Generating Contextualized Sentiment Lexica based on Latent Topics and User Ratings

Ralf Krestel
University of California, Irvine
Donald Bren Hall, Irvine, USA
krestel@uci.edu

Stefan Siersdorfer
L3S Research Center
Appelstr. 9a, Hannover, Germany
siersdorfer@l3s.de

ABSTRACT

Sentiment lexica are useful for analyzing opinions in Web collections, for domain-dependent sentiment classification, and as subcomponents of recommender systems. In this paper, we present a strategy for automatically generating topic-dependent lexica from large corpora of review articles by exploiting accompanying user ratings. Our approach combines text segmentation, discriminative feature analysis techniques, and latent topic extraction to infer the polarity of n-grams in a topical context. Our experiments on rating prediction demonstrate a substantial performance improvement in comparison with existing state-of-the-art sentiment lexica.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Dictionaries, Linguistic processing, Thesauruses*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Web 2.0*

General Terms

Algorithms, Experimentation, Measurement

Keywords

Sentiment Lexica, Topic Models, Latent Dirichlet Allocation, Sentiment Analysis, Rating Prediction

1. INTRODUCTION

The rapidly increasing popularity of user-generated content is based on the availability of suitable and easy to use mechanisms for publishing blog articles, product reviews, comments on news events, and contributions to discussion forums. The blogosphere has attracted an active web community in the recent years, and has become a popular environment for sharing experiences, opinions, and thoughts on a variety of issues. Topics discussed range from rather casual themes such as sports, concerts, and celebrities to more complex and polarizing political ones such as abortion, elections, and immigration. Large online-review communities on platforms such as Epinions, Amazon, or IMDB contain a variety of opinionated views on books, movies, and consumer electronics. Content sharing platforms such as YouTube and Flickr provide different social tools for community interaction, including the possibility to comment on existing resources.

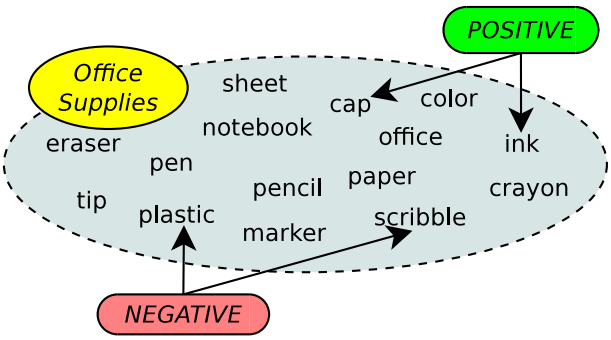


Figure 1: An example of a topic with topic-specific sentiment assignments

Techniques for automatic extraction of sentiments and opinions allow for a variety of applications such as opinion-oriented search, prediction of trends, summarization of product aspects, and filtering of flames in newsgroups. Liu et al. [23] make use of sentiment analysis techniques to predict movie incomes by mining blogs, carrying the potential to utilization in market analysis and business planning. Lu et al. [25] use short comments on products to provide an aggregated view on user opinions about thematic aspects such as “shipping”, “communication”, or “service”. Turney and Litman [37] mention several additional existing and potential applications of sentiment analysis such as corpus linguistics, aggregated views in the form of sentiment timelines, and even AI components in computer games responding in a more realistic way to the player’s textual input.

Many of the above mentioned applications build on *sentiment lexica* which provide information on the typical polarity of words. For instance, SentiWordNet [2], a lexical resource built on top of WordNet [13], assigns triples of *sentivalues* (corresponding to positive, negative, and neutral sentiment of a word). Turney and Litman [37] use a small seed set of polarized terms for automatically extracting additional sentiment terms from larger text corpora. Although there are general terms that almost always carry the same sentiment, their polarity can be highly *topic-dependent*, as pointed out, for instance, by Nowson [28] (“Scary Films Good, Scary Flights Bad”). This makes sentiment analysis across different domains and contexts a challenging task.

24th ACM Conference on Hypertext and Social Media
1–3 May 2013, Paris, France

The work on sentiment lexicon generation presented in this paper is placed in between lexica in broad domains [19] and fine grained lexica on specific opinion-entity pairs [6] or aspects [27]. We aim to determine the polarity of words in a *topical context*. In addition, we aim to assign sentiment scores to n-grams and named entities — something traditional lexica fail to do. To this end, we are the first to combine *topic modelling* with information obtained from *user ratings* in review articles through discriminative feature selection to extract topic-specific sentiment lexica. Experiments for finding topics and associate sentiment values to terms were conducted on a dataset containing $27,375$ Epinions reviews. We performed a classification-based evaluation on Epinions as well as on a well-known multi-domain sentiment dataset [5].

Figure 1 shows example entries from our lexicon. The top-15 terms composing the latent topic "office supply" are shown. Within this topical context, our algorithm identified a negative sentiment value attached with "plastic" and "scribble", whereas "cap" and "ink" are perceived as mainly positive. Especially the sentiment value of "plastic" is highly topic dependent, since for other products plastic is associated with a rather positive sentiment value due to its light weight and inexpensive production.

The rest of this paper is organized as follows: In Section 2 we discuss related work on sentiment lexica and context- or topic-specific sentiment analysis and classification techniques. We describe our technical approach for extracting topic-specific sentiment lexica in Section 3. In Section 4 we provide the results of the evaluation of our lexicon generation method through sentiment classification experiments, and we compare our approach to existing state-of-the-art sentiment lexica. We conclude and show directions of our future work in Section 5.

