# Cross-Domain Contextualization of Sentiment Lexicons

**Stefan Gindl**[1] and **Albert Weichselbraun**[2] and **Arno Scharl**[1]

**Abstract.** The simplicity of using Web 2.0 platforms and services has resulted in an abundance of user-generated content. A significant part of this content contains user opinions with clear economic relevance - customer and travel reviews, for example, or the articles of well-known and respected bloggers who influence purchase decisions. Analyzing and acting upon user-generated content is becoming imperative for marketers and social scientists who aim to gather feedback from very large user communities. Sentiment detection, as part of opinion mining, supports these efforts by identifying and aggregating polar opinions - i.e., positive or negative statements about facts. For achieving accurate results, sentiment detection requires a correct interpretation of language, which remains a challenging task due to the inherent ambiguities of human languages. Particular attention has to be directed to the context of opinionated terms when trying to resolve these ambiguities. Contextualized sentiment lexicons address this need by considering the sentiment term's context in their evaluation but are usually limited to one domain, as many contextualizations are not stable across domains. This paper introduces a method which identifies unstable contextualizations and refines the contextualized sentiment dictionaries accordingly, eliminating the need for specific training data for each individual domain. An extensive evaluation compares the accuracy of this approach with results obtained from domain-specific corpora.

## 1 INTRODUCTION

Sentiment detection - a sub-area of opinion mining - tries to detect whether sentences, phrases or documents are favorable (positive sentiment) or unfavorable (negative sentiment). Due to the explosive growth of opinions and reviews available in Web resources such as forums, blogs, shopping and travel portals, etc. research on sentiment detection has gained considerably in importance. This development is further accelerated by studies which suggest that a considerable fraction of customers consult online reviews and that this reviews have a significant impact on the their purchase decisions [12].

Work on sentiment detection suggests that state of the art machine learning approaches do not unfold their full potential when applied to sentiment detection [13]. Pang et al. [13] believe that this is due to features such as ambiguities and subtle changes of tonal expressions which are not considered sufficiently in current applications. Approaches addressing this issue by considering the text's context are often limited to specific domains due to unstable contextualized terms which change their sentiment across domains.

Therefore, we extend the contextualization of sentiment terms by distinguishing between unstable (domain-specific) and stable

(domain-independent) contextualized sentiment terms and consider them accordingly (Section 3). Our research shows that stable contextualized sentiment terms are valuable features for sentiment detection methods which significantly improve the detection across domains (Section 4). In contrast, domain-specific terms tend to hurt the detection process, if they are not applied to the correct domain.

Contextualized sentiment lexicons consider the context in which a term appears in determining its sentiment. This paper introduces a technique for identifying unstable contextualized sentiment terms and removes these harmful terms from the contextualized sentiment lexicon, creating a generic domain-independent lexicon. The evaluations demonstrate that such a preprocessing of contextualized sentiment lexicons significantly improves the performance of a subsequent sentiment detection method. We conclude with an outlook on future work based on the proposed approach in Section 5.

## 2 RELATED WORK

Sentiment Detection has a rather long history; in early work Wiebe [22] classified subjective sentences. Hatzivassiloglou and McKeown [8] used syntactical relations to identify new sentimental terms, which can be considered as an early form of context exploitation. In general, sentiment detection techniques can be roughly divided into two sub-areas. Lexical approaches use sentiment lexicons – lists of terms tagged with a value indicating their sentiment polarity (i.e. positive or negative) – to determine a document's sentiment. The second sub-area are machine-learning approaches which exploit either syntactical features, such as POS tags, or linguistic features, such as the terms of a sentiment lexicon. All these approaches suffer from the ambiguity of human language. Thus, it is beneficial to consider a term's context to unravel its true sentiment. The approaches presented in the following use contextual information for sentiment detection. According to Nasukawa and Yi [11] sentiment detection is a three step process, where the identification of sentiment expressions is followed by the determination of their polarity and strength. The last step of the procedure identifies the subject the sentiment terms are related to. They model such relationships for verbs, which either directly transfer their own sentiment or another term's sentiment to the subject. With this model they are capable of treating expressions such as $t_i$ *prevents trouble* [11]. The verb *prevents* passes the opposite sentiment of the term *trouble* to the target $t_i$. Sentence particles different from verbs directly transfer their sentiment to the subject. Kim and Hove [9] specify subjects with a Named-Entity-Recognition and assign them the overall sentiment value of the sentence. A list of 44 verbs and 34 adjectives expanded by WordNet [6] synonyms and antonyms serves as sentiment lexicon. To handle complex sentence structures such as "the California Supreme Court *disagreed* that the state's new term-limit law was *unconstitutional*" [9] they developed a strategy, where several negative sentiment terms in one and the same sentence eliminate each other.

