# Subsentential Sentiment on a Shoestring: A Crosslingual Analysis of Compositional Classification

**Michael Haas** and **Yannick Versley**
Institute for Computational Linguistics
University of Heidelberg
{haas,versley}@cl.uni-heidelberg.de

## Abstract

Sentiment analysis has undergone a shift from *document-level* analysis, where labels expresses the sentiment of a whole document or whole sentence, to subsentential approaches, which assess the contribution of individual phrases, in particular including the *composition* of sentiment terms and phrases such as negators and intensifiers.

Starting from a small sentiment treebank modeled after the Stanford Sentiment Treebank of Socher et al. (2013), we investigate suitable methods to perform compositional sentiment classification for German in a data-scarce setting, harnessing cross-lingual methods as well as existing general-domain lexical resources.

## 1 Introduction

In sentiment classification, we find a general tendency from document-level classification towards more fine-grained approaches that yield a more detailed appraisal of the judgement performed in the text - in particular, using composition over syntactic structure to get a more detailed approach over phrases.

For English movie reviews, work using the Stanford Sentiment Treebank (SSTb) has shown that such subsentential sentiment information can yield approaches with both very high accuracy (Socher et al., 2013; Dong et al., 2014; Hall et al., 2014) and precise information about the role of each phrase – information which can subsequently used for extracting or summarizing the sentiment expressed in the text.

The effort for creating a sentiment treebank such as the SSTb, however, seems prohibitive if we wanted to create such a resource for each pair of relevant domain and language: Compared to document-level annotations for sentiment, which are easy to come by (e.g., star ratings), annotating individual syntactic phrases requires considerable effort.

The main focus of this paper is the question if and how it is possible to reach sensible performance for compositional sentiment classification when we only have limited resources to spend on an in-language, in-domain sentiment treebank. For this goal, we use a new resource, the *Heidelberg Sentiment Treebank* (HeiST), which is a German-language counterpart to the Stanford Sentiment Treebank in the sense that it makes explicit the composition of sentiment expression over syntactic phrases. Our experiments on HeiST provide a direct comparison of different techniques for harnessing cross-lingual, cross-domain, or cross-task information, and are the first of this kind to specifically target *compositional* sentiment analysis.

Figure 1 (next page) shows a schematic overview of the experiments: beyond supervised baseline experiments using SVM classification and a supervised RNTN model (section 3), we evaluated cross-lingual projection (section 4), lexicon-based approaches (section 5), as well as semi-supervised approaches based on word clusters (section 6).

## 2 Related Work

The starting point for our research is the idea that the sentiment of larger stretches of text can be calculated through composition over smaller stretches of text, which was investigated in a learning framework by both Yessenalina and Cardie (2011) and Socher et al. (2011, 2012), both learning in a compositional

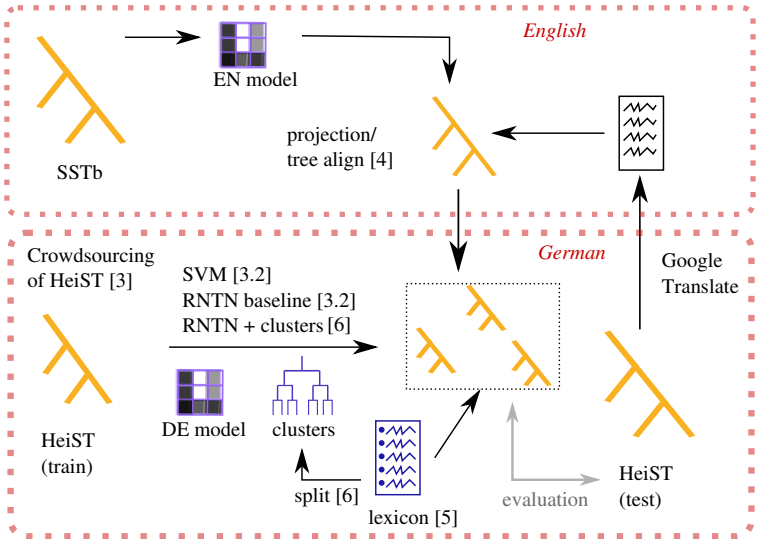



Figure 1: Plan to the experiments described in this paper

fashion from datasets that only have document-level sentiment annotation.

On the same dataset as Socher et al., Wang and Manning (2012) later demonstrated that unigram and bigram features in an SVM-based classification framework can reach a greater accuracy than the earlier recursive neural network approach of Socher et al. (2011, 2012), which calls into question the assumption that sentiment composition can be learned purely from sentence-level annotations.