2. RELATED WORK

Sentiment Lexica. There exist various domain-independent sentiment lexica, one of the most prominent being SentiWordNet [11], which was extended by exploiting the graph structure of the underlying WordNet lexicon using Page-Rank-like propagation of sentiment values [2]. Manual sentiment annotations can be found in the MPQA corpus [40]. Whitelaw et al. [39] created a lexicon consisting of around 2000 manually selected, general sentiment terms. Turney and Litman [37] make use of a small set of seed terms with positive and negative semantic orientation, and estimate the polarity of new terms by computing co-occurrence based statistics (using pointwise Mutual Information and Latent Semantic Indexing). Some of the technical components in this paper resemble the ones used by Turney and Litman but, in contrast, are applied in our work to identify *topical context* and exploit *rating scores* in reviews. In our experiments we will show comparisons to the above mentioned lexica. In addition, there are a couple of approaches generating sentiment lexica semi-automatically from the Web. Velikovic et al. [38] present a method based on graph propagation. This leads to large lexica which also include sentiment scores for non-standard terms, e.g. slang, spelling mistakes, and phrases or n-grams comparable to our approach. Kaji and Kitsuregawa [18] make use of structural clues to extract sentiment bearing sentences from Japanese Web sites. In contrast, we make use of *rated* review articles to generate *topic-dependent* lexica.

Context- and Topic-Dependent Sentiment Lexica. There is also work on generating sentiment lexica that take context and topic information into account. The problem of polarity shift of adjectives in certain domains [28] inspired Fahrni and Klenner [12] to identify domain-specific nouns, and create specific sentiment lexica of adjectives for these target nouns. Wikipedia is used to find the nouns and polarity is estimated through a bootstrapping approach for extracting patterns. Kanayama and Nasukawa [19] identify opinion terms using a seed set of general polarity terms and their connections to other entities in reviews in order to discover new polarity terms. They achieve domain orientation by applying their methods separately on discussion board corpora from different domains. Similarly, starting with a seed of sentiment words, Qiu et al. [32] iteratively expand sentiment lexica through connection to other terms for separate review corpora on categories like "digital cameras", "DVD players" or "cell phones". Also starting with a seed set of sentiment terms, Jijkoun et al. [16] extract syntactic patterns and potential targets from a background corpus. They compare the frequency of occurrences in the background corpus with the frequencies in a topic-specific set of documents using chi-square. The topic-specific corpus is obtained by querying a corpus with a topic keyword. For the top targets, sentiment terms are then extracted for the topic-specific lexicon. Yejin Choi and Claire Cardie [7] describe an approach to adapt a general sentiment lexicon for specific domains using integer linear programming. Bross and Ehrig [6] analyze review data with a specific *pros* and *cons* structure to identify the polarity of opinion tuples (o, p) of opinion words o and entities p (e.g. ("intuitive", "menu")) by exploiting the correlation of their occurrence in the *pros* and *cons* lists. A context-dependent sentiment lexicon also based on such tuples within a fixed set of domains is presented in [24]. For each domain, terms are grouped together into aspects like "service", "food", etc. and a lexicon is generated for tuples of aspect terms and opinion terms. An optimization framework is described which combines linguistic heuristics, document ratings, and a general domain-independent lexicon. Li et al. [22] propose the generation of domain-dependent lexica using cross-domain co-extraction. Given a well-labeled source domain, sentiment terms and topic terms are identified based on seed sets and bootstrapping. Co-extraction of patterns within the source and target domains is used to extend the seed set iteratively. The cross-domain learning algorithm employed is Transfer AdaBoost to learn different weights for the different domains.

In contrast to the described works, the topical context studied in our approach is placed in between a small number of large and fixed domains and very fine-grained entity specific sentiment assignments. Furthermore, our techniques are orthogonal to the described ones in the sense that we make use of numeric ratings accompanying review articles, and, on the other hand, do not rely on predefined domains as is the case in [24]. The work by Li et al. [21] introduces a joint model of latent topics and sentiment using a seed set of manually selected terms; however, they do not exploit information from review ratings to build their lexica. We compare to this approach in our experiments, and show that considering such ratings leads to a substantial performance boost.

Context-Dependent Sentiment Analysis. There are several works that analyze sentiment in the context of topics and aspects. Choi et al. [8] analyze corpora based on queries related to different domains like "business" or "politics" separately for each of these domains to improve sentiment oriented search. Xia and Zong [42] study cross-domain sentiment classification based on part-of-speech information. Lu et al. [25] describe a method for