[1] MODUL University Vienna, Austria, Department of New Media Technology, email: {stefan.gindl, arno.scharl}@modul.ac.at
[2] Vienna University of Economics and Business, Austria, Department of Information Systems and Operations, email: albert.weichselbraun@wu.ac.at

Wilson et al. [23] examine 28 syntactical and linguistic features in a machine learning approach. Several of those features are context-based, e.g. invoking the sentence preceding or succeeding the current one or the document topic. The features are tested using BoosTexter's AdaBoost.MH algorithm [16] on the Multi-perspective Question Answering () Opinion Corpus [21]. The approach has two steps: the first step filters subjective sentences from objective ones and the second assigns sentiment values to the subjective sentences. In their successive work [24] Wilson et al. use four different machine learning algorithms to test their feature selection and also use a larger version of the corpus. Agarwal et al. [1] use the corpus to test n-grams and syntactical label relations as context characteristics. Polanyi and Zaenen propose context handling strategies from a linguistic perspective [14]. They distinguish two main groups of context modifiers: *Sentence Based Contextual Valence Shifters* and *Discourse Based Contextual Valence Shifters*. Esuli and Sebastiani [5] implicitly invoke contextual information by propagating sentiment values across WordNet synsets and store the data in SentiWordNet. They first manually label all synsets containing 14 seed terms, which results in an amount of 47 synsets with positive label and 58 with negative. All synsets obtained from certain relations (e.g. *direct antonymy*, *similarity* and *derived-from*) with these seed synsets are labeled accordingly. Synsets without connection to the seed sets are classified as objective, as long as they do not have a different sentiment value in the General Inquirer. The so gathered data is used to train eight ternary classifiers, which classify the rest of WordNet.

Turney and Littman [19] use Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA) to identify sentiment terms in a large Web corpus. Terms with sufficient co-occurrence frequency with one of 14 paradigm terms (i.e. a gold standard list of seven positive and negative terms) are assigned the same sentiment value as the respective paradigm term. Evaluated on the General Inquirer [17] PMI shows results comparable with the algorithm of Hatzivassiloglou and McKeown [8]. Using three different extraction corpora and the sentiment lexicon of [8] Turney and Littman show that PMI does not outperform Hatzivassiloglou's and McKeown's algorithm but is more scalable [20]. LSA also provided better results, but was not as scalable as PMI too. In [18] Turney uses the same techniques to identify new sentiment terms from a paradigm list of only two terms (*excellent* and *poor*). This procedure performed well on the review corpus. Beineke et al. re-interpret the previously discussed mutual association as a Naïve Bayes approach [2]; they also expand this perspective (which is an unsupervised approach) and create a supervised approach using labeled data.

Lau et al. [10] prove the importance of context by applying three different language models, whereof one is an inferential language model sensible for context. According to their evaluation the inferential language model outperforms the other two models, emphasizing the importance of context. Bikel and Sorensen apply a simple feature selection together with a perceptron classifier to reviews from Amazon.com [3]. They use all tokens with an occurrence frequency higher than four and achieve an accuracy of 89% in their experiments. Denecke [4] applies a machine learning approach to multilingual sentiment detection using movie reviews from six different languages. Google Translator (www.google.com/language_tools) translates foreign-language documents into English. The feature selection procedure extracts a total of 77 features out of four superclasses [4]: (1) the frequency of word classes (i.e. the number of verbs, nouns, etc.), (2) polarity scores for the 20 most frequent words and the averages scores for all verbs, nouns and adjectives are calculated using SentiWordNet [5]; other features are (3) the frequency

of positive and negative words according to the General Inquirer and (4) textual features such as the number of question marks. Using all features the Simple Logistic classifier of the WEKA tool[7] reaches exorbitantly good results when applied to native English documents. When applied to non-native, translated documents the results are still higher than the baseline demonstrating the efficacy of using a lexical resource such as SentiWordNet.

## 3 METHOD

A term's sentiment is often influenced by the context ($C$) in which it occurs as for instance its part-of-speech tag ("The Smiths *like* Susan" versus "Countries *like* Germany and Spain...") or even its textual context as in the example below.

