A customer feedback sentiment dictionary: Towards automatic assessment of online reviews

Laurens Tetzlaff ^{1*}, Katrin Rulle ², Gero Szepannek ³ and Werner Gronau ⁴

Received: 14/08/2018 Accepted: 03/03/2019

¹ University Stralsund, Germany, Stralsund University of Applied Sciences, Zur Schwedenschanze 15 18435 Stralsund Germany, E-mail: laurens.m.tetzlaff@fh-stralsund.de. Tel: +49 171 555 333 2

² University Stralsund, Germany, E-mail: katrin.rulle@fh-stralsund.de

³ University Stralsund, Germany, E-mail: gero.szepannek@hochschule-stralsund.de

⁴ University Stralsund, Germany, E-mail: werner.gronau@hochschule-stralsund.de

* Corresponding author

Coordinating editor: Andreas Kagermeier

Abstract

This paper aims at creating a tool to automatically extract online customer reviews of hospitality businesses and to assign a reliable score to them, based on a specifically created sentiment dictionary for this purpose by means of a statistical learning method. The effect of the amount of available training data and their resulting dictionaries is investigated. As such, a practical approach for applying LASSO regression in the context of online hospitality reviews is presented resulting in a sentiment dictionary of 778 terms with their associated weights trained on 20 000 reviews. It is shown that the created dictionary is able to accurately predict online review scores set by consumers, therefore highlighting the practical relevance of the proposed approach.

Keywords: Customer Reviews, Dictionary Creation, Hotel Online Reviews, LASSO, Sentiment Analysis, Text Mining

Citation: Tetzlaff, L., K. Rulle, G. Szepannek and W. Gronau (2019) A customer feedback sentiment dictionary: Towards automatic assessment of online reviews. *European Journal of Tourism Research* 23, pp. 28-39

Introduction

The internet continues to be on the rise and the possibilities and tools it provides are numerous and still growing. Nowadays, digitalisation, big data and online bookings are playing a crucial

role in our everyday life and are getting taken for granted. This is a development which cannot be unseen or neglected by any business sector but has to be taken advantage of. In the tourism industry, there are many ways

This work is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0). To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

in which the internet enables customers and providers alike to make use of the data given and information available online.

Seen from a customer's point of view, this may happen in order to gather information about the destination and accommodations prior to or during the trip, to book the whole holiday or single elements and to finally publicly evaluate and share the experiences made (Schuckert, Liu, & Law, 2015). Users of well-known online review and booking platforms such as 'TripAdvisor' and 'booking.com' have a huge interest in customer-generated content like online reviews and blog entries whose numbers have risen significantly during the last years (eMarketer, 2013). The reason for this enormous growth is given by the nature of its content: It is based on (real) customer experiences rather than promises made by the providing company. Being transmitted from customer to customer, this kind of information is perceived as more reliable than the one provided by the company itself by means of website content or advertisement (Kitingan, 2016).

User-generated content is of high value for potential customers when it comes to services and products that are mostly based on personal experiences. As those offers lack a physical appearance that would allow consumers to carefully examine them prior to the purchase, the information provided by other clients has found to be an important parameter of quality and its price/performance ratio (Ye, Law, & Gu, 2009). This applies especially in the very much service-based tourism industry (Ye, Law, Gu, & Chen, 2011). People hope to get an honest, yet subjective impression of what they can expect prior to making the experiences themselves (Schuckert *et al.*, 2015). According to a study by Gretzel and Yoo (2008) reviews published in virtual communities such as 'TripAdvisor' and those at online travel agencies and booking platforms are the most frequently used ones for this purpose. The subjects of these touristic reviews are numerous and can range from a whole country to a specific tourist attraction to a small café. By no means, the value of these online customer-based reviews is only restricted to new potential clients. It is the companies

themselves that can make great use of the huge amount of data provided about their own business and that of their competitors (Kitingan, 2016).

