Preprint of: Graebner, D., Zanker M., Fliedl, G., Fuchs M.: Classification of Customer Reviews based on Sentiment Analysis.
19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.

Classification of Customer Reviews based on Sentiment Analysis

Dietmar Gräbner[a], Markus Zanker[b], Günther Fliedl[b] and Matthias Fuchs[c]

[a] econob Informationsdienstleistungs GmbH, Austria
dietmar.graebner@econob.com

[b] Institute of Applied Informatics
Alpen-Adria-Universität Klagenfurt, Austria
markus.zanker@uni-klu.ac.at, guenther.fliedl@uni-klu.ac.at

[c] Department of Social Sciences
Mid-Sweden University, Sweden
matthias.fuchs@miun.se

Abstract

In this paper we propose a system that performs the classification of customer reviews of hotels by means of a sentiment analysis. We elaborate on a process to extract a domain-specific lexicon of semantically relevant words based on a given corpus (Scharl et al., 2003; Pak & Paroubek, 2010). The resulting lexicon backs the sentiment analysis for generating a classification of the reviews. The evaluation of the classification on test data shows that the proposed system performs better compared to a predefined baseline: if a customer review is classified as *good* or *bad* the classification is correct with a probability of about 90%.

Keywords: Web 2.0, sentiment analysis, customer reviews, classification

1 Introduction

The degree of interactivity established by Web 2.0 applications shifted the priority of the Internet from an information source to an opinion source (Dippelreiter et al. 2007; Schmalegger & Carson, 2008). Every piece of information, whether it is a product offered in an online store or a post in your social network of choice, can be commented or rated in some way (Litvin et al., 2008; Xiang et al. 2010). Surveys show that the majority of Internet users do research on products they intend to buy (Pan et al. 2007; Vermeulen & Seegers, 2009). More precisely, 73% to 87% of consumers of product reviews within the tourism domain (e.g. hotel or restaurant reviews) denote that reviews influenced their purchase decision (Pang & Lee, 2008; Zehrer et al. 2011; Ye et al. 2011). Indeed, the exploitation of available opinions is interesting for companies as well as users (Lin, & Huang, 2006; Carson, 2008). The former may wish to automatically extract customer feedback from online sources, or emails. By contrast, the latter request a more concise representation of opinions (Bosangit et al., 2009). Sentiment analysis, typically, quantifies the degree of

Preprint of: Graebner, D., Zanker M., Fliedl, G., Fuchs M.: Classification of Customer Reviews based on Sentiment Analysis.
19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.

positivity or negativity towards the main subject of a text. Thereby it captures the subjectivity in terms of the semantic orientation associated with the constituents of a text (Taboada et al., 2011). In essence, sentiment analysis does what every user is required to do after writing a product review e.g. at amazon.com (http://amazon.com): to quantify the opinion represented by the text with stars.

The aim of this paper is to generate a reliable classification approach of customer reviews based on an existing domain-specific corpus by applying a lexicon-based sentiment analysis. The study comprises three steps: First, we build a lexicon of those text components with a semantic orientation. Second, we apply a sentiment analysis based on the lexicon in order to generate a classification of customer reviews. Finally, classification results are evaluated against a set of withheld reviews with quantitative ratings. We choose two different setups to demonstrate the flexibility of the proposed approach. A first analysis adopts the common classification scheme of the corpus and classifies reviews into five star-categories. The second analysis distinguishes between three categories only, thus, automatically identifying a positive, negative or neutral tendency, respectively. The paper is structured as follows: after briefly discussing related work (section 2), section 3 introduces the corpus (3.1), describes the process of lexicon construction (3.2) as well as the sentiment computation algorithm and classification (3.3). Evaluation results are presented in section 4. Finally, section 5 concludes and gives an outlook on further research activities.