- "Unless you like *annoying* sounds, do not purchase this product."
- "The text messaging can get *annoying* when trying to read and send text messages but you get use to it."
- "Older, bulkier phones got *annoying* after a while, so this was a nice change."

Ambiguous sentiment terms need contextualization (or disambiguation) to be useful for sentiment analysis. Such terms can be identified by their frequency graph. Figure 1 contrasts the graphs of unambiguous and ambiguous sentiment terms. Ambiguous terms such as "accident" have two maxima (positive as well as negative). Unambiguous sentiment terms have a strong focus on one particular polarity, whereas ambiguous terms have a more balanced frequency in both polarities.
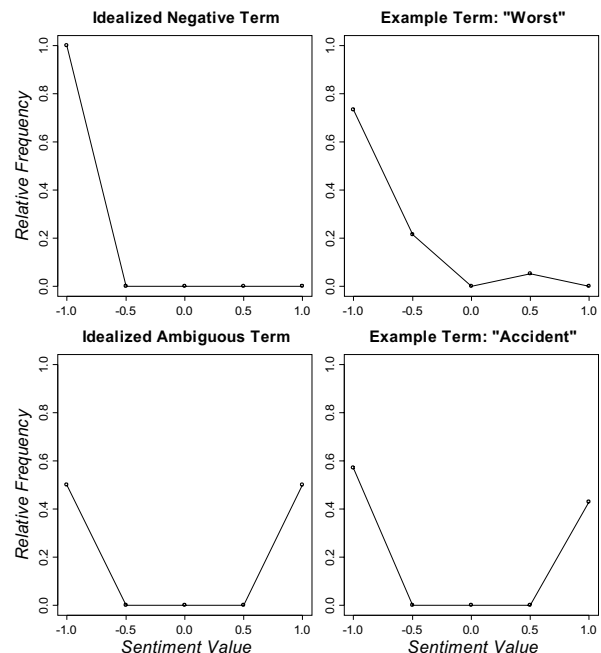


**Figure 1.** Comparison of the frequency graphs of idealized and real terms

Machine learning algorithms such as Naïve Bayes can help creating a contextualized sentiment lexicon which considers the term's context in the evaluation of its sentiment based on the context terms
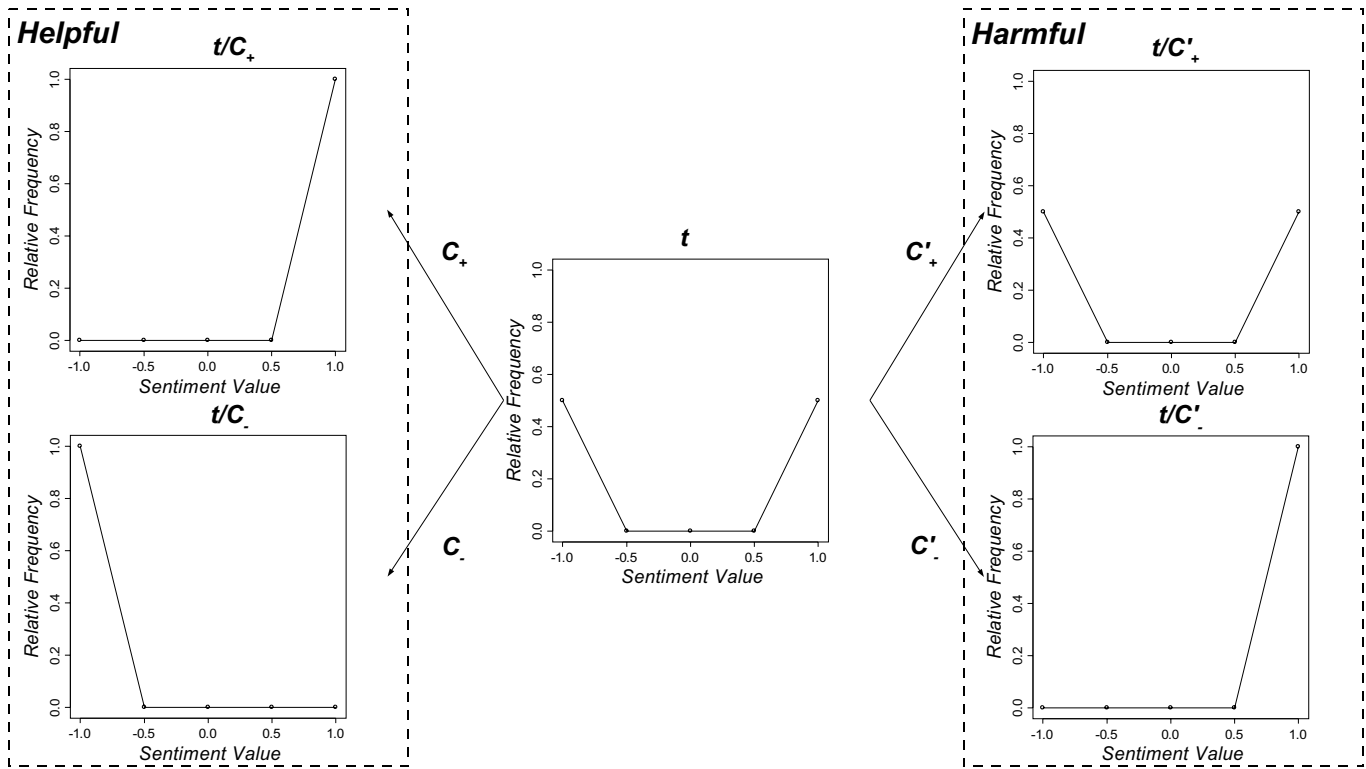
**Figure 2.** Frequency distributions of harmful and helpful terms

