., 2014), we stress the impact of the distribution of local sentiment. Accordingly, here we determine the frequencies of all types of local sentiment in the given text as well as of series of statements with the same type and of changes from one type to another. Also, we have features that denote the local sentiment at specific positions like the first and last two statements, and we compute the average local sentiment. For the latter, we map pos to 1.0, obj to 0.5, and neg to 0.0.

In addition, we follow Mao and Lebanon (2007) in that we capture the local sentiment flow based on the defined mapping. To preserve the original flows as far as possible, we length-normalize the sequence of values using non-linear interpolation with subsequent sampling. Figure 4 shows an example.

Discourse Relations (a2) We count the occurrences of different discourse relation types from (Mann and Thompson, 1988), e.g. *contrast* or *elaboration* (in Section 5, we distinguish a subset of ten types). To model connections between sentiment and discourse, we do the same for all frequently occurring combinations of discourse relation types and local sentiment of the related statements, e.g. *contrast(pos, neg)* or *contrast(neg, pos)*. By that, we imitate Lazaridou et al. (2013) to some extent.

Domain Concepts (a3) With the same intention, we determine the most frequent domain concepts in the given training set and we compute how often each concept cooccurs with each type of local sentiment. Examples from the sample text in Figure 1 are *hotel(obj)* or *price(neg)*. Moreover, we count the number of different domain concepts as well as the instances of all possibly distinguished types of domain concepts, which would be *product* (like "hotel") and *product feature* (like "price") in the given case.

Types a1–a3 refer to important characteristics of review argumentation. However, none of them captures a review's overall argumentation structure. Even the local sentiment flow in a1 rather measures the impact of local sentiment at different positions. The reason behind is that the flow positions are represented by individual features. Hence, common learning approaches like regression will naturally tend to assign positive weights to all positions, not considering the sentiment flow as a whole.

4.2 Learning of Structural Argumentation Aspects

To capture the impact of the structure of an argumentation, we introduce a novel feature type based on the local sentiment flows of texts only. The idea behind resembles *explicit semantic analysis* (Gabrilovich and Markovitch, 2007) in that every single feature represents the similarity to a complete flow:

Sentiment Flow Patterns (a4) We first construct a set of common *sentiment flow patterns* from a set of known training review texts, where each pattern denotes the average of a set of similar local sentiment flows of normalized length. Given an unknown review text, we then measure the similarity of its normalized local sentiment flow to each constructed pattern. The set of these similarities forms a4.

Figure 5 exemplifies the pattern construction. Our hypothesis behind sentiment flow patterns is that similar local sentiment flows entail similar sentiment scores. Accordingly, flows that construct a pattern should be as similar as possible and flows of different patterns as dissimilar as possible. Therefore, we apply *clustering* (Manning et al., 2008) to partition the flows of all texts from the given training set based on some flow similarity function (in Section 5, we use the manhattan distance). The centroid of each ob-

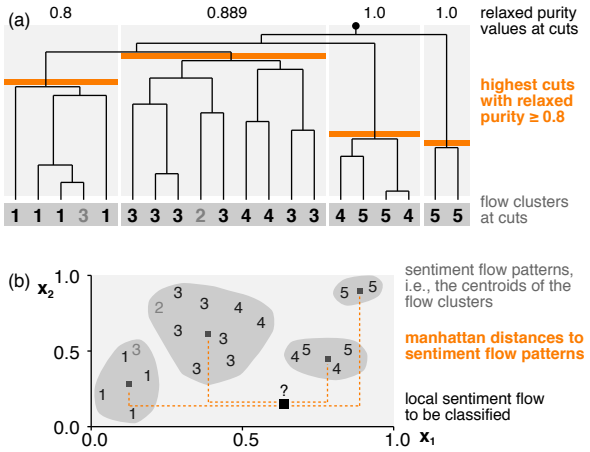

Figure 6: (a) A purity threshold of 0.8 derives four clusters from a hierarchical clustering of 20 local sentiment flows represented by their scores from 1 to 5. (b) 2D plot of computing distances to the sentiment flow patterns, i.e., the clusters' centroids.