Online reviews are a valuable source of information for customers and providers alike. They are full of first-hand information about how the customers perceive and evaluate the services provided which highlights the strengths and possible weaknesses that may not be detected by the company itself. This valuable assessment from outside the business allows for a profound evaluation and interesting insights into customer satisfaction alongside with possible areas for improvement (Schuckert *et al.*, 2015). A thorough and precise feedback evaluation and management are of great importance for companies not only in the tourism sector. It enables a fast detection and analysis of complaints which can then be overcome by working on the deficits that were pointed out by the clients (Barclays, 2016). These complaints might address easily detectable and changeable issues, like the noise level in hotels rooms or the variety of food at the breakfast buffet. Yet, they can also refer to the 'soft assets' of a company which are not always easy to detect from the inside. The friendliness of the staff or the feeling of being an appreciated and welcomed guest can serve as examples. Due to their way of distribution and publishing, online reviews are easily accessible for both parties and can be seen as a reliable, though subjective reflection of a company's performance. Those attributes make online reviews an important basis for evaluation. Still, the enormous amount of online reviews and the information and data they contain are not easy to handle and to utilize.

Working on the improvement of the company may ultimately lead to a better overall review score which can indeed affect the number of new online bookings proportionally (Ye *et al.*, 2009). The influence of online reviews should therefore not be neglected by tourism businesses as they can have a serious effect on the economic performance of a company (Ye *et al.*, 2011). In order to facilitate the analysis of hundreds of online reviews, this paper aims at providing a tool that creates a compressed, yet in-depth insight into the

satisfaction of customers as well as the company's strengths and weaknesses. A sentiment analysis of hotel reviews provides a dictionary of words which are weighted according to their positive or negative connotation. The analysis is carried out by the statistical software R (R Core Team, 2018). To create the weighted dictionary, the written text of the online reviews, as well as the score assigned to them by the author of the review, are used. With the help of such a tool for sentiment analysis based on online reviews, companies will get access to a competent, systematic and objective yet fast evaluation of their products and services as well as the customers' emotions and attitude towards the company (Gräbner, Zanker, Fliedl, & Fuchs, 2012). An analysis of strengths and weaknesses can entail and lead to an enhancement of the company's perception by its clients. In order to ensure the reliability of the created dictionary and the validation of the model, a sufficiently large sample size of 25 118 reviews is used.

Literature Review

In the following sections, an overview shall be provided over the related research that has been done so far mainly within the domain of online hotel reviews especially focussing on sentiment analysis and text mining and ultimately limiting the scope on dictionary generation for this purpose.

Online Reviews in Hotel Context

As previously highlighted, online hotel reviews need to be regarded as an important source of value creation by consumers that can directly be accessed by companies. In this context, the typical consumer persona shifts towards the figure of a value co-creator next to the organisation (Rihova, Buhalis, Moital, & Gouthro, 2015). Consequently, the generated benefit of an online review should be paid special attention to. Research has been done not only on the relationship between online reviews and online buying behaviour as well as satisfaction but also on sentiment mining approaches, i.e. sentiment analyses and respectively sentiment classifications (Govindarajan, 2014; Schuckert *et al.*, 2015). However, in order to keep up with the rapid increase in the number of customer reviews, scalable techniques are needed to perform

such analyses (Sarvabhotla, Pingali, & Varma, 2010; Shi & Li, 2011).

Sentiment Analysis and Text Mining

As online reviews are unstructured natural language texts, the aim of a sentiment analysis is to measure their subjective content (Pröllochs, Feuerriegel, & Neumann, 2015) by giving the components of the text and thereby the whole review a sentiment orientation. Respectively, research has been conducted to further classify given reviews in terms of a binary categorisation, positive and negative opinions, or as multiclass classification, e.g. when looking at star ratings from 1-5 or including a class for neutral sentiment (Gräbner *et al.*, 2012; Kasper & Vela, 2011; Sarvabhotla *et al.*, 2010; Shi & Li, 2011).

Another aim of a sentiment analysis might consist not only in predicting ratings by classifying the reviews but also in discovering patterns of key topics related to the domain addressed within the consumers' statements. Berezina *et al.* (2016) investigated a topic related analysis to derive insightful information for management. However, subjectivity detection in reviews is a challenging task, especially since the terms in a dictionary can possess different meanings depending on the surrounding text (Calheiros, Moro, & Rita, 2017). Büschken and Allenby (2016) try to bypass this problem by using sentence-constrained Latent Dirichlet allocation (LDA).