$\{c_1, ...c_n\}$ (Equation 1 and 2).

$$\mathbf{c} = \{c_1, ...c_n\} \tag{1}$$

$$p(C_+|\mathbf{c}) = \frac{p(C_+) \cdot \prod_{i=1}^{n} p(c_i|C_+)}{\prod_{i=1}^{n} p(c_i)} \tag{2}$$

A contextualized lexicon, therefore, contains (ambiguous) sentiment terms and the corresponding context terms ($c_i$) which help splitting sentiment terms with an ambiguous meaning into two unambiguous entries based on the given context ($C_+, C_-$).

In this work we apply a dictionary based method for sentiment detection which uses the Naïve Bayes algorithm to determine the context ($C$) of ambiguous terms and detects negations based on explicit linguistic features. A negation trigger (e.g. "not" or "never") preceding a sentiment term inverts its sentiment value. In other words, a negation trigger turns a positive term into a negative and vice-versa.

## 3.1 Domain-specific versus Domain-independent Sentiment Terms

Domain-specificity represents an inherent problem of contextualized sentiment lexicons. The contextualization might not remain valid across domains and sometimes even reduce the method's accuracy. Figure 2 outlines this problem. An ambiguous term is contextualized using the contexts $C_+$ and $C_-$. Transferring the term to another domain might preserve the contextualization's distinguishing properties (helpful), or it might loosen or even revert the term's polarity (harmful) as the contexts $C'_-$ and $C'_+$.

The latter group of terms is harmful to the sentiment analysis because their expected sentiment does not correspond to their real usage. Based on this insight we suggest the preprocessing process outlined below to identify and remove harmful terms from contextualized sentiment dictionaries.

## 3.2 Preprocessing Contextualized Sentiment Lexicons

Learning contextualized sentiment lexicons yields a knowledge base (KB) which contains ambiguous terms and their context terms. This knowledge base is typically linked to the training set or, put differently, non-generic. To obtain a generic knowledge base by combining two or more non-generic knowledge bases. For each knowledge base, we identify the terms which have been helpful or harmful in the sentiment detection step and those which have no real impact, i.e. the neutral ones. Afterwards, we generate the generic knowledge base by including helpful and neutral terms and eliminating harmful ones. Our hypothesis is that this procedure yields concepts (i.e. context terms $c_i$ related to ambiguous terms) common to several domains. In other words, they are considered universal concepts used to express sentiment. In the following we give a more detailed description of the procedure.

The generation of the generic knowledge base, assuming two corpora *A* and *B*, is accomplished in three steps (Figure 3):

1. Perform a cross-corpus evaluation to identify helpful, neutral and harmful context terms for both corpora.

2. From corpus *A* remove those helpful and neutral terms which are harmful in corpus *B* and vice versa.
3. Merge the remaining helpful and neutral terms into the generic knowledge base