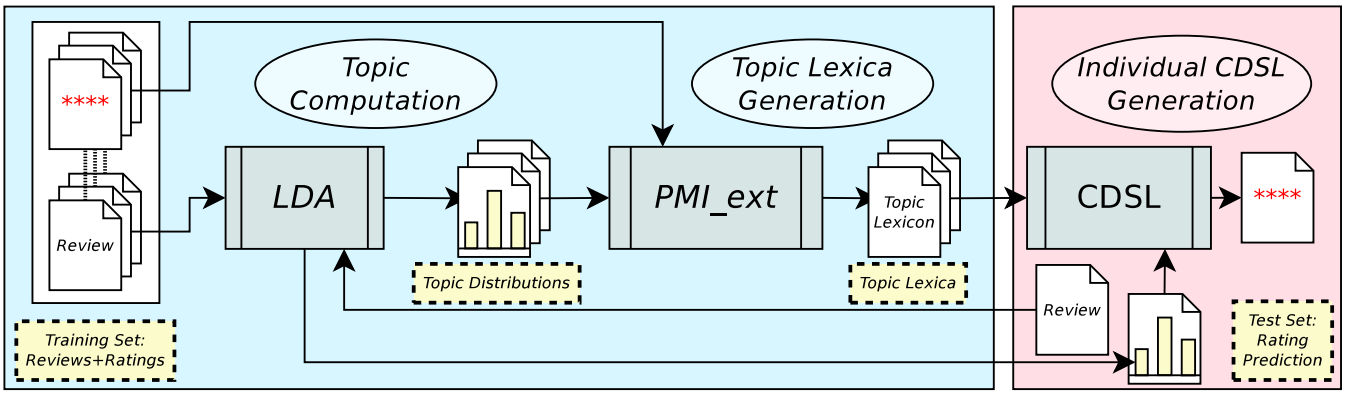


Figure 2: System Overview

summarizing the opinion on product aspects from a set of short comments. They employ a three-phase approach in which they first extract latent topics using PLSI, then classify the sentiment of short comments, and finally aggregate over the obtained sentiment scores. Titov and McDonald [35] focus on the first of these steps, and extract multi-grain topics from online reviews. They apply an LDA-like approach coined MG-LDA to distinguish between global topics (e.g. "London") and ratable aspects within these topics (e.g. "transportation"). Moghaddam and Ester [27] use an interdependent LDA model (ILDA) to extract product aspects and identify implicit ratings for these aspects (e.g. aspect "zoom" is assigned a rating of 5 based on the review text "...excellent zoom..."). In contrast, we focus on the task of automatic sentiment *lexicon* generation based on topical context.

Sentiment Classification. Finally, there is a plethora of work on sentiment classification and rating prediction. Sentiment classification (described, for instance, in [30]) deals with the problem of automatically assigning opinion values (e.g. "positive" vs. "negative" vs. "neutral") to documents or topics using various text-oriented and linguistic features. Recent work in this area makes also use of SentiWordNet to improve classification performance [10]. Qu et. al [33] apply regression models learned on training sets of review-rating pairs to predict product ratings. Blei and McAuliffe [3] and Yohan and Oh [17] modify latent topic models to consider information from additional response variables (e.g. rating scores of product reviews), and use their models for rating prediction. Cross-domain sentiment classification was studied in [29] where spectral graph analysis is used to infer links between domain-independent and domain-specific terms. Similar to many of the aforementioned works we borrow from latent topic modelling, especially latent Dirichlet allocation (LDA) [4], as well as discriminative feature analysis and selection [44]. However, the focus of our work is on the generation of topic-specific sentiment *lexica* rather than on predictions or summaries of ratings and sentiments.

3. LEXICON GENERATION

A context-dependent sentiment lexicon assigns sentiment values to terms depending on their topical context. This context can be defined by the surrounding terms [6], the comprising sentence [41], paragraph, document, or the whole domain [14]. In our approach, we propose to consider the document as the decisive level of granularity. Within a corpus of user-generated reviews, this allows for identifying different product categories (or services) and computing sentiment lexica for them or their aspects. How fine-grained a context is defined depends on the number of topics for the whole corpus, and is set as a parameter for the topic detection algorithm. Depending on the document context, sentiment values can significantly differ for individual terms. If, for instance, the term "train station" occurs in a hotel review, it is mostly associated with a negative sentiment due to the noise of the trains. On the other hand, in a football stadium review, "train station" will more likely be associated with a positive sentiment because of improved accessibility. In order to find the positive and negative terms within a context or product category, we conduct a discriminative analysis of the review terms exploiting user-assigned ratings.

Figure 2 gives an overview of our system. Our method for generating a context-dependent sentiment lexicon (CDSL) for a test document can be broken down into four steps: 3.1 Preprocessing; 3.2 Topic Computation; 3.3 Topic Lexica Generation; 3.4 Individual CDSL Generation. In the following we elaborate on the different steps in detail.

3.1 Preprocessing

Review data contains many stop words and function words. We used the Stanford part-of-speech tagger[1] to identify nouns, verbs, adjectives, and adverbs and discard other types. Although adjectives carry the most sentiment, other part-of-speech classes exhibit sentiment as well [30, 34], thus should not be discarded. We further used WordNet[2] to find the lemmas of each term. After that, we generated a list of n-grams from each review; in our experiments we obtained the best results for a combination of unigrams, bigrams, and trigrams. Finally, we discarded all n-grams occurring less then 5 times in our corpus in order to eliminate idiosyncratic terms. We also experimented with the removal of stop words using stop word lists or term frequencies within the corpus. However, the best results were achieved by removing only the verbs "to be" and "to have". After these steps, each review was represented as a list of POS-tagged, lemmatized n-grams.

3.2 Topic Computation

Instead of considering explicitly given categories, we automatically extract topics using latent topic analysis. Even in the special case of reviews, where each review is – often in a hierarchical manner – categorized into exactly one product group, automatic

[1] http://nlp.stanford.edu/software/tagger.shtml
[2] http://wordnet.princeton.edu/

topic/domain identification can be beneficial (see results in Section 4.4). For instance, for books, the category "Media" might not appropriately reflect the topics for our sentiment lexicon, as the category is possibly too broad. Since our objective is to provide topic-specific lexica for general purposes we aim to detect the topic(s) of a text automatically. For this step, we employ latent Dirichlet allocation [4], which additionally allows for a probabilistic assignment of different topics to a single review.

Latent Dirichlet allocation (LDA) identifies a given number of $|Z|$ topics within a corpus. Being the most important parameter for LDA, this number determines the granularity of the resulting topics. In order to find the latent topics, LDA relies on probabilistic modeling. This process can be described as determining a mixture of topics z for each document d in the corpus, i.e., $P(z|d)$, where each topic is described by n-grams w following another probability distribution, i.e., $P(w|z)$. This can be formalized as

$$P(w_i|d) = \sum_{j=1}^{|Z|} P(w_i|z_j)P(z_j|d), \quad (1)$$

where $P(w_i|d)$ is the probability of the ith n-gram for a given document d, $P(w_i|z_j)$ is the probability of w_i within latent topic z_j, and $P(z_j|d)$ is the probability of picking a term from topic z_j in the document. We make use of Gibbs Sampling [15] for computing the topic model[3].