Several approaches to perform a sentiment analysis have been evaluated so far. Mainly, they can be divided into lexicon-based approaches and machine learning based approaches (Calheiros *et al.*, 2017). When referring to the former, an existing lexicon of manually selected or rather predefined terms, which are supposed to be of relevance for the analysed domain, is used as a dictionary to evaluate the review's overall sentiment orientation. Using this approach, the quality of the lexicon is the key issue regarding the quality of the resulting sentiment analysis (Gräbner *et al.*, 2012). Hence, research such as Dang *et al.* (2010) investigates the possibilities in improving the needed lexicon towards the specific domain it is applied to by combining dictionaries containing ex-ante

selected words with both content-free as well as content-specific features discovered via machine learning techniques. Yet, using this approach challenges may arise when scalability is needed.

It has been concluded that machine learning based approaches tend to give more accurate predictions (Karsi, Zaim, & El Alami, 2017; Pröllochs *et al.*, 2015; Sarvabhotla *et al.*, 2010). A natural question based on these observations concerns the amount of text data that is needed for construction of a stable model. Various machine learning methods have been evaluated for Sentiment Analysis in different contexts such as Genetic Algorithms, Naïve Bayes, K-nearest neighbours (KNN) as well as Logistic Regression and Support Vector Machines (SVM) (Dang *et al.*, 2010; Govindarajan, 2014; Karsi *et al.*, 2017; Sarvabhotla *et al.*, 2010). Shi and Li (2011) use SVMs for a binary classification applying different term weighting strategies. They concluded that using term frequency-inverse document frequency (TF-IDF) instead of term frequency prior to the classification improves its results. Other research such as Zhang and Yu (2017) perform sentiment classification via eXtreme Gradient Boosting (XGBoost) while investigating the possibilities of deep learning in form of Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013) to derive word vectors and further word clusters that can be fed to the classification method.

Dictionary Generation

There have been several ways of constructing or accessing dictionaries for sentiment analysis. Many of the aforementioned studies refer to already existing dictionaries containing previously selected terms which serve as a baseline and in the course of their work are further edited (Calheiros *et al.*, 2017; Kasper & Vela, 2011). However, as Pröllochs *et al.* (2015) point out, using dictionaries of ex-ante selected terms can implicate two major challenges. First, such dictionaries depend on the domain-specific context they were created for and consequently, the included terms can have different connotations when put into a different context. Second, within some predefined dictionaries weighting the terms according to their relevance is not accounted

for. Thus, the authors use various Bayesian variable selection methods to construct dictionaries adjusted to the financial context their study focusses on. Others, such as Leung *et al.* (2006) employ a relative frequency based method to construct a sentiment dictionary containing information about the sentiment orientation of a term, meaning which category (positive, negative or if star rating 1-5) the term can be associated with, and its strength, i.e. the power of influence the word has on assigning a review to the respective class of the word. Also other research aims at constructing domain-specific dictionaries (Gräbner *et al.*, 2012). Yet, the dataset used in this study consists of as few as 180 reviews for dictionary construction which is quite small and the approach of utilizing a POS-tagger is not very scalable (Sarvabhotla *et al.*, 2010).

Contribution of this Work

Our research relates to the one of Pröllochs *et al.* (2015) by applying the least absolute shrinkage and selection operator (LASSO) as a statistical method to the domain of online hotel reviews. As a regression method, the algorithm assigns domain specific weights to the terms while at the same time performing an implicit variable selection which supports the creation of the dictionary. By transferring the methodology of Pröllochs *et al.* (2015) it is achieved to create a domain-specific dictionary for the specific business application.

Further, in addition to existing research, terms in the context of positive and negative feedback are treated separately. Finally, as an extension to the initial work of Pröllochs *et al.* (2015) a systematic evaluation of the required amount of data for dictionary creation based on a proper training and validation subsetting has been investigated. As a result, a dictionary for customer hospitality online reviews has been created that can be used for further applications in the business context as well as a basis for future research.