**Compositionality through Tensors** In subsequent work, Socher et al. (2013) introduce the Stanford Sentiment Treebank, which contains detailed annotations of sentiment values for individual syntactic phrases in a binarized tree, and an approach based on recursive neural tensor networks (RNTN) which yields significant improvements over the earlier approaches using token-level features.

The RNTN represents the contribution of individual nodes as vectors of reals and achieves its precision by using a tensor $V^{[1:d]} \in \mathbb{R}^{2d \times 2d \times d}$ as well as a matrix $W \in \mathbb{R}^{2d \times d}$ to capture second-order dependencies between the two children of a node in the tree (with vectors $a$, $b$), yielding first a vector $h$ by

$$h_i = \begin{bmatrix} a \\ b \end{bmatrix}^T V^{[i]} \begin{bmatrix} a \\ b \end{bmatrix} + W_i \begin{bmatrix} a \\ b \end{bmatrix}$$

then using a monotonic nonlinear function on $h$ (here: *tanh*) to yield the vector for this node. The sentiment label of a node is then gained by multiplying these *hidden vectors* by a matrix $W_s$, yielding a five-dimensional vector with the classification. Using hidden vectors for each node and capturing second-order interaction between the two child vectors $a$ and $b$, the RNTN model achieves descriptive power greater than that of TreeCRFs (Nakagawa et al., 2010), and similar to latent-variable models that have been very successful in syntactic parsing (Petrov et al., 2006).

In later work, Zhu et al. (2014) show that the RNTN's lexicalized modeling of negators and their behaviour leads to increased descriptive power of the model, which results in an improved treatment of negation. Dong et al. (2014) introduce an approach that chooses between multiple composition tensors (AdaMC-RNTN), which yields further gains with respect to RNTN performance.

In contrast to the lexicalized and high-dimensional RNTN model, there are several lines of work that attempt to work in a more data-scarce setting.

**Lexicon-based approaches** The classical approach for performing sentiment classification in a setting where training data is sparse can be seen in the SO-CAL approach of Taboada et al. (2011): Using a manually curated dictionary with sentiment values for multiple parts of speech, and a set of heuristics that predict how intensifiers, nega-

tors/shifters as well as nonveridical moods affect the sentiment of a phrase, they show that it is possible to reach good results across different domains.

Choi and Cardie (2009) show that it is possible to adapt an existing general-domain sentiment lexicon to a specific domain using an approach that optimizes a joint objective of classification loss, sparsity of the changes made to the lexicon, and ambiguity of lexicon entries. Their approach yields appreciable gains over the general-domain lexicon, both with CRF-based machine learning classification and with a simpler "vote & flip" algorithm that is based on majority voting and negators.

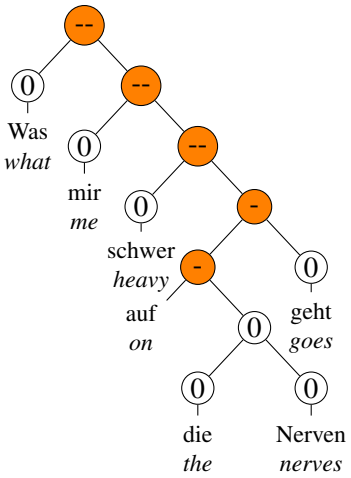
**Crosslingual Sentiment Analysis** involves the usage of a dataset in one language to perform sentiment analysis in another language; in their work, Banea et al. (2013) show that translating text in the target language to the source language and applying a well-tuned sentiment classification system works better than either translating the training corpus or the lexicon used by the system.

In research by other groups, Wan (2009) advocates a bootstrapping approach that combines source-side and target-side features in one classifier; Duh et al. (2011) note that crosslingual sentiment analysis techniques always incur a loss due to the shift in language from the source language texts to the target language even though the general domain is the same. Popat et al. (2013) argue that full machine translation is not useful for resource-scarce languages, and propose to use cross-lingual clustering both to improve the generalization capability within a single language as well as for crosslingual projection, which works better than machine translation with the English-Hindi language pair.

It should be noted that most of the work presented in the last two paragraph works with document-level sentiment, or (in the case of Choi and Cardie) with shallower annotations, and offers additional challenges in the case of sentiment composition.

## 3 Low-Budget Treebanking for Sentiment

For both supervised training and for evaluation, we created a German dataset that is close in domain to the Stanford Sentiment treebank (Socher et al., 2013), covering opinionated sentences from movie reviews with phrase-level sentiment annotations.