2 Related work

A recent publication on lexicon-based sentiment analysis by Taboada et al (2011) shows the relevance of the research area for major text analysis tasks. By applying the Semantic Orientation CALculator (SO-CAL) the authors present a system performing sentiment analysis using manually created lexicons. They show that lexicon-based methods are superior to current state-of-the-art (i.e. statistically trained) classifiers. Moreover, extending lexicon content by linguistic information increases the robustness of a system, particularly, when texts stem from different domains. Both are important conclusions motivating our proposed lexicon-based approach. A survey on sentiment analysis is provided by Pang and Lee (2008). The authors focus on applications of sentiment analysis that go beyond extracting a sentiment value from a single text. Their applications range from sentiment computation towards identifying topics of a text, the visualization of sentiments as well as automatically defining the usefulness of a customer review. Pak and Paroubek (2010) use Twitter's micro-blogging service (http://twitter.com) as opinion source to generate a corpus. Although their methodology is not lexicon-based, the corpus statistics indicate that linguistic analysis of a corpus is the key for generating lexicons of superior quality: the correlation between occurrences of a certain word category and the overall rating of a text is clearly pointed. Below, we can show from corpus statistics of our study that the

Preprint of: Graebner, D., Zanker M., Fliedl, G., Fuchs M.: Classification of Customer Reviews based on Sentiment Analysis.
19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.

same correlation exists on a semantic layer. Finally, Taboada and Grieve (2004) defined an alternative corpus that might be reused for the evaluation of our sentiment analysis approach. 400 customer reviews extracted from epinions.com (http://epinions.com), each associated with a category, are classified either positive or negative. Although the corpus is rather small and the classification is binary, the subset of hotel reviews may be reused. Boiy and Moens (2009) apply machine learning techniques to classify web texts into positive, negative and neutral. In contrast to our token-based sentiment calculation the authors use the sentence as sentiment unit in order to determine the semantic orientation of a document. Gindl, Weichselbraun and Scharl (2010) focus on generating domain independent lexicons for sentiment analysis. The required disambiguation is achieved by considering the context of sentiment terms in contextualized dictionaries. The domain independent lexicon is filled with those terms identified as generic with respect to two different corpora of separate domains. Ohana and Tierney (2009) present an approach integrating data from SentiWordNet, a WordNet-based dictionary associating terms with positivity and negativity values, into the sentiment computation. Again the focus is the integration of general purpose lexicons to improve the classification performance. Contrasting their approaches we currently aim to build a domain-specific lexicon with a minimum amount of costly preprocessing steps.

3 A Lexicon-based Classification

As the vocabulary of domain-specific documents is limited we suggest that the sentiment analysis of domain-specific documents is ideally achieved through a lexicon-based method (Taboada et al. 2011). This section summarizes our proposed approach.

3.1 The corpus

The applied corpus comprises customer reviews of TripAdvisor (http://tripadvisor.com), the major web 2.0-platform with focus on travel and vacation services (O'Connor, 2008). Customers can book, rank and review hotels, flights and restaurants. The focus of the portal is to filter content based on rankings that are derived from user ratings. Thus, rankings are split into several categories, like value, rooms, location, cleanliness and sleep quality. Available rating categories are determined by the type of the reviewed object. A rating scale contains five values, ranging from '*terrible*' to '*excellent*'. These values are further referred to as *1star* to *5star*. A separate mandatory overall rating summarizes the total customer satisfaction. Finally, the natural language part of the review comprises a title and a text. The title is displayed in quotation marks and users are invited to use concise formulations, like *"We loved it and we'll be back!"*, or *"There were things I hated."*. The text is of variable size.

Preprint of: Graebner, D., Zanker M., Fliedl, G., Fuchs M.: Classification of Customer Reviews based on Sentiment Analysis.
19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.

The used corpus from TripAdvisor is restricted to reviews of hotels written in English, Each record contains a hotel category, the overall rating, the title and the review text. Furthermore, the entries of the subcategories value, rooms, location, cleanliness, check-in, service, and business are available. These subcategories may contain null values denoting that the user didn't care about that detail. In total, the corpus comprises over 80 000 reviews from various large tourism cities in different continents. Texts and ratings were automatically extracted from TripAdvisor. Finally, a subset was chosen by restricting reviews to hotels located in New York. From Table 1 emerges that the number of available ratings increases with the positivity of the overall rating.