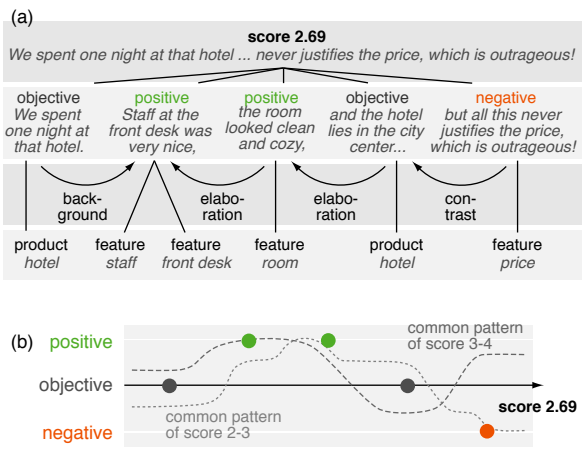

Figure 7: Two possible explanations of scoring the sample text from Figure 1: (a) Graph visualization of our model of review argumentation. (b) Comparison of the local sentiment flow of the text with the two most similar sentiment flow patterns.

tained cluster ω then becomes a sentiment flow pattern. Since we know the sentiment scores associated to the flows in the training set, we can measure the *purity* of a cluster ω, which here denotes the fraction of those flows ω_{y^*} in ω whose score equals the majority score y^* in ω (Manning et al., 2008). The original purity definition, however, assumes exactly one correct score for each flow. Here, this would mean that a flow alone decides a score. Instead, for larger sentiment scales, we propose to relax the purity measure by assuming also the dominant neighbor of the majority score as correct:

$$relaxed\ purity(\omega) \;=\; \big(|\omega_{y^*}| + \max(|\omega_{y^*-1}|,|\omega_{y^*+1}|)\big) \;/\; |\omega|$$

We seek for clusters with a high purity, because such clusters support that similarities between flows and patterns indicate specific sentiment scores. At the same time, the number of clusters should be small in order to achieve a high average cluster size and, thus, a high commonness of the patterns. For this purpose, we rely on *hierarchical clustering*, where we can easily find a flat clustering with a certain number of clusters through cuts at appropriate nodes in the binary tree of the associated hierarchy. Pattern construction profits from compact clusters, suggesting to compute distances between clusters from their *group-average links* (Manning et al., 2008). To minimize the number of clusters, we search for all nodes closest to the tree's root that represent clusters with a purity above some threshold, e.g. 0.8 in the example in Figure 6(a). The centroids of these clusters become sentiment flow patterns, if they are made up of some minimum number of flows.

At the end of the clustering process, we remain with one feature for each constructed sentiment flow pattern. Given a review text to be classified, we then compute its normalized local sentiment flow and we measure the flow's distance to all patterns. Each distance represents one similarity in the feature type a4. Figure 6(b) sketches the computation of the distances, mapped into two dimensions.

4.3 Comparison with Baseline Approaches

In the evaluation below, we compare the feature types a1 to a4 with the well-known sentiment scoring approach of Pang and Lee (2005) in terms of effectiveness. Our focus, however, is the robustness of modeling argumentation structure in contrast to standard text classification features employed in many other approaches, such as n-grams or variations of them (Qu et al., 2010). To this end, we also integrate the following baseline features, most of which model a text at the lexical and syntactic level:

Baseline Distributions (b1) We compute the distributions of all word and part-of-speech unigrams, bigrams, and trigrams as well as of all character trigrams that frequently occur in a given training set. In addition, we determine the length of the given text in different units and some average *SentiWordNet* scores with respect to both the first and the average senses of its words (Baccianella et al., 2010).

4.4 Explanation of Sentiment Scores

We propose a shallow statistical argumentation analysis that learns to predict sentiment scores on a training set of review texts. After prediction, we can directly exploit the information captured in our model as well as the values of the feature types a1–a4 in order to explain the predicted score. Two possible explanations are visualized in Figure 7, while a combination of them is exemplified in Figure 1. We believe that such explanations can increase a user's confidence in statistical analysis results and, hence, the acceptance of corresponding applications. To demonstrate the analysis and explanation of review argumentation, we provide a free-to-use tool and webservice at http://www.arguana.com.

5 Evaluation of Modeling Argumentation for Sentiment Scoring

In this section, we evaluate the effectiveness and domain robustness of modeling argumentation for sentiment scoring with a focus on the sentiment flow patterns. The source code of the evaluation can be found at http://www.arguana.com. Our experiments are based on two English text corpora with reviews from the hotel domain and the movie domain, respectively. In both cases, we leave out the review titles for generality, because our approach targets at arbitrary reviews including those without a title.