"*What really gets on my nerves . . .*"

Figure 2: A multiword expression in HeiST

The original Stanford Sentiment Treebank is based on the dataset of Pang and Lee (2005), which includes 10,662 sentences from excerpts of movie reviews published on `rottentomatoes.com`. It should be noted that these excerpts are much more likely to express an opinion than general text or even the main body of a movie review since they contain precisely a summary of the opinion.

In order to match both domain and role of these sentences most precisely, we collected creative-commons-licensed reviews from a German movie review site, `filmrezensionen.de`, and used only the summary part of these documents, yielding 1184 sentences, for which we crowdsourced annotation for each individual phrase in the binary tree (see Figure 2 for an example tree fragment).

For the purpose of getting binary phrase trees, sentences were processed with the Berkeley Parser (Petrov et al., 2006), NP nodes were added inside PPs (Samuelsson and Volk, 2004) and the resulting parse trees binarized using the head table in CoreNLP (Rafferty and Manning, 2008), yielding 14,321 unique phrases.

Annotation was outsourced via the CrowdFlower service, which collects three judgements for each phrase and computes an end result through voting, using unambiguous test items (which we composed from strongly positive or strongly negative adjective-noun combinations) to filter out annotators lacking the requisite understanding of German.

| *Experiment A: Features, Confidence* | Prec | Recl |
|---|---|---|
| SentiWS | 0.882 | 0.959 |
| SentiWS+Regression | 0.894 | 0.967 |
| SentiWS+Regression, @50% | 0.935 | 0.985 |
| SentiWS+Regression+POS | 0.912 | 0.960 |
| SentiWS+Regression+POS, @50% | 0.978 | 0.997 |
| *Experiment B: Classifier (@50%)* | Prec | Recl |
| Linear SVM | 0.975 | 0.980 |
| Random Forest | 0.984 | 0.992 |
| Gradient Boosting | 0.978 | 0.997 |

Table 1: Filtering out objective phrases

The HeiST treebank, as well as the code used in these experiments, are available for research purposes.[1]

### 3.1 Selecting Subjective Phrases

One possible approach to reduce annotation effort would be to annotate only those phrases that a classification model deems to have sentiment content in the first place. As a more extreme example of such an approach, consider the MLSA sentiment dataset for German, where 270 sentences were selected that already contained two words from an existing sentiment lexicon (Clematide et al., 2012), with the goal of getting sentences with interesting interactions between sentiment words. Given the potential benefits (getting more data for the same annotation effort), an approach that filters out non-interesting (confidently objective) phrases would be highly appealing.

For the pre-classification experiment, we used cross-validation on 20 to assess the potential impact of strategies for saving. For the corresponding classifier, we used features from a German general-domain sentiment lexicon, a regression model for document-level sentiment (see section 5.2), as well as part-of-speech tag features in a gradient boosting classifier. As seen in table 1, the sentiment lexicon, especially in conjunction with the regression model and a POS-based filter, would allow to detect uninteresting (objective) phrases with high accuracy. We limit ourselves to the 50% of most confident classifications, and as a measure of caution, the filter is bypassed for any phrase that contains a word in one of several sentiment dictionaries (see section 5). The classifier has a precision of 96.5% for objective phrases while catching about 66.7% of all objective nodes. While this would correspond to substantial savings (about a quarter of all nodes would be assigned the "neutral" label and not annotated), we would also lose a fraction of non-neutral phrase and introduce an unwanted bias (towards lexicon-based resources) into our dataset.

[1] http://www.cl.uni-heidelberg.de/~versley/HeiST/

| System | Node Acc. | Root Acc. |
|---|---|---|
| **HeiST, only pos+neg sentences** | | |
| *supervised* | | |
| RNTN (tuned) | 0.776 | 0.687 |
| SVM (unigrams, coarse) | **0.850**$^{**}$ | **0.774**$^{**}$ |
| SVM (unigrams, fine) | 0.835$^{**}$ | 0.735 |
| *cross-lingual* | | |
| CLSA (simple feat.) | 0.823$^{**}$ | 0.737 |
| CLSA (complex feat.) | 0.810$^{**}$ | 0.738$^{*}$ |
| **Comparison: HeiST, all sentences** | | |
| *supervised* | | |
| RNTN (tuned) | 0.803 | 0.703 |

*/**: significantly better than RNTN ($p < 0.05$ / $p < 0.005$)

Table 2: HeiST baseline, cross-lingual projection, SVM.