By applying LDA we are able to represent latent topics as a list of n-grams with a probability for each n-gram indicating the membership degree for the topic. Furthermore, for each document in our corpus (reviews in our case) we can determine through topic probabilities $P(z_j|d_i)$ to which topics it belongs and to which degree. In the next step, we assign the documents in the corpus to latent topics. To this end, we iterate for each latent topic over the reviews and assign all reviews to this topic based on the topic probability. Thus, our corpus is divided into overlapping sub-corpora with each one representing one latent topic. Based on the document collections for each topic, we generate the topic-dependent lexica in the next step.

3.3 Topic Lexica Generation

In order to build topic-dependent lexica we employ statistical methods for analyzing a large amount of product reviews covering a variety of products and services. We claim that product reviews together with their star rating can be used to identify terms with positive or negative semantic orientation. A user assigning the maximum number of stars to a product is likely to write a positive review using positive terms to describe the product. Conversely, a low rating is likely to be reflected by usage of rather negative terms. In the following we combine these star rating with probabilities assigned to terms and documents through LDA in order to construct sentiment lexica.

The LDA-based approach described in the previous subsection generates latent topics for the whole corpus. Each document d_i is modeled as a distribution over topics Z, and can be represented as a mixture of topics as follows: $\sum_{z_j \in Z} P(z_j|d_i) = 1$. Reciprocally, each topic z_j is generated by different documents. By defining a threshold on the topic probability $P(z_j|d_i)$ for each document d_i, we can employ these topic generating documents to obtain a topic-specific corpus of reviews. The probability $P(z_j|d_i)$ is used as a weighting factor for each document d_i. This results in a corpus for each latent topic with weighted documents and assigned star ratings. To compute sentiment values for a latent topic we assign documents to the "positive" or "negative" class $C = \{pos, neg\}$ with a certain probability depending on the assigned star ratings (class probability). In order to identify the most discriminative terms we extend the mutual information measure. Pointwise mutual information (PMI) is an information theoretic measure to compute the mutual dependence between two random variables. Mutual information has been used in the past to do feature selection for text categorization [44, 46] and for exploiting the occurrence of emoticons in blogs to build an emotion lexicon [43]. In our case, we are interested in the dependence between the occurrence of a n-gram w and the membership to the positive or negative class $c \in C$. In its general form, pointwise mutual information is defined as

$$PMI(w, c) = \log\left(\frac{P(w, c)}{P(w) \cdot P(c)}\right) \quad (2)$$

where w is a n-gram, c is a class (in our case in $\{pos, neg\}$ corresponding to positive or negative sentiment), and $P(w, c)$ is the probability that n-gram w occurs in a document of class c. $P(w)$ is computed using maximum likelihood estimation, which returns the document frequency in this case; $P(c)$ is estimated analogously using the class frequency. In order to account for different degrees of topic membership and rating polarity, we incorporate the topic probability of a document and its rating class probability into the general pointwise mutual information computation (Eq. 2). Instead of a hard rating class assignment, we compute probability $P(c|d)$ for the class membership of a document d based on its star rating x using a sigmoid function, which is commonly used for mapping scores to probabilities [31]:

$$f(x) = \frac{1}{1 + a^{b-x}} \quad (3)$$

$P(c|d) = f(x)$ for $c = pos$ and $P(c|d) = 1 - f(x)$ for $c = neg$. Parameter a determines how strong the positive and negative star ratings should be discriminated, and parameter b defines the neutral star rating. In our experiments we used a setting of $a = 4$ and $b = 3$ (i.e. the median of the 5 possible distinct ratings). We extend pointwise mutual information as follows to incorporate the topic and rating probabilities:

$$\mathrm{PMI}_{ext}(w, c, z) = \log\left(\frac{\frac{1}{|D|}\sum_{d \in D} P(w|d)P(c|d)P(z|d)}{P(w) \cdot P(c) \cdot P(z)}\right) \quad (4)$$

where $|D|$ is the total number of documents, and $P(w|d)$ is estimated using a Jelinek-Mercer smoothed language model representation [45]. $P(z)$ is estimated by the fraction of terms assigned to topic z divided by the total number of terms. The equation makes use of language model, rating information, and topic mixture of the documents in our corpus, in order to determine the dependence between a term and a sentiment category for a given topic. For computing a sentiment score for an n-gram w in a given topical context, we compute $\mathrm{PMI}_{ext}(w, pos, z)$ for the positive class and $\mathrm{PMI}_{ext}(w, neg, z)$ for the negative class. The *topic-dependent sentiment value* (CDSV) is then computed as

$$\mathrm{CDSV}(w, z) = \mathrm{PMI}_{ext}(w, pos, z) - \mathrm{PMI}_{ext}(w, neg, z) \quad (5)$$

with $\mathrm{CDSV}(w, z) = 0$ indicating that, in the context of topic z, term w is neutral with respect to sentiment, and negative/positive CDSV scores indicating negative/positive sentiment. For each latent topic we can compute the CDSV for each term and accordingly identify the discriminative n-grams between the positive and

[3]For basic LDA computation we used MALLET [26]

Table 1: Sample entries from our lexicon for topic "fast food" with automatically assigned sentiment scores along with membership probabilities to this topic and average SentiWordNet scores (green: positive sentiment score; red: negative; and yellow neutral).