Text Corpora On the one hand, we process the *ArguAna TripAdvisor corpus* that we have introduced in (Wachsmuth et al., 2014). This corpus compiles a collection of 2,100 reviews of hotels from seven locations, balanced with respect to their sentiment scores between 1 and 5. All of the reviews' texts are segmented into statements with an average of 14.8 statements per text. Each statement is annotated as an objective fact, a positive, or a negative opinion. Moreover, all mentions of domain concepts are marked as such. The corpus is also available at http://www.arguana.com, free for scientific use. In the experiments, we rely on the provided corpus split with 900 reviews from three hotel locations in the training set, and 600 reviews from two locations in the validation set and test set each.

On the other hand, we use the *Sentiment Scale dataset* (Pang and Lee, 2005) consisting of 5,006 movie reviews that are split into four text corpora according to their authors (*Author a*, *b*, *c*, and *d*). From these, we have discarded eight reviews due to encoding problems. We choose the provided sentiment scale from 0 to 2, so we can logically map the scale of the hotel reviews (1–5) to it for a domain transfer. In particular, scores 1–2 are mapped to 0, 3 to 1, and 4–5 to 2. On average, the movie reviews are much longer with 36.1 statements per text. Since no local sentiment annotations are given, we also process the *subjectivity dataset* (Pang and Lee, 2004) and the *sentence polarity dataset* (Pang and Lee, 2005) in order to develop classifiers for sentence sentiment. Accordingly, we assume each movie review sentence to denote one statement. To directly compare our results to those of Pang and Lee (2005), we perform 10-fold cross-validation separately on the dataset of each single author, averaged over five runs.

Preprocessing For feature computations, we preprocess all texts with a tokenizer, a sentence splitter, and the part-of-speech tagger from (Schmid, 1995). We employ lexicon-based extractors for discourse relations and domain concepts, which aim at a high precision while not being able to recognize unseen instances. The former resembles the lightweight approach of Mukherjee and Bhattacharyya (2012). Primarily, it looks for conjunctions that indicate certain discourse relations, such as "but" or "because". The latter detects exactly those domain concepts that are annotated largely consistently in the training set of the ArguAna TripAdvisor corpus. Thus, it helps only on the hotel reviews. These reviews are segmented into statements with a respective algorithm that comes with the corpus.

For both domains, we have trained linear support vector machines (SVMs) from Chang and Lin (2011) that classify the subjectivity of each statement (opinion or fact) and the polarity of each opinion (positive or negative). They use 1k to 2k features of different types: word and part-of-speech unigrams, character trigrams, SentiWordNet scores (Baccianella et al., 2010), and some special features like the first word of a statement or its position in the text. On the test set of the hotel domain, the classifiers have an accuracy of 78.1% for subjectivity and of 80.4% for polarity. In the movie domain, we achieve a subjectivity accuracy of 91.1%, but a polarity accuracy of only 73.8% (measured through 10-fold cross-validation).

Feature Computation We determine one distinct feature set for each evaluated text corpus made up of the feature types presented in Section 4. Where necessary, we divide the computed feature values by

the length of the text (in tokens or statements, as appropriate), in order to ensure that all feature values always lie between 0 and 1.

Local sentiment flows are normalized to length 30 in case of the hotel reviews and to length 60 in case of the movie reviews, which allows us to represent most of the original flows without loss. Altogether, feature type a1 sums up to 50 and 80 features, respectively. For a2, a3, and b1, we consider only those features whose frequency in the training texts exceeds some specified threshold. For instance, a word unigram is taken into account within b1 only if it occurs in at least 5% of the hotel reviews or 10% of the movie reviews, respectively. As a result, the number of evaluated features varies depending on the processed text corpus. Concretely, we obtain 64 to 78 features for discourse relations (a2), 78 to 114 for domain concepts (a3), and 1026 to 2071 baseline features (b1). More details are given in the instruction and configuration files that come with the provided source code.