For the first step, we split both corpora *A* and *B* into a training and a test set, ending up with $A_{training}$, $A_{test}$, $B_{training}$ and $B_{test}$. The system now creates the two non-generic knowledge bases on $A_{training}$ and $B_{training}$. Afterwards, it uses the Naïve Bayes algorithm on the respective test set, i.e. $KB(A_{training}) \rightarrow A_{test}$ and $KB(B_{training}) \rightarrow B_{test}$. This step yields correctly and incorrectly classified reviews. We compare the result from the baseline – which does not consider contextualized sentiment terms – with the previously obtained result from the Naïve Bayes approach. The identification of helpful, neutral and harmful terms is performed as follows:

- **Helpful terms:** sentiment terms in reviews which have been incorrectly classified by the baseline but correctly by Naïve Bayes.
- **Neutral terms:** sentiment terms in reviews where the contextualized knowledge base and the baseline yield the same result (we do not differentiate if that result was correct or incorrect).
- **Harmful terms:** sentiment terms in reviews which have been correctly classified by the baseline but incorrectly by Naïve Bayes.
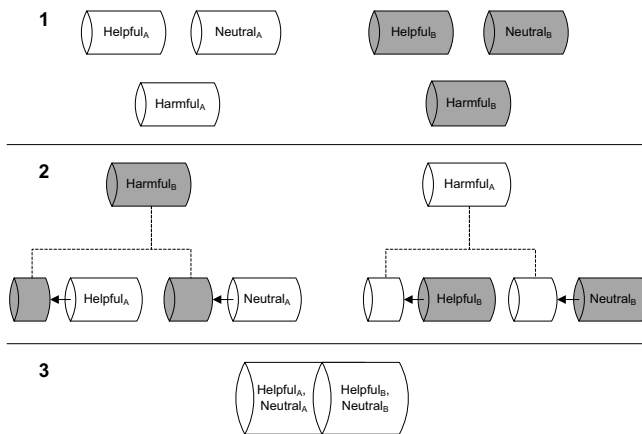


**Figure 3.** Extraction of harmful and helpful terms

In the next step the generic knowledge base is created by using only terms helpful or neutral in both corpora. Terms helpful (neutral) in one corpus but harmful in the other one are discarded, as well as terms harmful in both corpora. Thus, a smaller set of stable contextualized sentiment terms and their context information remains in each non-generic knowledge base. Merging the stable contextualized sentiment terms and re-calculating the probabilities creates the generic knowledge base.

### 3.3 Examples

In general, it is hard to find intuitive examples based on features learned by an automatic process – nevertheless the following real-world examples illustrate the need for preprocessing contextualized sentiment dictionaries. For both corpora we picked out the same ambiguous term affected by the same helpful context term. Context terms need not necessarily occur in the same sentence, as we define 'context' as the whole document a sentiment term is embedded in. For TripAdvisor, the example is straightforward:

> Sorry wish we could write **better** :-( Hotel lovely but service very *poor*, very *poor* for a five star!!!

The originally positive sentiment term **better** switches its sentiment value through two occurrences of *poor* in the subsequent sentence. The example for the Amazon corpus even spans several sentences:

> *Poor* support for Macs. ... I suggest HP abandon its support for Macs and it is **better** off than claiming its support but actually there is no support at all.

As mentioned before, examples intuitively comprehensible for humans are hard to find. The 'small' number of samples (i.e. the limited size of the training and test corpora) intensifies this effect even more. The next example, which was extracted from the TripAdvisor corpus, shows the impact of harmful terms:

> This hotel is **clean**, *efficient*, and lacking in human *hospitality*. (TripAdvisor)

The context terms *efficient* and *hospitality* establish a positive context ($C_+$) for the term **clean** (it is a positive term in the sentiment lexicon), although the sentence expresses a negative opinion.

We found a more complex example in the Amazon corpus. The sentence contains the ambiguous term *service*. The terms *small* and *fairly* indicate a positive context and would therefore assign **service** the correct positive sentiment value. Yet their influence is eliminated by the two harmful terms *cartridges* and *customer* suggesting a negative context ($C_-$). More harmful terms in the rest of the review (which is pruned to a sentence here for space reasons) intensify the effect and turn **service** into a negative sentiment term.

> It's *small* and *fairly* quiet, and since it is from Hewlett-Packard you know ink *cartridges* and great *customer* **service** will always be available. (Amazon)

Pruning such ambiguous entries from the contextualized sentiment lexicon improves the accuracy of sentiment detection. The examples above demonstrate that the Naïve Bayes approach for context detection will benefit from the preprocessing step.