N-Gram	POS	Topic Probability	Sentiment Score	Average SentiWordNet Score		
				Positive	Negative	Objective
food	noun	0.035	-0.004	0.00	0.04	0.96
burger	noun	0.015	0.007	0.00	0.00	1.00
eat	verb	0.014	-0.020	0.04	0.00	0.96
sandwich	noun	0.009	0.029	0.00	0.00	1.00
meal	noun	0.007	0.013	0.00	0.00	1.00
restaurant	noun	0.006	-0.001	0.00	0.00	1.00
service	noun	0.006	-0.001	0.00	0.00	1.00
fast food	adjective,noun	0.006	0.001	0.08	0.06	0.86
cheese	noun	0.004	0.031	0.00	0.00	1.00
mcdonalds	noun	0.004	-0.063	n.a.	n.a.	n.a.
burger king	noun,noun	0.003	-0.005	0.01	0.00	0.99
fresh	adjective	0.002	0.056	0.16	0.27	0.57
french fry	adjective,noun	0.002	0.037	0.00	0.00	1.00
tasty	adjective	0.001	0.054	0.62	0.25	0.12
cold	adjective	0.001	-0.056	0.15	0.37	0.48
kfc	noun	0.001	-0.110	n.a	n.a.	n.a.
variety	noun	0.001	0.073	0.15	0.08	0.77
grease	noun	0.001	-0.103	0.00	0.06	0.94
atmosphere	noun	0.001	0.066	0.00	0.00	1.00
fast	adverb	0.001	0.047	0.00	0.00	1.00
hot dog	adjective,noun	0.001	0.074	0.09	0.14	0.77
dirty	adjective	0.001	-0.139	0.08	0.47	0.45

the negative ratings. In addition to the individual sentiment lexica for each topic, we also computed a general, topic-independent sentiment lexicon using pointwise mutual information on the whole corpus (see results in Section 4.4).

Topic Lexica Examples. Some non-trivial and highly topic-dependent sample entries from our lexicon covering the "fast food" topic are shown in Table 1. The terms "burger" and "fast food" are rather neutral in this topical context whereas "grease", "cold", or "dirty" are highly negative. Interestingly, the verb "eat" is slightly negative in the fast food context. When looking at other topics, the adjective "cold" which is negative in the fast food context, can get a positive connotation. For example, our lexicon contains entries for a latent topic corresponding to "ski resorts" topic; in this context cold weather is associated with good skiing conditions and thus the term gets a positive sentiment score. Further, we found that the noun "cold" has a positive sentiment score in the context of a latent topic related to "medicine", where drugs providing relief from colds are perceived positively.

3.4 Individual CDSL Generation

As mentioned earlier, the assignment of a document to a latent topic is based on the document's topic distribution $p(z \mid d)$. To generate an individual sentiment lexicon we use the trained LDA model to infer this topic distribution. We can then combine the general topic lexica according to the individual document's topic distribution to get this document's context-dependent sentiment lexicon (CDSL).

In our experiments we aim to infer the polarity of a term in a review by considering the review text as context. To this end, we compute the context-dependent sentiment value for a term w within a document d by combining the topic lexica as follows:

$$\text{CDSV}(w, d) = \sum_{i=1}^{|Z|} p(z_i|d)\ \text{CDSV}(w, z) \tag{6}$$

Optionally, introducing additional weights, combining the score with a context-independent component, or normalizing the values could be beneficial depending on the application area of the lexicon. For example, one could use a static, context-independent sentiment lexicon such as SentiWordNet and increase or decrease its sentiment scores according to our topic-dependent analysis.

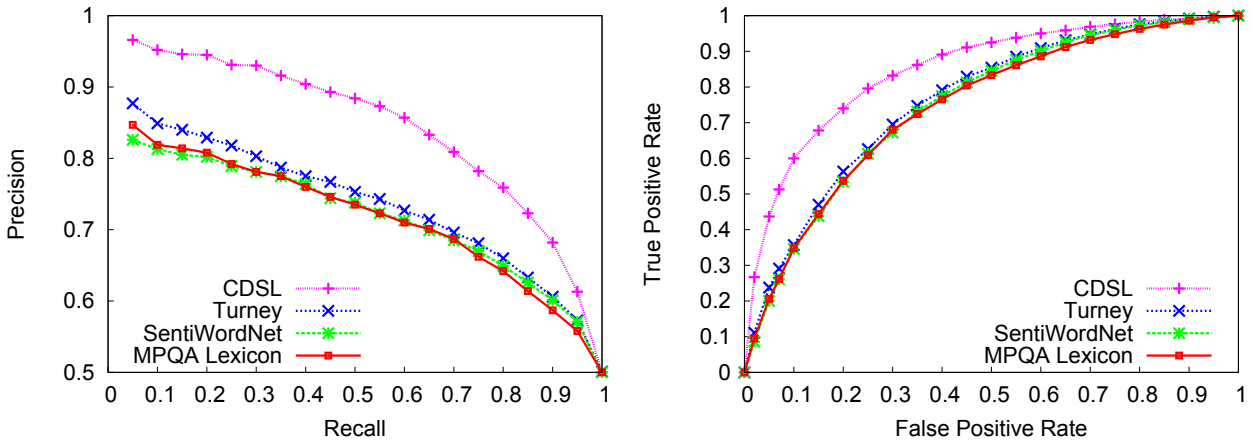
4. EXPERIMENTS

In order to evaluate our context-dependent lexicon we tested its applicability for predicting the star rating of reviews. This problem is known in the literature as sentiment classification. Methods usually include machine learning [30], sentiment lexica [36], or both [1, 9]. Note that this paper is *not* about improving the performance of sentiment classification per se; we rather use the sentiment classification task as the standard task for measuring and comparing the performance of different lexicon generation methods [21]. We therefore do not compare to sentiment classification methods that are *not* lexicon-based.