To construct sentiment flow patterns (a4), we have developed an agglomerative hierarchical clusterer that implements the approach from Section 4.2. After some tests with different settings, we decided to measure flow and cluster similarity using group-average link clustering based on the manhattan distance between the length-normalized local sentiment flows. For the hierarchy tree cuts, we use a purity threshold of 0.8, where we take the relaxed purity for the sentiment scale 1–5 of the hotel reviews, but the original purity for the movie reviews (because of the limited scale from 0 to 2). All centroids of clusters with at least three flows become a sentiment flow pattern, resulting in 16 to 86 features in a4.

Sentiment Scoring On the hotel reviews, we compute the root mean squared error of linear sentiment score regression trained using stochastic gradient descent (SGD) from *Weka 3.7.5* (Hall et al., 2009). Both the regularization parameter and the learning rate of SGD are set to 10^{-5}, whereas we determine the epochs parameter of SGD on the validation set. Then, we measure the error on the test set.

For the comparison to (Pang and Lee, 2005), we predict the scores of the movie reviews using classification, which additionally stresses the domain change. In particular, we measure the accuracy of a linear 1-vs.-1 multi-class SVM with probability estimates and normalization. While we optimize the cost parameter of the SVMs in the in-domain task, we rely on the default value (1.0) for the domain transfer.

5.1 Effectiveness of Modeling Argumentation

First, we measure the theoretically possible scoring effectiveness of all feature types within one domain. To this end, we compare the feature types based on the ground-truth annotations of the ArguAna TripAdvisor corpus. The column *Corpus* of Table 1 lists the resulting root mean squared errors. As can be seen, all argumentation feature types clearly outperform the baseline distributions (b1) and improve strongly over random guessing. The distributional local sentiment (a1) does best with an error of 0.77, whereas the domain concepts perform worst among a1 to a4. Still, they result in an 0.12 lower root mean squared error than the baseline distributions (b1). Overall, the lowest observed error is 0.75, achieved by the SVM with all features as well as by two subsets of the argumentation features alone.

In practice, no ground-truth annotations are given, so we need to create annotations in the review texts ourselves using the preprocessing described above. This in turn changes the feature set and the respective values of the argumentation features. The third column of Table 1 (*Self*) shows that such a resort to self-created annotations leads to a root mean squared error increase of 0.14 to 0.22 for the types a1 to a4. Nevertheless, the argumentation features succeed over the baseline distributions with 0.94 as opposed to 1.11, which demonstrates the effectiveness of modeling the argumentation of hotel reviews.

5.2 Robustness of Modeling Argumentation Structure

We hypothesize that the developed structure-based argumentation features are robust against domain transfer to a wide extent. To investigate this, we classify sentiment scores using SVMs based either on all or on one single feature type (except for a3, for lack of movie domain concept extractors) in two tasks on the four movie datasets: (1) with training in the movie domain (through 10-fold cross-validation), and (2) with training out-of-domain on the hotel review training set.

Figure 8 contrasts the accuracy results for the two tasks and compares them to the best SVM approach of Pang and Lee (2005), i.e., *ova* (open squares). In the in-domain task, our SVM based on all feature types (black squares) is significantly better than *ova* on one dataset (*Author a*) and a little worse on

Feature type		Corpus	Self
none	Random guessing	1.41	1.41
a1	Local sentiment	0.77	0.99
a2	Discourse relations	0.84	1.01
a3	Domain concepts	0.99	1.13
a4	Sentiment flow patterns	0.86	1.07
b1	Baseline distributions	1.11	1.11
a1–a4	Argumentation features	0.76	0.94
a2, a3, a4	w/o local sentiment	0.79	0.99
a1, a3, a4	w/o discourse relations	0.76	0.97
a1, a2, a4	w/o domain concepts	**0.75**	0.95
a1, a2, a3	w/o sentiment flow patterns	**0.75**	0.95
all	All features	**0.75**	**0.93**

Table 1: Root mean squared error of sentiment score regression on the hotel review test set for all evaluated features types and for different combinations of these types. Features are computed based on ground-truth annotations (*Corpus*) or based on self-created annotations (*Self*).