| System | Node Acc. | Root Acc. |
|---|---|---|
| **Comparison: SSTb sample, pos+neg sentences** | | |
| *lexicon-based* | | |
| General Inquirer | 0.824 | 0.715 |
| SubjectivityClues | 0.820 | 0.695 |
| *supervised* | | |
| RNTN, 500 sent. | 0.704 | 0.526 |
| RNTN, 1000 sent. | 0.738 | 0.539 |
| RNTN, 1500 sent. | 0.756 | 0.569 |
| SVM, 500 sent. | 0.803 | 0.652 |
| SVM, 1000 sent. | 0.814 | 0.675 |
| SVM, 1500 sent. | 0.823 | 0.683 |
| RNTN, 6920 sent.[a] | 0.876 | 0.854 |
| SVM, 6920 sent.[a] | 0.846 | 0.794 |

[a]: published figures from Socher et al. (2013)

Table 3: Comparison figures on subsets of the Stanford Sentiment Treebank

### 3.2 Baseline results

We use the existing RNTN implementation of Socher et al. (2013) to train and test supervised learning for sentiment composition, using cross-validation. For parameter tuning, we varied the number of vector dimensions as well as the size of the minibatches used in training, and found that the resulting classifier yields very sensible results compared to a similarly-sized sample from SSTb (see

Tables 2 and 3). We evaluate our results as in Socher et al. (2013): we consider the recall of positive and negative nodes while ignoring both neutral nodes and the difference between positive (+) and strongly positive (++) or between negative (-) and strongly negative (--) nodes, respectively. Socher et al. remove sentences with neutral overall sentiment in training as well as in testing, which seems to worsen the RNTN performance on our dataset (see Table 2), although other methods seem to be less affected by it. For comparability reasons, all other reported figures are based on Socher's non-neutral-sentences-only setting. In comparison results on SSTb (see Table 3), classification experiments from the English data also show poor results for the RNTN classifier at small data sizes, in parallel with anecdotal evidence on recurrent neural networks having trouble with small dataset sizes.[2]

## 4 Crosslingual Projection for Compositional Sentiment

Our crosslingual approach follows Banea et al. (2013) in assuming that machine translation of the target documents to the source language, then applying a source-language sentiment analysis, and finally projecting the result back to the target side will yield usable sentiment classification. In difference to previous approaches for cross-lingual sentiment analysis, however, our annotation transfer concerns not just analysis results for the complete sentence, but for individual syntactic nodes.

After translating the target-language trees using the Google Translate API, we parsed the sentences using the English model of the Stanford parser, and applied the RNTN model of Socher et al. (2013) trained on the English Stanford Sentiment Treebank, yielding a labeling for each syntactic node with a sentiment value. We then performed word alignment using the PostCAT word aligner (Ganchev et al., 2008) with a model trained on the OPUS version of the EuroParl corpus (Tiedemann, 2012), and alignment of syntactic nodes using the `Lingua::Align` toolbox for tree alignment Tiedemann (2010) with a model trained on the Smultron parallel treebank (Volk et al., 2010).

[2]Alec Radford (2015): *General Sequence Learning using Recurrent Neural Nets*, https://indico.io/blog/general-sequence-learning-using-recurrent-neural-nets/

| System | Node Acc. | Root Acc. |
|---|---|---|
| *Vote-only* | | |
| Klenner *et al.* | 0.769 | 0.646 |
| GermanPolarityClues | **0.815** | 0.648 |
| SentiWS | **0.815** | **0.711** |
| SentiMerge [0.0] | 0.660 | 0.577 |
| SentiMerge [0.23] | 0.718 | 0.604 |
| SentiMerge [0.4] | 0.724 | 0.604 |
| Amazon+Lasso | 0.499 | 0.426 |
| *Vote-and-flip* | | |
| Klenner *et al.* | 0.780 | 0.646 |
| GermanPolarityClues | 0.802 | **0.680** |
| SentiWS | **0.807** | 0.665 |
| SentiMerge [0.0] | 0.653 | 0.582 |
| SentiMerge [0.23] | 0.717 | 0.607 |
| SentiMerge [0.4] | 0.723 | 0.603 |
| Amazon+Lasso | 0.471 | 0.413 |

Table 4: Lexicon-based phrase labeling

Using the word alignment and our `Lingua::Align` model, we are able to map 98.6% of the target-language nodes to a corresponding node on the source (English) side, whereas the remaining nodes are assigned the same sentiment label as the root. As can be seen in table 2, a model that uses simpler features for `Lingua::Align` works less well than the full feature model. Considering that the RNTN on the Stanford Sentiment Treebank reaches 87.6% node accuracy and 85.4% root accuracy, we see that the crosslingual projection step induces a loss in accuracy, but still performs well in comparison to the approaches that use the HeiST training data.

## 5 Lexicon-based Approaches

Considering that the size of HeiST creates a sparse data problem for the RNTN learner, it is natural to ask whether we can improve the generalization capabilities of the system by either using a less-supervised approach or by generalizing over individual word forms to alleviate the sparse data problem.