4.1 Data

For our experiments, we made use of a large dataset crawled from Epinions[4] in 2010. Epinions is a rating and review platform for a variety of different products and services, ranging from cars to football stadiums, hotels, and DVDs. Users can rate products on a five star scale and write a review about their experience to justify the given rating. The products are classified into 16 distinct categories with 318 sub-categories. We evaluated our approach using

[4] www.epinions.com



(a) Precision-Recall Curves (b) ROC Curves

Figure 3: Precision-recall curves and ROC curves for different algorithms on the Blitzer test dataset

a dataset of 27,375 reviews from 11 categories containing approx. 500 reviews for each star rating and category.

In order to compare our results with previous work we additionally used the test set designed by Blitzer et. al. [5] to evaluate domain adaption for sentiment classification. It consists of product reviews from Amazon[5] for four different product categories: books, DVDs, electronics, and kitchen appliances. Each review has an associated star rating, with reviews having a rating higher than three labeled positive and lower than three labeled negative. Reviews with a rating of 3 were discarded because their polarity was considered ambiguous. The whole dataset consists of 1,000 positive and 1,000 negative reviews for each category.

4.2 Setup for Classification

In order to evaluate our approach in a large-scale and automatic fashion, we conducted experiments on using our lexicon to classify reviews. We evaluated the effectiveness of different methods to predict the star rating associated with a review. This is a typical application scenario for sentiment lexica. Positive words in a review indicate in most cases a positive rating for the discussed product whereas negative words indicate that the author of the review was not satisfied with the product. The ground truth for classification is the star rating assigned by a user to a product, and the input data consists of the review written by this user. The algorithms should rank the given test reviews from negative to positive, and approximate the ground truth ranking as closely as possible. In our experiments we employed 5-fold cross-validation on the Epinions data described in Section 4.1. In addition, we evaluated our algorithm on the test set described in [5].

4.3 Methods

We compare the results for classifying reviews using our context-dependent sentiment lexicon (CDSL) with the results using two static, domain-independent lexica (SentiWordNet [2] and MPQA [41]), an unsupervised approach based on pointwise mutual information by Turney [37], and finally two other domain-dependent approaches by Li et al. [21] and Denecke [10].

[5]www.amazon.com

Static Lexica. For the baseline using static, domain-independent lexica, we computed a score for a review by averaging over the sentiment scores of its words. For **SentiWordNet [2]** we averaged over different WordNet synsets if a term had more than one sense. For the **MPQA Lexicon [41]** we assigned a score of 1.0 to all positive terms labeled "strongly subjective" and 0.75 to the positive terms labeled "weakly subjective" (-1.0 and -0.75 for negative terms respectively). We also computed sentiment values for bi-grams, tri-grams, etc. by averaging over the single terms of the n-grams.

Lexicon Generation using Term Co-Occurrences. This domain-independent baseline is based on **Turney [37]**. For each term in the review to be classified we computed a sentiment score by comparing the co-occurrence of this term with terms from a list of positive and negative seed words. Co-occurrence between two terms was computed based on pointwise mutual information by looking at co-occurrences on sentence level. [6] We also experimented with computing the co-occurrence on a document level but results on a sentence level proved to be superior. We discarded the original seed sets that consisted of only 7 positive and 7 negative terms in favor of a larger seed set of around 1,000 positive and 1,000 negative adjectives and adverbs described in [39]. The optimal threshold for considering a review positive or negative was determined using cross-validation since the "natural" threshold of 0.0 would have classified most reviews as positive.

Existing Context Dependent Approaches. We compare our results with two sate-of-the art systems for review classification based on lexica. The first approach from **Li et al. [21]** uses a joint sentiment and topic model to do domain-dependent analysis based on static sentiment lexica. The second approach from **Denecke [10]** is a fully supervised approach based on machine learning. For each domain a classifier is trained using, among others, SentiWordNet scores as features. Although we did not implement these approaches, we compare to the accuracy values reported in these papers using an identical experimental setup.

[6]Note that this usage of PMI is completely different from our approach described in Section 3 in that we apply PMI over distributions of terms and *sentiment categories*, instead.

Table 2: Accuracy and correlation of the original ranking with the rankings produced using our context-dependent sentiment lexicon (CDSL) and other approaches

Approach	Accuracy (+/-0.01; 95% CI)		Kendall's τ_b	
	Blitzer Test Set	Epinions Cross-Val.	Blitzer Test Set	Epinions Cross-Val.
MPQA Lexicon [41]	0.669	0.653	0.326	0.313
SentiWordNet [2]	0.687	0.681	0.345	0.327
Turney [37]	0.702	0.687	0.364	0.341
Li et al. [21]	0.690	n.a.	n.a.	n.a.
Denecke [10]	0.707	n.a.	n.a.	n.a.
CDSL	**0.775**	**0.806**	**0.481**	**0.513**

Table 3: Accuracy and correlation of the original ranking with the rankings produced using different approaches and a small training dataset from Epinions

Approach	Epinions Cross-Val.	
	Accuracy (+/-0.02; 95% CI)	Kendall's τ_b
MPQA Lexicon [41]	0.650	0.324
SentiWordNet [2]	0.679	0.312
Turney [37] Small	0.667	0.293
CDSL Small	**0.685**	**0.466**

Our Context Dependent Sentiment Lexicon. Generating our topic-specific lexicon **CDSL** as described in Section 3 provided us with a latent topic model of the data and a sentiment lexicon for each latent topic. For our experiments we use $|Z| = 75$ topics. We show in Section 4.4 how varying the number of topics influences the results. In order to compute a cumulated sentiment score for each review in the test set we had to combine different topic-specific lexica. Therefore, we needed to infer the topics for each review along with their probabilities. We then combined the sentiment lexica of different topics using the topic probabilities of the document as weights. For each n-gram in the document we assigned a score based on the sentiment values of the generated lexicon. Finally, we summed up the sentiment values for all terms w in review d, normalized by the number of terms, and obtained a sentiment score for review d.