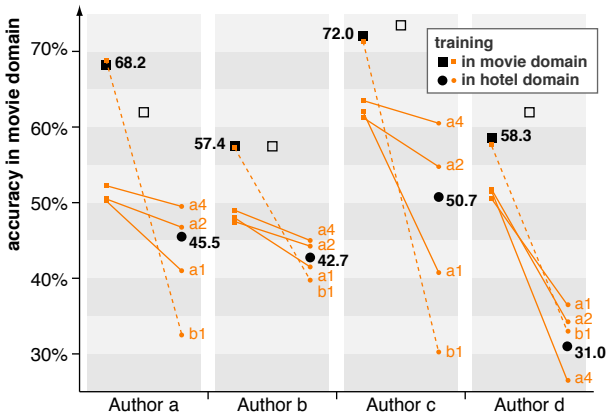

Figure 8: Sentiment scoring accuracy of the SVM based on feature type a1, a2, a4, and b1 (black icons) and of each SVM based on one of these types (orange icons) on the four movie datasets, when trained on movie reviews (squares) or on hotel reviews (circles). Open squares: *ova* from (Pang and Lee, 2005).

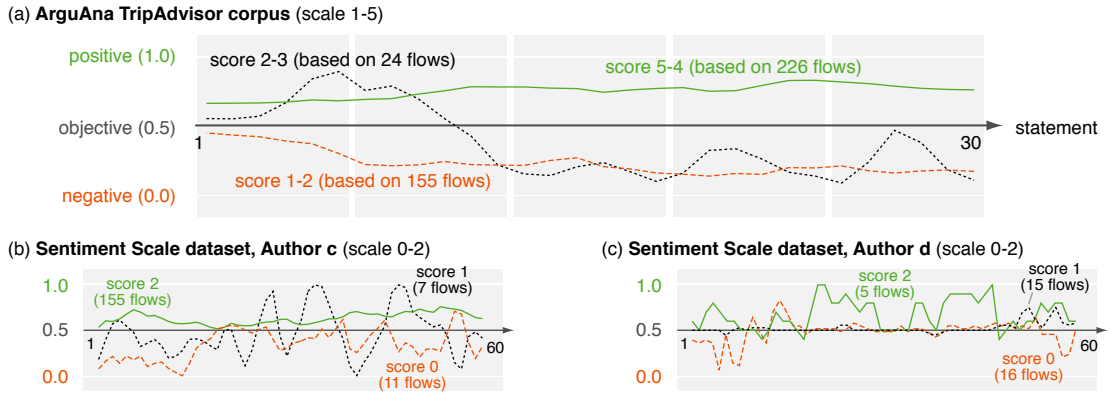

Figure 9: (a) The three most common sentiment flow patterns in the training set of the ArguAna TripAdvisor corpus, labeled with their associated sentiment scores. (b–c) The according sentiment flow pattern for each possible score of the texts of *Author c* and *Author d* in the Sentiment Scale dataset, respectively.

two other datasets (*Author c* and *Author d*). On all four datasets, a4 classifies sentiment scores more accurately than both a1 and a2, but none of the argumentation feature types can compete with the baseline distributions (b1). We suppose that the reason behind mainly lies in the limited effectiveness of our opinion polarity classifier, which reduces the impact of all features that rely on statement sentiment.

Conversely, b1 fails completely in the out-of-domain task (from squares to circles) with accuracy drops of up to 41% (on *Author c*). This indicates a large *covariate shift* (Shimodaira, 2000) in the distribution of the baseline features. In contrast, a1, a2, and a4 suffer much less from the domain transfer. Especially the accuracy of the sentiment flow patterns (a4) remains stable on three of the four datasets and, hence, provides strong support for our hypothesis. In case of *Author c*, the SVM based on a4 alone even achieves a significantly higher accuracy than the SVM based on all features (60.5% as opposed to 50.7%), thus offering evidence for the decisiveness of the structure of an argumentation. Only on *Author d*, all four evaluated feature types similarly fail when trained on hotel reviews with a4 being the worst. Apparently, the argumentation structure in the texts of *Author d* differs from the others, which is reflected by the found sentiment flow patterns and which we therefore finally analyze.

5.3 Insights into Sentiment Flow Patterns

In Figure 9, we plot the three most common sentiment flow patterns in the training set of the ArguAna TripAdvisor corpus (with self-created annotations) as well as the respective patterns in the movie reviews

of *Author c* and *Author d* for each possible sentiment score. In total, we found 38 sentiment flow patterns in the hotel reviews, meaning that a4 consists of 38 features in this case. As depicted in Figure 9(a), they are constructed from the local sentiment flows of up to 226 texts. One of the 75 patterns of *Author c* results from 155 flows, whereas each of the 41 patterns of *Author d* represents at most 16 flows.