## 4 EVALUATION

For the evaluation of the proposed approach we used 2 500 Amazon product reviews (1 250 positive and 1 250 negative reviews) and 1 800 travel reviews (900 positive and 900 negative reviews) from TripAdvisor. These corpora cover reviews from different domains (product reviews vs. travel reviews), which is essential to identify generic (i.e. domain-independent) contextualized sentiment terms. The evaluation is accomplished as a 10-fold cross validation. For each run, the training set has a size of 2 250 reviews, the test set contains 250 reviews. Both sets of each run contain a similar number of positive and negative reviews. The evaluation addresses the following two research questions:

- **Intra-domain sentiment detection** – how does the preprocessing step impact the performance of the intra-domain sentiment detection? Does the removal of unstable contextualized sentiment terms degrade the sentiment detection's performance for evaluations performed in the same domain?

- **Cross-domain sentiment detection** – how does the preprocessing step affect the accuracy of the cross-domain sentiment detection? Does our approach outperform contextualized dictionaries learned from domains different from the test domain?

The first research question ensures that the preprocessing step does not reduce the performance of the sentiment detection by removing important sentiment terms from the contextualized sentiment lexicon. The second question addresses the suitability of the generated generic knowledge base for cross-domain sentiment detection by verifying that the generic contextualized sentiment lexicon which was created by the preprocessing step outperforms the baseline and the domain-specific lexicons in cross-domain sentiment detection tasks.

All significance values were obtained with Wilcoxon's rank sum test implemented in R [15]. Figure 4 contains the result graphs for each test set with all three possible training combinations.
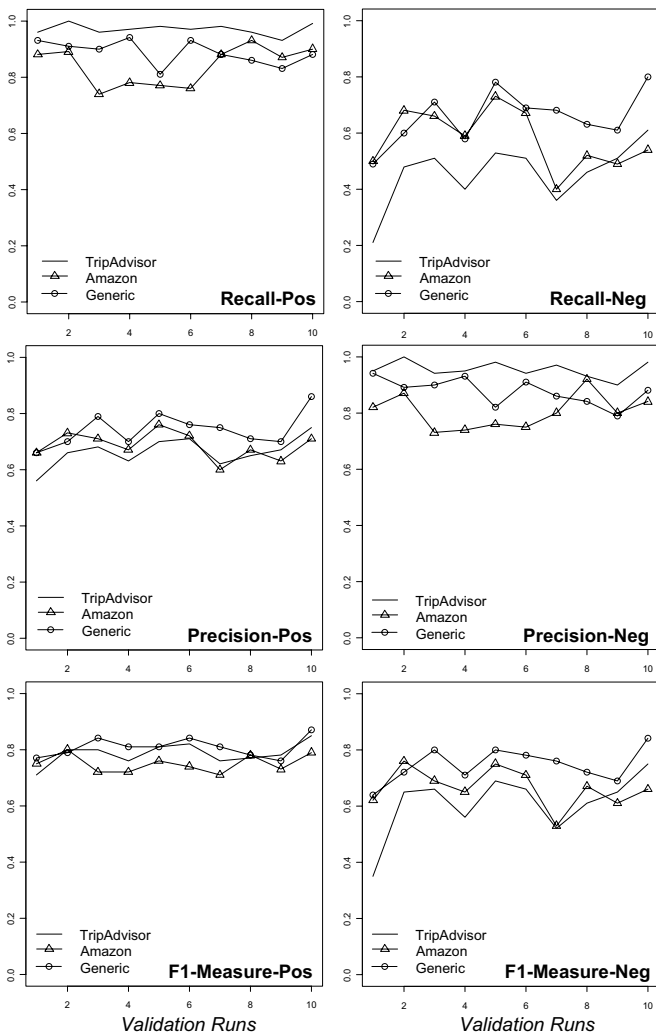


**Figure 4.** Graphical overview over all cross-validation results for all three knowledge bases (Test corpus: TripAdvisor)

## 4.1 Intra-Domain Sentiment Detection

This evaluation uses the same training and test corpus to assess the impact of the pre-processing on intra-domain sentiment detection.

The experiments show that the presented approach does not lower the sentiment detection's performance compared to the domain-specific sentiment lexicons. Table 1 contains the evaluation results. Amazon data did not reflect a significant change in the method's performance – the generic version performed as well as the non-generic version on this corpus. Tested on TripAdvisor, the results for recall and precision are contradictory. Significant losses in positive recall and negative precision are accompanied by significant gains in positive precision and negative recall. This should nevertheless considered an improvement since (i) the results are more balanced - the unfiltered lexicon yields a very high precision ($\approx 95\%$) for negative polarity at the cost of a low recall ($\approx 46\%$); and (ii) the F1-Measure has increased significantly, for both polarities.