Tab. 1. Number of reviews for hotels in New York

Class label	**1star**	**2star**	**3star**	**4star**	**5star**
#Reviews	495	613	1316	3695	4850

For the conducted sentiment analysis a sample consisting of 200 reviews was randomly taken from each class. This restriction intends to ease the process of document analysis and lexicon construction as well as to provide the same amount of training data for each class label (Pang & Lee, 2008). Sample sets were further split into a training set of 180 and a test set of 20 reviews. As the title tends to summarize the review, title and text were merged. Table 2 displays corpus statistics of the generated training set after using commercial natural language processing (NLP) applications of econob Informationsdienstleistungs GmbH (http://econob.com). Quality metrics for the components producing the results are not available. Nevertheless we assume that the results correspond to current state-of-commerce. The latter firm provides several NLP components that generate text annotations. An annotation, basically, is a markup of a text portion from the review, thus, identifying structural or semantic properties (Manning & Schuetze, 1999). The majority of annotations are structural annotations, like tokens or sentences, generated with a standard tokenizer and sentence detector. Semantic annotations are generated with several components responsible for named entity recognition. Amongst the most frequent semantic annotations are facilities, cities, position, and money amounts (Tab. 2).

Preprint of: Graebner, D., Zanker M., Fliedl, G., Fuchs M.: Classification of Customer Reviews based on Sentiment Analysis.
19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.

Table 2. Corpus statistics of documents in the training set grouped by class label.

Annotation	1star	2star	3star	4star	5star	Overall
# Character	190156	215115	229302	175941	175695	986209
# Token	40891	46435	49083	37053	36683	210145
# Annotations	48829	54595	57898	44048	44747	250117
# Sentences	2669	2933	2892	2300	2126	12920
#Document	180	180	180	180	180	900
#Facility	100	102	170	240	282	894
#Money	172	186	258	74	20	710
#City	46	47	88	70	187	438
#Position	118	101	89	55	55	418

By splitting the statistics according to their class label interesting tendencies become apparent: the size of customer reviews is highest for texts labeled with *3star*, while texts with higher ratings tend to be shorter. The occurrences of money amounts indicate a negative rating, whereas occurrences of cities imply a higher rating. Hence, even from this small sample set, certain types of entities are distinctive for a specific rating - or at least indicate a positive or negative review tendency.

3.2 Lexicon construction

As the overall sentiment value of the document is solely derived from the entries in the lexicon, its quality is the key issue in a lexicon-based sentiment analysis (Taboada et al., 2011). The lexicon in the present study was generated on the base of the vocabulary in the training set only. An entry in the lexicon is defined by a token and its part-of-speech (POS) tag. By considering no additional data sources during lexicon construction, we aim to demonstrate that generating a customized lexicon is straightforward and easy to automate. The lexicon is highly domain specific and the ensuing sentiment analysis reveals the usefulness of such a dictionary. The lexicon contains a list of tokens, each associated with a sentiment value. Values above zero denote positivity, values below zero denote negativity and zero indicates neutrality. Let us consider the following example: in the context of hotel reviews the token 'rat' is typically associated with a strong negative value. By contrast, the token 'beautiful' is clearly representing positivity. The assignment of sentiment values to concrete lexicon entries is done prior to classification and is further described in the next section. For the construction of the lexicon the meta-data generated by NLP components is used to select the relevant subset of tokens. In detail, the processing steps include a tokenizer and part-of-speech tagger. The former identifies relevant lexical units of the text, while the POS tagger assigns a word category to each token. In order to capture the domain and to generate a lexicon of significant size all verbs

Preprint of: Graebner, D., Zanker M., Fliedl, G., Fuchs M.: Classification of Customer Reviews based on Sentiment Analysis.
19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.

and nouns are considered relevant (Taboada et al., 2011). Table 3 shows an excerpt of the most frequent tokens in the training set.

Table 3. The most frequent tokens from the training set.

Token	Frequency	POS
hotel	1851	noun
room	1842	noun
staff	618	noun
location	510	noun
stay	426	verb
breakfa	354	noun

For the sake of classification, it is important to consider only those tokens in the lexicon that are discriminating between the different class labels. Thus, for each class label a separate lexicon containing the characteristic tokens is constructed. The metric used is based on the relative token frequency with respect to POS tag and the class label. Accordingly, a token is relevant for a class label *X*, if the relative frequency of the token for the class label is higher than the relative frequency of the same token for all other class labels. Moreover, an additional parameter α is used to control the size of the lexicon. That means, if α=0 each token is assigned to those class labels it occurred at least once. For α=1 each token is assigned to the class label with the highest relative frequency. Finally for α>1 each token is assigned to a class label with a relative frequency that is α times higher than the relative frequency of the same token in all other classes. Thus, for high values of α only those tokens will remain that occur only once among the five classes. This method guarantees that the lexicons are disjoint sets of tokens with a parameter $\alpha \geq 1$ (Taboada et al., 2011).