With respect to the depicted sentiment flow patterns, the movie reviews show less clear sentiment but more changes of local sentiment than the hotel reviews. While there appears to be a certain similarity in the overall argumentation structure between the hotel reviews and the movie reviews of *Author c*, two of the three patterns of *Author d* contain only little clear sentiment at all, especially in the middle parts. The disparity of the *Author d* dataset is additionally emphasized by the different proportions of opinions in the evaluated text corpora. In particular, 79.7% of all statements in the ArguAna TripAdvisor corpus are opinions, but only 36.5% of the sentences of *Author d* are classified as subjective. The proportions of the three other movie datasets at least range between 58.4% and 66.5%. These numbers also serve as a general explanation for the limited accuracy of a1, a2, and a4 in the movie domain.

A solution to achieve higher accuracy and to further improve the domain robustness of the structure-based argumentation features might be to construct flow patterns from the subjective statements or from the changes of local sentiments only, which we leave for future work. Here, we conclude that our novel feature type a4 does not yet solve the domain dependency problem, but it still defines a promising step towards a more domain-robust sentiment analysis.

6 Conclusion

Text classification tasks like sentiment analysis are domain-dependent and tend to be hard on texts that comprise an involved argumentation, such as reviews. To classify the sentiment scores of reviews, we model a review's text as a composition of local sentiment, discourse relations, and domain concepts. Based on this shallow model of argumentation, we combine existing sentiment analysis approaches with novel features that capture the abstract overall argumentation structure of reviews irrespective of their domain and their linguistic style. In particular, we learn common sequences of local sentiment in reviews through clustering in order to then compare a given review to each of these learned *sentiment flow patterns*. Our evaluation on hotel and movie reviews suggests that the sentiment flow patterns generalize well across domains and it indicates the effectiveness of modeling argumentation. In addition, both the patterns and our model help to explain sentiment scoring results, as exemplified.

Due to errors in the preprocessing of texts, some obtained effectiveness gains are rather small, though. In the future, we seek to develop features that are less affected from preprocessing. A promising variation in this respect is e.g. to learn patterns based on the changes of local sentiment only. Also, we plan to analyze common sequences of discourse relations in order to capture the argumentation structure of a text in an even more domain- and language-independent manner. By that, we contribute to the general research on robust and explainable text classification. As outlined in Section 3, many text classification tasks can profit from modeling argumentation. For this purpose, other types of statements, relations, and domain concepts will be needed as well as, in some cases, a deeper argumentation analysis.

Acknowledgments

This work was funded by the German Federal Ministry of Education and Research (BMBF) under contract number 01IS11016A as part of the project "ArguAna", http://www.arguana.com.

References

Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval*, pages 981–990.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2200–2204.

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

Elena Cabrio and Serena Villata. 2012. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 208–212.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Semire Dikli. 2006. An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment*, 5(1).

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1606–1611.

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.

Bas Heerschop, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. 2011. Polarity Analysis of Texts Using Discourse Structure. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1061–1070.

Thorsten Joachims. 2001. A Statistical Learning Model of Text Classification for Support Vector Machines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 128–136.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2nd edition.

Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M. Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented End-user Debugging of Naive Bayes Text Classification. *ACM Transactions on Interactive Intelligent Systems*, 1(1):2:1–2:31.

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.

Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, pages 195–204.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.

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.

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

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1847–1864.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity. In *Proceedings of 42th Annual Meeting of the Association for Computational Linguistics*, pages 271–278.

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 of the Association for Computational Linguistics*, pages 115–124.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 79–86.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein. 2013. Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1212–1221.

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.

Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 913–921.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Hidetoshi Shimodaira. 2000. Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

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.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of American Society for Information Science and Technology*, 60(3):538–556.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards Discipline-independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

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 and Kathrin Bujna. 2011. Back to the Roots of Genres: Text Classification by Language Function. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 632–640.

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

Douglas Walton and David M. Godden, 2006. *Considering Pragma-Dialectics*, chapter The Impact of Argumentation on Artificial Intelligence, pages 287–299. Erlbaum.

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.