4.4 Results

The overall results for our lexica and the comparison to baseline approaches are shown in Table 2. We evaluated the results on two datasets: A test dataset from Amazon and via cross-validation on our Epinions dataset (see Section 4.1). For ranking, we computed Kendall's τ_b [20] for each approach by sorting the test documents in a descending order from predicted negative to positive and compared the ordering with the original one based on user ratings.

Classification and ranking results show that our approach (CDSL) clearly outperforms the others for both datasets (Table 2). We consider our approach not fully supervised since we do not need new annotated training data for new datasets or domains. Compared with other entirely supervised approaches like Denecke [10], or unsupervised classification using existing lexica (e.g. Li et al. [21]) we improve the accuracy by approximately 10 percent.

Table 3 shows the influence of the size of the training data set. We generated a smaller subset of the Epinions dataset containing 7,500 reviews from 15 of the categories with 100 reviews for each star rating and category and compared the results training on the full dataset of 27,375 reviews. As expected, results improve with larger training set size. However, even for the very small training set, our method outperforms the baseline approaches.

Table 4: Break-even points and ROC-AUC values for CDSL and the baseline approaches

Approach	Blitzer Test Set		Epinions Cross-Val.	
	BEP	AUC	BEP	AUC
MPQA Lexicon [41]	0.692	0.746	0.675	0.750
SentiWordNet [2]	0.688	0.753	0.683	0.737
Turney [37]	0.697	0.768	0.688	0.757
CDSL	**0.775**	**0.849**	**0.808**	**0.882**

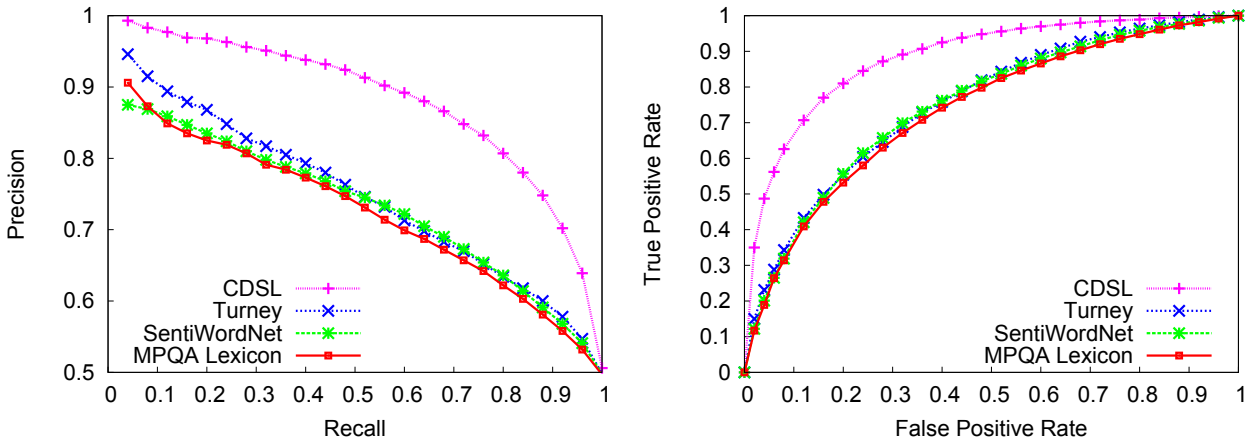
Table 5: Accuracy and correlation of the original ranking with the rankings produced using our context-dependent sentiment lexicon (CDSL) for the Blitzer test dataset

Num. of Topics	Accuracy (+/-0.01; 95% CI)	Kendall's τ_b
Orig. Categories	0.611	0.203
1 (Topic-Indep.)	0.758	0.454
25	0.746	0.442
50	0.769	0.476
75	**0.775**	**0.481**
100	0.765	0.472
200	0.766	0.469
500	0.750	0.447

Figure 3(a) shows the precision-recall curves for the binary classification task on the Blitzer test set where a review is considered positive if the rating is four or five, and negative for a rating of one or two. The break-even point (BEP) of CDSL (0.775) is substantially higher than for the other approaches (PMI is second best with a BEP of 0.697). Figure 3(b) shows the ROC curves for the same setting, with our lexicon exhibiting the best performance and a ROC-AUC value of 0.849 vs. 0.768 for PMI. For the Epinions dataset (omitted due to space constraints) results are comparable (BEP of CDSL is 0.808 vs. 0.688 for PMI Large).

Detailed BEP and ROC-AUC values are shown in Table 4. The results for the large Epinions dataset using cross-validation are shown in Figures 4(a) and 4(b). Our CDSL approach achieves a BEP of 0.808 which is considerably higher than for the other approaches.

In order to evaluate the influence of the number of latent topics, we varied this number between 25 and 500 (Table 5). The performance of our approach is rather robust with respect to the number of latent topics (accuracy between 0.746 and 0.775). Additionally, we experimented with using only one topic, i.e. making our lexicon topic-independent. In this case we rely solely on the user-assigned ratings and do not exploit the latent topics in the corpus; thus, we made only use of pointwise mutual information exploiting user ratings without taking latent topics into account. Our topic-independent lexicon exhibits good performance in this classification tasks, since most sentiment terms do not switch polarity completely across domains/topics [42]. Still, considering topics to gain topic-dependent sentiment scores improves accuracy by additional three percent. The first row of Table 5 shows the results using the original Epinions categories instead of latent topics. For each cat-



(a) Precision-Recall Curves (b) ROC Curves

Figure 4: Precision-recall curves and ROC curves for different algorithms using 5-fold cross-validation on the Epinions test dataset

egory, we built a lexicon and mapped the Amazon categories of the test set to the Epinions categories. We observe that our automatic topic extraction approach works better and does not rely on category information, which might not be available for other datasets like blogs or comments. A possible reason for the rather poor performance of using the original categories as topical context might be the size of training data per category. Instead of exploiting the whole dataset, this approach only uses the reviews for the given category, therefore even general (context-independent sentiment terms) are not recognized.