Table 4. The size of the lexicon per category.

α	1star	2star	3star	4star	5star
0	3481	3513	3565	2919	2781
1	1558	1510	1534	1026	1209
2	1197	1098	1121	805	963
3	1141	1032	1048	774	890
100	1096	1007	1007	753	829

Preprint of: Graebner, D., Zanker M., Fliedl, G., Fuchs M.: Classification of Customer Reviews based on Sentiment Analysis.
19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.

Table 4 shows the size of the lexicon per class with respect to the parameter α. Entries with α=100 show a problematic aspect of the lexicon: obviously, the majority of the entries are tokens that occurred only once in the training set. Thus, for future experiments it will be necessary to increase the number of samples in the training set. Nevertheless, the experiments described in subsequent sections are based on the lexicons generated with the parameter α=2. This value ensures that besides all tokens specific for a class, a minimum amount of overlapping tokens are included in the lexicon of one class.

The second analysis is based on a restructured training set distinguishing only between three class labels. For a broad range of applications distinguishing between positive, negative and neutral texts seems sufficient, as the difference between ratings of finer grained rating schemes is difficult to define anyway. Nevertheless, it is important to state that our proposed approach aims to be flexible with respect to defined classes. More concretely, the class label *bad* contains all texts from the categories *1star* and *2star*, the class label *neutral* comprises the reviews from *3star* and the class label *good* includes the remaining texts from *4star* and *5star*. As we are interested in the best possible classification performance for the classes *good* and *bad*, we only generate two dictionaries. An additional lexicon for the *neutral* class label reduces the other lexicons by terms indicating slightly positive or negative sentiment and thus reduces the accuracy results for our approach. Table 5 shows the size of the lexicons according to the second analysis.

Table 5. The size of the lexicon per category for a second analysis.

α	**good**	**bad**
0	4327	5216
1	2809	3983
2	2384	3455
3	2240	3225
100	2058	2947

Obviously, decreasing the number of target labels affects the size of lexicons, as the individual training corpora are larger. Again, we choose the lexicons generated with the parameter α=2 throughout the remaining sections to ensure comparability.

3.3 Classification

In order to compute a sentiment value from reviews the lexicon entries are associated with a semantic orientation (Taboada et al., 2011). Each of the five distinct lexicons and, thus, all of its entries are assigned separate values. The basic assumption is that a document without any values identified is neutral and, thus, has a sentiment value of

0. Hence, no prior probabilities for a customer review that belongs to certain class labels are considered. Values are assigned straightforward: *1star* dictionary entries are weighted negative -2, *2star* are weighted -1, *3star* are associated with the neutral 0, *4star* are slightly positive 1 and, finally, *5star* are excellent +2.

The classification function is computed based on the sentiment analysis of the documents in the training set and the lexicons generated. In order to compute the sentiment value for one document, the sum of all identified sentiment values is generated (Pang & Lee, 2008). More sophisticated algorithms (Taboada et al., 2011) are not necessary, as we currently do not account for negation or intensification. For the classification functions the average sentiment value of all documents per class label is used. Figure 1 shows the average sentiment values for all class labels. These functions can subsequently be used to classify new customer reviews by first computing the sentiment score and then calculating the distance of the score to the classes' average. The class with minimal distance to the documents' sentiment value is used as class label.