**Table 1.** Equivalence comparison on the Amazon and TripAdvisor dataset

|  | Intra-domain (Amazon) | | | Generic | | |
|---|---|---|---|---|---|---|
|  | $\overline{R}$ | $\overline{P}$ | $\overline{F}_1$ | $\overline{R}$ | $\overline{P}$ | $\overline{F}_1$ |
| Pos | 0.75 | 0.75 | 0.74 | 0.77 | 0.72 | 0.74 |
| Neg | 0.71 | 0.79 | 0.73 | 0.67 | 0.77 | 0.72 |

|  | Significance | | |
|---|---|---|---|
|  | $p_R$ | $p_P$ | $p_{F_1}$ |
| Pos | 0.63 | 0.11 | 0.91 |
| Neg | 0.51 | 0.56 | 0.56 |

|  | Intra-domain (TripAdvisor) | | | Generic | | |
|---|---|---|---|---|---|---|
|  | $\overline{R}$ | $\overline{P}$ | $\overline{F}_1$ | $\overline{R}$ | $\overline{P}$ | $\overline{F}_1$ |
| Pos | 0.97 | 0.66 | 0.79 | 0.89 | 0.74 | 0.81 |
| Neg | 0.46 | 0.95 | 0.61 | 0.66 | 0.87 | 0.75 |

|  | Significance | | |
|---|---|---|---|
|  | $p_R$ | $p_P$ | $p_{F_1}$ |
| Pos | 0.01 | 0.01 | 0.05 |
| Neg | 0.01 | 0.01 | 0.00 |

## 4.2 Cross-Domain Sentiment Detection

To assess whether non-generic contextualized sentiment lexicons can outperform generic ones, we applied the Amazon knowledge base to TripAdvisor data and the TripAdvisor knowledge base on Amazon data, and compared the results with the results of the generic knowledge base on both test sets. The comparison shows that the pre-processing step considerably improves the method's performance in cross-domain settings. Table 2 presents the averages for recall, precision and the F1-measure, and the significance values when compared with the generic results in Table 1. In the case of Amazon, only positive recall did not show a significant increase in the number of correctly identified reviews. In the case of TripAdvisor, precision and F1-measure increased significantly for both polarity classes. The average values for positive and negative recall also increased, but not significantly.

## 5 CONCLUSION AND FUTURE WORK

This paper presents an approach that addresses the problem of polarity shifts of contextualized sentiment terms across domains. We suggest a method which identifies unstable contextualizations and removes them from the contextualized sentiment lexicon creating a domain-independent (generic) contextualized sentiment lexicon. An extensive evaluation shows that the generic lexicon performs on our

**Table 2.** Cross-domain comparison: The first dataset served for training, the second for testing

| | TripAdvisor on Amazon | | | Significance | | |
|---|---|---|---|---|---|---|
| | $\overline{R}$ | $\overline{P}$ | $\overline{F}_1$ | $p_R$ | $p_P$ | $p_{F_1}$ |
| Pos | 0.76 | 0.67 | 0.71 | 0.76 | 0.01 | 0.03 |
| Neg | 0.58 | 0.73 | 0.64 | 0.02 | 0.03 | 0.01 |
| | Amazon on TripAdvisor | | | Significance | | |
| | $\overline{R}$ | $\overline{P}$ | $\overline{F}_1$ | $p_R$ | $p_P$ | $p_{F_1}$ |
| Pos | 0.84 | 0.69 | 0.75 | 0.19 | 0.02 | 0.01 |
| Neg | 0.58 | 0.8 | 0.66 | 0.07 | 0.04 | 0.01 |

data as well as domain-specific lexicons and clearly outperforms domain-specific lexicons in cross-domain evaluations.

The main contributions of this paper are (i) the introduction of the concepts of domain-independent (stable) and domain-dependent (unstable) contextualized sentiment terms, (ii) presenting an approach which identifies unstable contextualized sentiment terms and removes them from the sentiment lexicon, and (iii) performing an extensive evaluation which shows that the created domain-independent contextualized sentiment lexicons perform significantly better on our data than the domain-specific ones.

The evaluation showed the positive effect of the pre-processing process on the sentiment detection's performance. Removing unstable contextualized sentiment terms significantly improves the knowledge base's cross-domain applicability. A positive side-effect of the pre-processing step is that the number of ambiguous terms and context terms is reduced by a fair amount which reduces storage requirements and improves the method's throughput.