5. SUMMARY & FUTURE WORK

Contributions. We presented a novel approach for automatically assigning sentiment values to terms and n-grams based on the topical context. To the best of our knowledge, we are the first to exploit user-generated reviews and star ratings for products in combination with latent topics to build topic-specific sentiment lexica.

Granularity of our sentiment lexica. Our approach is placed in between lexica in very broad domains or context-independent ones and fine grained lexica on specific opinion-entity pairs or aspects. We determine the topic-specific polarity of words in a more flexible way using latent Dirichlet allocation.

Sentiment classification for evaluation. Our goal was not to improve the performance of sentiment classification per se; we rather used the sentiment classification task as the standard evaluation task for measuring and comparing the performance of different lexicon generation methods.

Influence of latent topics. Many sentiment terms do not switch polarity across domains/topics; therefore the topic-independent version of our approach works already quite well. However, making our PMI based lexica topic dependent clearly results in an additional performance boost.

Future Work. We are interested in testing alternative methods for discriminative analysis and feature selection within our framework. Introducing special rules for sentences containing negations might lead to further improvements. Furthermore, it might also be interesting to distinguish between general sentiment terms and sentiment terms relying on context. Finally, we believe that a promising area for future research is the combination of our method with techniques for sentiment lexicon generation based on term co-occurrences, grammatical analysis, seed sets of manually assessed sentiment terms, or glossary descriptions. We consider our method as orthogonal to these approaches in the sense that we make use of review ratings as a new source of information for sentiment lexicon generation.

Acknowledgments

This work was partially funded by the European Commission FP7 under grant agreement No. 287704 for the CUBRIK project.

6. REFERENCES

[1] A. Andreevskaia and S. Bergler. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 290–298. ACL, 2008.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 2200–2204. ELRA, 2010.

[3] D. Blei and J. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, 2007.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[5] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447. ACL, 2007.

[6] J. Bross and H. Ehrig. Generating a context-aware sentiment lexicon for aspect-based product review mining. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 435–439. IEEE CS, 2010.

[7] Y. Choi and C. Cardie. 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*, pages 590–598. ACL, 2009.

[8] Y. Choi, Y. Kim, and S.-H. Myaeng. Domain-specific sentiment analysis using contextual feature generation. In *Proc. of the 1st Intl. CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, pages 37–44. ACM, 2009.

[9] Y. Dang, Y. Zhang, and H. Chen. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25:46–53, 2010.

[10] K. Denecke. Are sentiwordnet scores suited for multi-domain sentiment classification? In *4th IEEE International Conference on Digital Information Management*, pages 33–38. IEEE, 2009.

[11] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422, 2006.

[12] A. Fahrni and M. Klenner. Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proceedings of the Symposium on Affective Language in Human and Machine*, pages 60–63, April 2008.

[13] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[14] S. Gindl, A. Weichselbraun, and A. Scharl. Cross-domain contextualisation of sentiment lexicons. In *19th European Conference on Artificial Intelligence*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 771–776. IOS Press, 2010.

[15] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, 2004.

[16] V. Jijkoun, M. de Rijke, and W. Weerkamp. Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 585–594. ACL, 2010.

[17] Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 815–824. ACM, 2011.

[18] N. Kaji and M. Kitsuregawa. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning)*, pages 1075–1083. ACL, 2007.

[19] H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. ACL, 2006.

[20] W. H. Kruskal. Ordinal measures of association. *J. of the American Statistical Association*, 53(284):814–861, 1958.

[21] F. Li, M. Huang, and X. Zhu. Sentiment analysis with global topics and local dependency. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*. AAAI Press, 2010.

[22] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu. Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 410–419. ACL, 2012.

[23] Y. Liu, X. Huang, A. An, and X. Yu. Arsa: A sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 607–614. ACM, 2007.

[24] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th International Conference on World Wide Web*, pages 347–356. ACM, 2011.

[25] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, pages 131–140. ACM, 2009.

[26] A. K. McCallum. Mallet: A machine learning for language toolkit, 2002. http://mallet.cs.umass.edu.

[27] S. Moghaddam and M. Ester. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information*, pages 665–674. ACM, 2011.

[28] S. Nowson. Scary films good, scary flights bad: Topic driven feature selection for classification of sentiment. In *Proceeding of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, pages 17–24. ACM, 2009.

[29] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, pages 751–760. ACM, 2010.

[30] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pages 79–86. ACL, 2002.

[31] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[32] G. Qiu, B. Liu, J. Bu, and C. Chen. Expanding domain sentiment lexicon through double propagation. In *Proc. of the 21st International Jont Conference on Artifical Intelligence*, pages 1199–1204. Morgan Kaufmann, 2009.

[33] L. Qu, G. Ifrim, and G. Weikum. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 913–921, Beijing, China, August 2010. ACL.

[34] E. Riloff, J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*, pages 25–32. ACL, 2003.

[35] I. Titov and R. T. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 308–316. ACL, 2008.

[36] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. ACL, 2002.

[37] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346, 2003.

[38] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785. ACL, 2010.

[39] C. Whitelaw, N. Garg, and S. Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 625–631. ACM, 2005.

[40] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[41] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, 2009.

[42] R. Xia and C. Zong. A pos-based ensemble model for cross-domain sentiment classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 614–622. AFNLP, 2011.

[43] C. Yang, K. H.-Y. Lin, and H.-H. Chen. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136. ACL, 2007.

[44] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann, 1997.

[45] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342. ACM, 2001.

[46] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. *SIGKDD Explor. Newsl.*, 6(1):80–89, 2004.