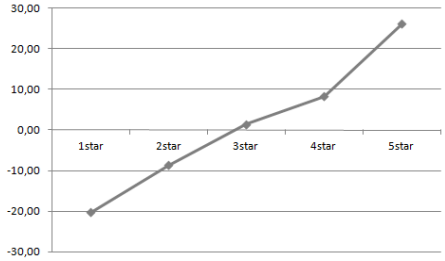

Fig. 1 Average sentiment values for five classes

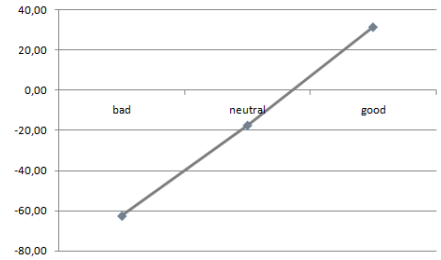

Fig. 2 Average sentiment values for three classes

Functions for different labels are well separated and should, thus, well reflect the characteristics of the different classes. The second classification function limited to three class labels is comparably computed. The entries of the lexicon with target label *bad* are associated with the value -2 and *good* are assigned with +2. Figure 2 shows the resulting classification function.

The high distance between the different class labels is mainly due to the higher distance between the sentiment values of the dictionaries as well as the larger amount of entries. In the next section the computed classification functions are evaluated.

4 Evaluation

Preceding the computation of the classification the data set was split into a training set containing 90% and a test set containing 10% of all reviews of a given class label. Thus, five distinct test sets are evaluated separately, each containing about 20

Preprint of: Graebner, D., Zanker M., Fliedl, G., Fuchs M.: Classification of Customer Reviews based on Sentiment Analysis.
19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.

documents. As evaluation metrics precision and recall measures were chosen (Manning & Schuetze, 1999). Precision defines the proportion of reviews the system classified correctly to all reviews classified. Recall describes the proportion of reviews selected correctly to all reviews selected. An additional F-measure combines both precision and recall into a single measure by computing the harmonic mean. The F-measure uses a parameter controlling the influence of precision and recall. As we have no concrete application in mind, we assume precision and recall equally important and, thus, set the weighting parameter to 0.5. The performance of the proposed system is compared to a baseline computed from randomly assigning class labels. Figure 3 summarizes the evaluation results.

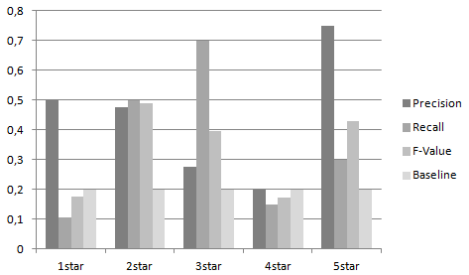

Fig. 3 P,R and F1 for five classes

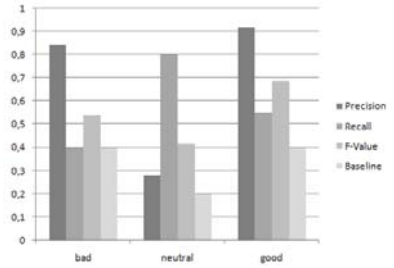

Fig. 4 P,R and F1 for three classes

The precision, recall and F-measure values over all class labels (i.e. *1star* to *5star*) are considerably higher than the baseline values. A closer look at individual performance values of the classes uncovers several interesting issues. The recall for the classes *1star* and *5star* is very low. This, clearly, is due to the high (respectively low) sentiment values needed for a document to be labeled correctly. A high (resp. low) score is achieved with a lot of either positive or negative sentiment bearing tokens identified in the document. Hence, the lexicon is much too small or not distinctive enough to encounter reviews that are not part of the training set. On the other hand, precision is highest for *5star* and *1star* because very few reviews from other classes are scored with a high value that fall into this category. Class *4star* is a negative outlier: 70% of the documents belonging to this class were incorrectly classified as *3star* documents. Again, this indicates that a very small amount of semantically relevant tokens were found, leading to the conclusion that the quality of the lexicon should be improved. Sentiment scores near 0 indicate either that the semantic orientation is neutral (positive and negative values neutralize each other) or that no sentiment indicators were found. This again reveals the weakness of the first used lexicon.