### 5.1 General-domain lexicon

Several general-domain sentiment lexicons exist for German, including those of Klenner et al. (2009), Waltinger (2010a), Remus et al. (2010), and Emerson and Declerck (2014).

Klenner et al. (2009) created their polarity lexicon by semiautomatic extension of an existing one: starting from a set of 2866 adjective seeds, they looked for adjectives that often co-occur in coordinations with known sentiment-bearing adjectives, which were added to the lexicon after a manual filtering step. The current version of Klenner et al.'s PolArt lexicon also contains other parts of speech, and a list of *shifters* and *intensifiers* that interact with subjective terms.

The GermanPolarityClues lexicon of Waltinger (2010a) combines translation from English lexicons with a semi-automatic approach for merging and manually correcting lexicon entries.

The SentiWS lexicon (Remus et al., 2010) contains translations of the English General Inquirer (Stone et al., 1966), which have been translated via Google Translate, as well as a small number of terms that were mined from positive and negative product reviews, expanded using a collocation dictionary.

Finally, the SentiMerge lexicon (Emerson and Declerck, 2014) has been constructed as a Bayesian combination (i.e., averaging with imputation for missing entries) of the three resources above together with the *German SentiSpin* resource of Waltinger (2010b), which contains automatic (dictionary-based) translations of the SentiSpin lexicon of Tamura et al. (2005).

We tested all lexicons using two approaches: In the **vote-only** approach, the sentiment of a phrase is determined by the sum of the scores of the words in that phrase as they are assigned in the sentiment lexicon. In the **vote-and-flip** approach, we consider the average of the sentiment terms, but invert the sentiment value whenever a term from the *shifter* category of Klenner et al.'s lexicon is found within the yield of the node. A similar strategy was used in many papers on sentiment composition, usually with a performance rather close to the best system (see e.g. the CompoMC baseline in Choi and Cardie, 2008, or the Vote-and-Flip baseline in Choi and Cardie, 2009).

### 5.2 Near-domain lexicon construction

While the *filmrezensionen* web site offers a good number of reviews, the final collection is rather small. To complement our small in-domain dataset we use the most common way of getting text with document-level annotations, namely customer-written reviews from the movies section of `amazon.de` web site.

Perhaps expectedly, customer reviews do not focus exclusively on the film and its performance. Rather, it often occurs that customer reviews include a discussion of the physical (or other) medium that the film came on:[3]

(1) *I am with Lovefilm (now Prime) and tried to stream the series. Terrible! Always [issues with] loading time and loss of the stream. It seems that Amazon hasn't come to terms with the technology yet.*

Other reviews on Amazon match our domain fairly well, as in the following:

(2) *If this is truly a sequel to "Speed", it only shows in the second hour of the film. It's only then that deBont shows why he would be an action [film] specialist. Admittedly, even then we don't get the same tension as in the predecessor, but in any case it's better than the first hour of the film.*

While we found that a small quantity of data (20+20 hand-classified sentences) together with a 300-class LDA representation was sufficient to reach 100% accuracy in separating content-related versus media-related text, we found that filtering out the irrelevant texts made no difference for the mean square error, in sharp contrast to L1/Lasso regularization, which allows to learn a sparse lexicon.

## 6 Variants of the RNTN Model

While the RNTN model certainly performs well on the full Stanford Sentiment Treebank, it is likely that its performance on HeiST is suffering from sparse data problems, and that both words and particular constructions can be novel and unseen.

In syntactic parsing, Koo et al. (2008) and Candito and Seddah (2010) have shown that using Brown clusters can be beneficial for alleviating sparse data problems in parsing. In a similar vein, Popat et al. (2013) have successfully applied crosslingual clustering to generalizing over potentially unseen words

[3] German original text has been omitted for space reasons

| System | Node Acc. | Root Acc. |
|---|---|---|
| *supervised baseline* | | |
| RNTN, supervised | 0.776 | 0.687 |
| *RNTN + clusters* | | |
| newspaper text+Brown | 0.708 | 0.649 |
| movie reviews+Brown | **0.780** | **0.677** |
| features+k-means | 0.755 | 0.674 |
| *RNTN + split movie-review clusters* | | |
| split *SentiWS* | 0.774 | 0.676 |
| split *GermanPolarityClues* | **0.807** | **0.689** |
| *RNTN + lexicon-based replacement* | | |
| *replace-gold* | *0.844* | *0.730* |
| repl-GermanPolarityClues | 0.789** | 0.681 |
| repl-SentiMerge[0.23] | 0.780* | 0.648 |

Table 5: Incorporating additional information

in (document-level) sentiment analysis for English, Hindi and Marathi.