While this work limits its observation of the impact of unstable contextualized sentiment terms on sentiment detection accuracy to the approach introduced in this paper, future work will explore their impact on other machine learning techniques and evaluate the extent to which a more accurate handling of unstable contextualized sentiment terms helps unlocking the full potential of machine learning approaches for sentiment detection.

## ACKNOWLEDGEMENTS

## REFERENCES

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown, 'Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams', in *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 24–32, Morristown, NJ, USA, (2009). Association for Computational Linguistics.

[2] Philip Beineke, Trevor Hastie, and Shivakumar Vaithyanathan, 'The sentimental factor: improving review classification via human-provided information', in *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 263, Morristown, NJ, USA, (2004). Association for Computational Linguistics.

[3] Daniel M. Bikel and Jeffrey Sorensen, 'If we want your opinion', in *ICSC 2007. International Conference on Semantic Computing*, pp. 493 – 500, Irvine, CA, (September 2007).

[4] Kerstin Denecke, 'How to assess customer opinions beyond language barriers?', in *Third International Conference on Digital Information Management*, pp. 430 – 435. IEEE, (November 2008).

[5] Andrea Esuli and Fabrizio Sebastiani, 'Sentiwordnet: a publicly available lexical resource for opinion mining', in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 417–422, (2006).

[6] C. Fellbaum, 'Wordnet - an electronic lexical database', *Computational Linguistics*, **25**(2), 292–296, (1998).

[7] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, 'The weka data mining software: an update', *SIGKDD Explorations Newsletter*, **11**(1), 10–18, (2009).

[8] Vasileios Hatzivassiloglou and Kathleen R. McKeown, 'Predicting the semantic orientation of adjectives', in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pp. 174–181, Morristown, NJ, USA, (1997). Association for Computational Linguistics.

[9] Soo-Min Kim and Eduard Hovy, 'Determining the sentiment of opinions', in *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, p. 1367, Morristown, NJ, USA, (2004). Association for Computational Linguistics.

[10] R.Y.K. Lau, C.L. Lai, and Yuefeng Li, 'Leveraging the web context for context-sensitive opinion mining', in *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, pp. 467–471, (Aug. 2009).

[11] Tetsuya Nasukawa and Jeonghee Yi, 'Sentiment analysis: capturing favorability using natural language processing', in *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, pp. 70–77, New York, NY, USA, (2003). ACM.

[12] B. Pang and L. Lee, 'Opinion mining and sentiment analysis', *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135, (2008).

[13] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, 'Thumbs up? sentiment classification using machine learning techniques', in *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pp. 79–86, Morristown, NJ, USA, (2002). Association for Computational Linguistics.

[14] Livia Polanyi and Annie Zaenen, 'Contextual valence shifters', in *Computing Attitude and Affect in Text: Theory and Applications*, The Information Retrieval Series, (2006).

[15] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[16] Robert E. Schapire and Yoram Singer, 'Boostexter: A boosting-based system for text categorization', *Machine Learning*, **39**(2-3), 135–168, (November 2000).

[17] Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith, *The general inquirer : a computer approach to content analysis*, M.I.T. Press, Cambridge, Mass. [u.a.], 1966.

[18] Peter D. Turney, 'Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews', in *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417–424, Morristown, NJ, USA, (2002). Association for Computational Linguistics.

[19] Peter D. Turney and Michael L. Littman, 'Unsupervised learning of semantic orientation from a hundred-billion-word corpus', Technical report, National Research Council, Institute for Information Technology, (2002).

[20] Peter D. Turney and Michael L. Littman, 'Measuring praise and criticism: Inference of semantic orientation from association', *ACM Transactions on Information Systems*, **21**(4), 315–346, (2003).

[21] Janyce Wiebe, Theresa Wilson, and Claire Cardie, 'Annotating expressions of opinions and emotions in language', *Language Resources and Evaluation (formerly Computers and the Humanities)*, **39**(2/3), 164–210, (2005).

[22] Janyce M. Wiebe, 'Tracking point of view in narrative', *Computational Linguistics*, **20**(2), 233–287, (1994).

[23] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 'Recognizing contextual polarity in phrase-level sentiment analysis', in *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347–354, Morristown, NJ, USA, (2005). Association for Computational Linguistics.

[24] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 'Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis', *Computational Linguistics*, **35**(3), 399–433, (2009).