Figure 4 shows the results of the evaluation of the second classification function covering only three distinct classes. As in the dictionary generation process the

Preprint of: Graebner, D., Zanker M., Fliedl, G., Fuchs M.: Classification of Customer Reviews based on Sentiment Analysis.
19th Conference on Information and Communication Technologies in Tourism (ENTER), Springer, Helsingborg, Sweden, 2012.

reviews of multiple training sets were merged, the size of the test set for both class labels *bad* and *good* comprises 40 documents each. The obtained results for the classes *bad* and *good* are nearly similar: a precision of 84% respectively 92% signals the high quality of a positive classification. By contrast, the recall values of 40% for *bad* and 55% for *good* indicate that many examples were falsely classified as *neutral*. However, the false classification of a review belonging to the class *bad* into *good* (or the other way around) was observed only once in the test set, leading to the conclusion that those classes are very well separated. The high recall of 80% for the class label *neutral* paired with a precision of 28% approves the observation from the analysis using five classes: for false positive reviews too few tokens with semantic orientation were identified or the positive and negative tokens neutralize each other. Thus, by decreasing the distance in the classification function between class labels, the recall increases at expense of precision. Nevertheless, the average F-value of 55% is significantly higher than the computed baseline. To sum up, comparing both presented evaluations shows that decreasing the number of class labels accompanied with a larger training set and a larger lexicon would significantly increase the overall classification performance. The system was also evaluated with one lexicon per target label, but the performance was inferior to the current solution using two lexicons for *bad* and *good*.

Gindl et al. (2010) also present an evaluation based on reviews from TripAdvisor. Their evaluation only considers the classification of positive and negative reviews, ignoring neutral texts for training and testing. The classification of positive reviews has a precision of 66% and a recall of 97%. The classification of negative reviews has a precision of 95% and a recall of 46%. Our precision values are constantly high (84% and 92%) for both positive and negative reviews. Our recall values are lower because a lot of reviews are classified as neutral. Directly comparing the performance of both systems is difficult, because Gindl et al. do not consider neutral reviews. But keeping an application in mind, it is seems crucial to also consider neutral customer reviews.

5 Conclusion and future work

In our study on sentiment analysis we proposed a lexicon-based approach to classify customer reviews in the tourism domain. With precision and recall values significantly exceeding the given baseline our proposed methodology for constructing a domain specific lexicon paired with the algorithm for sentiment analysis proved to be successful. Especially the analysis using only three target labels (i.e. *good, neutral* and *bad*) may be used in real world applications to extract sentiment from text resources, as the precision for the class labels representing positive and negative sentiment showed to be remarkably high. Put simply, if a customer review is classified as *good* or *bad* the classification is correct with a probability of about 90%. Furthermore, since literature is still scarce, the outlined system performance could be

used as a new baseline for future evaluations. Finally, the study supports the definition of future research designs and optimization goals to improve the overall performance of customer classifiers based on sentiment values. In more detail, such improvements are sketched below:

First, the samples taken from the corpus determine the quality of the lexicon. The analysis of classification results and the size of the lexicons show that increasing the sample size affects the specific vocabulary used for customer reviews. This is corroborated by the fact that customer reviews are rather short documents. As a consequence, sample size should be much larger. Second, the system is backed by five, respectively three, lexicons each containing disjoint semantic indicators for corresponding class labels. However, the set of lexicons has to be extended by a domain specific lexicon relaxing the disjoint criteria combined with lexicons introducing intensifiers, downgraders and negators (Pang & Lee, 2008). Further domain independent lexica of commonly used positive or negative tokens may be integrated. These efforts will be positive for the quality of the classification at the expense of higher costs for lexicon construction. Additionally, the information gained from corpus analysis (section 3.1; e.g. the occurrence of an entity representing a city is a positive indicator), may be used to generate additional lexicons encoding encyclopedic knowledge. Experiments with the value of the parameter α should also help in optimizing the quality of the lexicon currently used. Third, the further refinement of sentiment values, currently ranging from -2 to +2, might also help to optimize the lexicon. Fourth, for the computation of the classification performance a priori probabilities for customer reviews belonging to certain categories may be considered as the corpus shows that positive reviews are much more likely than negative ones. Additional corpus statistics, such as text length, should be incorporated. Finally, the corpus grounding the analysis comprises customer reviews from hotels around the world. However, for the presented analysis only reviews for hotels located in New York were considered. Thus, further experiments should be conducted in the future to evaluate the impact of various domain and region specific parameters. As described in section 3.1 the corpus contains, besides the overall rating, a subcategory rating for each review. Extending sentiment analysis to these subcategories may similarly lead to interesting new results.