In our experiments, we follow Candito and Seddah (2010) in simply replacing words by clusters: in their experiments, even this simple procedure can yield an improvement, with improved results when the unlabeled data stems from the target domain. Since Brown clusters are mostly syntactic/semantic in nature and do not automatically distinguish positive or negative sentiment, we additionally performed multiple experiments to use clusters while incorporating additional sentiment information:

On one hand, we try to incorporate the judgements on the Amazon near-domain dataset more directly into the clusters by using the repeated bisecting K-Means algorithm as implemented in CLUTO (Zhao and Karypis, 2005), with previous/next word, part-of-speech tag, and the score of the containing review as features. On the other hand, we split the Brown clusters according to the sentiment value that they have in a particular sentiment lexicon (e.g. SentiMerge), yielding three clusters *01101+*, *01101-* and *01101?* instead of the original cluster *01101*.

As a final experiment, we consider replacing *only sentiment words* by a concatenation of their part-of-speech and the sentiment class (turning "*a great film*" into "*a JJ++ film*"), and leaving neutral words intact. As an upper baseline for this approach, we can get words' sentiment polarity directly from training and testing data, which yields the *replace-gold* entry in table 5.

| *rule type* | # in SSTb | # in HeiST |
|---|---|---|
| AVG | 119468 | 19228 |
| INV | 2158 | 289 |
| INT | 6614 | 646 |
| MWE | 18235 | 1936 |

Table 6: Rule types in SSTb and HeiST

## 7 Results and Error analysis

Looking at the results in tables 2, 4 and 5, we see that simple support vector machine classification is very effective for reproducing the positive/negative sentiment of nodes and complete sentences, followed by crosslingual projection and a simple averaging approach; we also see that handling negation in the *vote-and-flip* approach seems to lower the score, just as the best model with word clusters and splitting (using the *GermanPolarityClues* lexicon) performs better than the word-based RNTN approach, but less well than the lexicon by itself. Even the *replace-gold* upper baseline – replacing sentiment-carrying words by their sentiment label, which raises the performance substantially – gives results below the simpler SVM approach, which is counterintuitive.

### 7.1 Is it about Compositionality?

One motivation for using sub-sentence structure both in approaches for rule-based composition (as, e.g. in Taboada et al. (2011) and other lexicon-based approaches) as well as in more complex learning approaches such as RNN (Socher et al., 2011) and RNTN (Socher et al., 2013) is the idea that such approaches are able to model the interaction between sentiment-bearing words and sentiment-modifying words. An example for investigations based on this assumption is the work by Zhu et al. (2014), who contrast different lexicon-based approaches for handling negation with an RNTN model of negation and a modification of said model.

Given the results using a lexicon-based approach implementing the *vote-and-flip* heuristic in comparison to the *vote-only* heuristic, we found it worth investigating what specific types of interaction exist in compositional sentiment treebanks, also considering that Zhu et al.'s investigations yielded a more precise picture of the sentiment-shifting action of negators as a highly lexicalized phenomenon.

For our analysis, we grouped the production rules $s_p \rightarrow s_l s_r$ in a sentiment treebank into one of the following categories:

**AVG** A production is said to be *averaging* if the parent category is within the range of either daughter category. (e.g. *mind-numbingly good* would be the composition of a negative term and a positive term to a positive term, which still fits the averaging heuristic).

**INV** A production is said to be *inverting* if one daughter category is neutral and the other daughter category is on the other side on the spectrum (e.g. "not great" landing on the negative side)

**INT** A production is said to be *intensifying* if the parent category is on the same side of the scale as the daughters but more extreme.

**MWE** A production is said to be a *multi-word* production if the daughter categories are classified as neutral while the parent category is not.[4]

As can be seen in table 6, the number of *inverting* and *intensifying* productions is dwarfed, both for the SST and for HeiST, by the number of *multi-word* rules. While it is likely that these counts are slightly distorted by noise in the annotation (as both datasets are the product of crowdsourcing), this fact is remarkable and merits further investigation.

**Types of multiword expressions** If we try to group the nodes with a "multiword" production, we can distinguish at least the following categories:

- **aspect descriptions**: In some cases, an adjective is specifically used to describe a (positive or negative) aspect of the movie, such as an *elaborate continuation*, or an *expanded vision*, where individual words have a neutral sentiment label (and conceivable could have been used in a non-aspect-specific way to convey a neutral or negative sentiment, such as an *elaborate perversion*, or an *expanded nightshift*). Similarly, *wenig Handlung* (not much action) has a negative meaning as a construction despite "*wenig*" (few/not much) not having a negative meaning itself.
- **expression strengthening** is a phenomenon that occurs when a term is judged as neutral by annotators by itself, but gains a sentiment value when paired with an intensifier or negator. For example, *intrusive* was labeled as neutral in SSTb, but *simply intrusive* as negative.
- **comparatives** are a very regular construction where too much of something is almost always bad: *too long*, *too insistent*, *too much*, *too many* are all negative in SSTb, just as *zu viel* (too many) and *zu wenig* (not enough) and other counterparts in HeiST are negative.
- **true constructions** such as *plot holes* or *historically significant* in SSTB, or *ruhigen Gewissens* (with a calm conscience) and *Finger weg* (don't touch it) in HeiST are both a problem for approaches relying purely on composition and not regular enough that we would expect to model it as a regular construction.

Some of the neutral-to-positive or neutral-to-negative transitions don't seem well-motivated and may be regarded as artifacts from the crowdsourcing, as *does n't*, *is n't* and *are n't* are negative in SSTb whereas *'s not*, *do n't* and *did n't* get a neutral label. In HeiST, *nicht immer* (not always) as well as *nicht ganz* (not quite) are negative, whereas *auch nicht* (neither) and *nicht so* (not as) or *nicht unbedingt* (not necessarily) are neutral.

The MWE productions seem to overlap with well-known linguistic phenomena – consider Fahrni and Klenner (2008) and their claim that most adjectives have a polarity that is dependent on the target they modify instead of having a '*prior*' polarity that holds independently of the target, or the observation of Su and Markert (2009) that sentiment should be dependent on word senses instead of word forms (which would capture a large number of examples within the *expression strengthening* category). Yet, others may be idiosyncracies introduced by the crowdsourcing process, and powerful learners such as RNTN or the approach of Hall et al. (2014) will gain performance from simply memorizing the idiosyncracies of the data when there is

[4]The MWE category also contains a small number – about 5% of total MWE productions – of positive-to-neutral and negative-to-neutral productions, which we found to be predominantly noise from the crowdsourcing process.

| | | SVM | | CLSA | | RNTN | | +*replace-gold* | |
|---|---|---|---|---|---|---|---|---|---|
| Type | Total | Corr | Prec | Corr | Prec | Corr | Prec | Corr | Prec |
| AVG | 6341 | 3408 | 0.546 | 3158 | 0.506 | 3604 | 0.577 | 4309 | 0.690 |
| MWE | 1638 | 369 | 0.225 | 538 | 0.328 | 538 | 0.359 | 546 | 0.333 |
| INT | 612 | 370 | 0.605 | 362 | 0.592 | 413 | 0.675 | 470 | 0.768 |
| ID | 392 | 283 | 0.722 | 269 | 0.686 | 286 | 0.730 | 323 | 0.824 |
| INV | 259 | 76 | 0.293 | 65 | 0.251 | 93 | 0.359 | 79 | 0.305 |

Table 7: Precision of rules with non-neutral parent label (ID: daughters and parent have identical labels)

enough of it – because of the way the Stanford Sentiment Treebank is constructed, phrases always have the same (context-independent) label, while we may get a more accurate (and possibly different) picture from introducing additional means of quality control (which in turn increases the necessary investment for such a sentiment treebank).

### 7.2 Contrasting SVM and RNTN behaviour

In table 7, we tabulated the classification accuracy for the parent node in different types of productions in HeiST. In this evaluation, we counted a production as correct whenever the parent node has the right sentiment label (in parallel with the *labeled recall* in syntactic evaluation), ignoring for the moment the question whether the production produced by a system falls into the same category. It is easy to see that AVG-type productions are the least error-prone for all classifiers, whereas MWE and INV productions pose a significant challenge for the models. We also see that on these challenging production, the RNTN performs better than the other methods.

## 8 Summary

We presented a novel dataset for subsentential sentiment classification, which uses the same conventions as the Stanford Sentiment Treebank (SSTb), which is the only German resource of this type besides the smaller (270 sentences) MLSA corpus (Clematide et al., 2012). We performed a systematic exploration into supervised, cross-lingual, and lexicon-based approaches on this dataset and found that, paradoxically, the performance of the state-of-the-art recursive neural tensor network (RNTN) models are severely impeded in this data-sparse situation, unlike latent-variable models for syntax which can deal with such conditions quite well: Lavelli and Corazza (2009), for example, reports that the best results for parsing on the very small TUT treebank (slightly more than 2000 sentences) can be achieved using a PCFG-LA model.

We showed that a wide variety of models – from lexicon-based sentiment prediction over SVM with unigram features to crosslingual classification – performs better than the RNTN, and that methods to improve RNTN performance that work in other settings (syntax) do not offer any easy fix.