References

Bosangit, C., McCabe, S. & Hibbert, S. (2009). What is Told in Travel Blogs? Exploring Travel Blogs for Consumer Narrative Analysis. In W. Höpken, U. Gretzel & R. Law, *eds. Information and Communication Technologies in Tourism 2009*. Wien New York: Springer. pp.61-71.

Boiy, E. & Moens, M. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval* 12(5): 526-558.

Carson, D. (2008). The Blogosphere as a Market Research Tool for Tourism Destinations. *Journal of Vacation Marketing* 14(2): 111-119.

Dippelreiter, B., Grün, Chr., Pöttler, M., Seidel, I., Berger, H., Dittenbach, M. & Pesenhofer, A. (2007). Online Tourism Communities on the Path to Web 2.0 - An Evaluation, Virtual Communities in Travel and Tourism. *Information Technology & Tourism*, 10(4): 329-353.

Gindl, S., Weichselbraun, A. & Scharl, A. (2010). Cross-Domain Contextualisation of Sentiment Lexicons. *Proceeding of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*. pp. 771-776.

Lin, Y.S. & Huang, J.Y., (2006). Internet blogs as a tourism marketing medium: A case study. *Journal of Business Research*, pp.1201-05.

Litvin, S.W., Goldsmith, R.E. & Pan, B. (2008). Electronic Word-Of-Mouth in Hospitality and Tourism Management. *Tourism Management*, 29(3): 458-68.

Manning, C.D. & Schuetze, H. (1999). Foundations of Statistical Natural Language Processing, MIT Press, 1 edition.

O'Connor, P. (2008). User-Generated Content and Travel - A Case Study on Tripadvisor.Com. In O'Connor, P., Höpken, W. and Gretzel, U (Eds.), *Information and Communication Technologies in Tourism 2008*, Springer, New York, pp. 47-58.

Ohana, B. & Tierney, B. (2009). Sentiment Classification of Reviews Using SentiWordNet *Proceedings of the 9th IT&T Conference*, Dublin Institute of Technology, Dublin, Ireland.

Pak, A. & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining, *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10), Valletta, Malta.

Pan, B., MacLaurin, T. & Crotts, J. C. (2007). Travel Blogs and the Implications for Destination Marketing. *Journal of Travel Research* 46(4)): 35-45.

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis, *Foundations and Trends in Information Retrieval* 2(1-2): 1–135.

Scharl, A., Pollach, I. & Bauer, C. (2003). Determining the Semantic Orientation of Web-based Corpora. In J. Liu, Y. Cheung & H. Yin (eds.) *Intelligent Data Engineering and Automated Learning*, 4th Int. Conference, Ideal, 2003, Hong Kong, (Lecture Notes in Computer Science Vol. 2690) (pp. 840-849). Berlin: Springer.

Schmallegger, D. & Carson, D. (2008). Blogs in Tourism: Changing Approaches to Information Exchange. *Journal of Vacation Marketing* 14(2): 99-110.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K.D. & Stede, M., (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* 37(2): 267–307.

Taboada, M. & Grieve, J. (2004). Analyzing Appraisal Automatically, *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Vermeulen, I.E. & Seegers, D. (2009). Tried and Tested: The Impact of Online Hotel Reviews on Consumer Consideration. *Tourism Management*, 30(2): 123-127.

Xiang, Z. & Gretzel, U. (2010). Role of Social Media in Online Travel Information Search. *Tourism Management* 31(2):179-188

Ye, Q., Law, R., Gu, B. & Chen, W. (2011). The Influence of User-Generated Content on Traveler Behavior: An Empirical Investigation on the Effects of e-Word-Of-Mouth to Hotel Online Bookings. *Computers in Human Behavior* 27(2): 634-39.

Zehrer, A., Crotts, J.C. & Magnini, V.P. (2011). The Perceived Usefulness of Blog Postings: An Extension of the Expectancy-Disconfirmation Paradigm. *Tourism Management* 32(1): 106-13.

Acknowledgement: Special thanks to econob (http://www.econob.com) for providing infrastructure, interesting ideas and the natural language processing suite.