In a second step, we took a closer look at the crowdsourced data in order to explain certain counterintuitive results (such as the fact that most sentiment lexicons do not benefit from negation handling, or that the *upper* baseline achievable with the RNTN by getting gold-standard information on positive and negative words is at about the same level as our SVM classifier), and found that SSTb-type resources show marked differences from e.g., the MLSA dataset as they incorporate *multiword* items, but seem to be challenging for the study of compositionality due to noise that is not present in expert-annotated resources.

## References


Banea, C., Mihalcea, R., and Wiebe, J. (2013). Porting multilingual subjectivity resources across languages. *IEEE Transactions on Affective Computing*, 2(4):211–225.

Candito, M.-H. and Seddah, D. (2010). Parsing word clusters. In *Proceedings of the First Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL 2010)*.

Choi, Y. and Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*.

Choi, Y. and Cardie, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U., and Wiegand, M. (2012). MLSA – a multi-layered reference corpus for german sentiment analysis. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*.

Dong, L., Wei, F., Zhou, M., and Xu, K. (2014). Adaptive multi-compositionality for recursive neural models with applications to sentiment analysis. In *AAAI 2014*.

Duh, K., Fujino, A., and Nagata, M. (2011). Is machine translation ripe for cross-lingual sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers*.

Emerson, G. and Declerck, T. (2014). SentiMerge: Combining sentiment lexicons in a Bayesian framework. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*.

Fahrni, A. and Klenner, M. (2008). Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Affective Language in Human and Machine (AISB 2008)*.

Ganchev, K., Graca, J., and Taskar, B. (2008). Better alignments = better translations? In *Procecedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL 2008)*.

Hall, D., Durrett, G., and Klein, D. (2014). Less grammar, more features. In *ACL 2014*.

Klenner, M., Petrakis, S., and Fahrni, A. (2009). PolArt: A robust tool for sentiment analysis. In *NODALIDA 2009*.

Koo, T., Carreras, X., and Collins, M. (2008). Simple semi-supervised dependency parsing. In *ACL 2008*.

Lavelli, A. and Corazza, A. (2009). The Berkeley Parser at the EVALITA constituency parsing task. In *Proceedings of the Workshop on Evaluation of NLP Tools for Italian (EVALITA 2009)*.

Nakagawa, T., Inui, K., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005*.

Petrov, S., Barett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *COLING-ACL 2006*.

Popat, K., Balamurali, A. R., Bhattacharyya, P., and Haffari, G. (2013). The haves and the have-nots: Leveraging unlabelled corpora for sentiment analysis. In *Proceedings of ACL 2013*.

Rafferty, A. and Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *ACL'08 workshop on Parsing German*.

Remus, R., Quasthoff, U., and Heyer, G. (2010). SentiWS - a publicly-available German-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation Conference (LREC 2010)*.

Samuelsson, Y. and Volk, M. (2004). Automatic node insertion for treebank deepening. In *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories (TLT 2004)*.

Socher, R., Hurval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *EMNLP 2012*.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP 2011*.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966).

*The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Su, F. and Markert, K. (2009). Subjectivity recognition on word senses via semi-supervised mincuts. In *Proceedings of NAACL 2009*.

Taboada, M., Brooke, J., Tofilofski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Tamura, H., Inui, T., and Okumura, M. (2005). Extracting semantic orientation of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.

Tiedemann, J. (2010). Lingua-align: An experimental toolbox for automatic tree-to-tree alignment. In *Proceedings of LREC 2010*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*.

Volk, M., Göhring, A., Marek, T., and Samuelsson, Y. (2010). SMULTRON (version 3.0) – The Stockholm MULtilingual parallel TReebank. `http://www.cl.uzh.ch/research/parallelcorpora/paralleltreebanks_en.html`.

Waltinger, U. (2010a). GermanPolarityClues: A lexical resource for German sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation Conference (LREC 2010)*.

Waltinger, U. (2010b). Sentiment analysis reloaded - an empirical study on machine learning.based sentiment classification using polarity clues. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST 2010)*.

Wan, X. (2009). Co-training for cross-lingual sentiment classification. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL-IJCNLP 2009)*.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of ACL 2012*.

Yessenalina, A. and Cardie, C. (2011). Compositional matrix-space models for sentiment analysis. In *EMNLP 2011*.

Zhao, Y. and Karypis, G. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10:141–168.

Zhu, X., Guo, H., Mohammad, S., and Kiritchenko, S. (2014). An empirical study on the effect of negation words on sentiment. In *ACL 2014*.