Methodology

In order to apply LASSO regression, the review data has to be gathered and pre-processed. Hence in this section, we introduce our general approach and all steps related to creating the final sentiment dictionary.

Framework

By use of the R package *rvest* (Wickham, 2016) and web-crawling, the online hotel reviews of predefined websites are scraped as represented in Figure 1. The data store information for each review such as its title, date, a positive and a negative feedback text as well as an overall rating score. Finally, the reviews are cleaned using standard pre-processing steps of text mining to end up with the common structured form of a document-term matrix (DTM) which is typical for predictive modelling in text mining (Feinerer, Hornik, & Meyer, 2008). We include the following steps of pre-processing:

- **Removal of punctuation, numbers, and emoticons.**
- **Removal of stop words.** Stop words are function words which are unimportant to the sentiment orientation of a review and hence do not add further information value to our generated dictionary (Manning & Schütze, 2005). They were eliminated using a list of words contained in the R package *tm* (Feinerer *et al.*, 2008).
- **Stemming.** This is a process that reduces words in inflected forms to their actual word stem by removing affixes (Manning & Schütze, 2005). However, it has to be noted that for some cases this process can distort the meaning of words or whole sentences in morphology-rich languages such as German which has to be considered during evaluation.

- **Feedback Separation.** In addition to former research and in order to take into account whether a term has been used in a positive or negative context, the terms within both positive and negative feedback are separated by means of capitalisation. This approach extends standard sentiment analysis and text mining approaches and allows the same term to have a different sentiment connotation depending on the category it belongs to. Within this paper, terms belonging to the negative context of a review are capitalised.

- **Weighting.** Common term frequencies are used as weights within the DTM.

Afterwards, the DTM is split up into training- and validation data (20.3% of the total data corresponding to 5118 reviews) for an independent performance evaluation. In order to obtain reliable results, the latter data are not used for model training (i.e. sparsity reduction, dictionary generation and assignment of weights) but only the training data (of 20K reviews) are used to create the dictionary. The final document-term matrices (DTM) contain all terms that appeared in the reviews with their respective frequencies. As a supplement to the work of Pröllochs *et al.* (2015) several dictionaries are created based on subsamples of the training data of different size. This allows analysing the effect of the available amount of text data on the performance of the resulting sentiment score model.

Figure 1. Steps to obtain dataset and create domain-specific sentiment dictionary

Dictionary Creation using LASSO

LASSO regression of the document-term matrix as a method for dictionary creation provides numerous advantages. Every term is a possible regressor (Pröllochs *et al.*, 2015) and, moreover, it leads to an implicit variable selection. Therefore, it is both a regularisation and a variable selection method (Hastie, Tibshirani, & Friedman, 2017) that, in addition, provides an easily interpretable result (in our case: the dictionary). The resulting dictionary consists of terms assigned with a positive or negative connotation in form of a numeric weight. To create the dictionary, the score rating of each review from the training dataset is used as a target criterion for regression which is predicted by the term frequencies contained in the document-term matrix and the following equation:

$$\text{Target Criterion: } RSS + \lambda \sum_{j=1}^p |\beta_j|$$

The regression minimizes the residual sum of squares (RSS) as a trade-off with the ℓ_1 -norm penalisation term $\lambda \sum_{j=1}^p |\beta_j|$. The latter is used to shrink the regression coefficients β_j in order to avoid overfitting of the training data (James, Witten, Hastie, & Tibshirani, 2017). As a consequence of the ℓ_1 regularisation, some coefficient estimates are shrunk to exactly zero, leading to an implicit variable selection (Tibshirani, 1996). By this kind of implicit variable selection, non-informative noise variables are removed from the dictionary, reducing the complexity of the overall model which is crucial in text mining as document-term matrices typically consist of many columns.

In our case, the full training corpus of 20 000 observations has 32 541 terms. For this reason, a preceding sparsity removal of terms that occur in less than a certain percentage of the training data is necessary (cf. Section 4). The size of the regularisation parameter λ is automatically optimized by an internal 10-fold cross validation loop on the training data according to the common '1-standard-error-rule': the simplest model (i.e. the one with the largest parameter λ) is selected where the performance is at maximum one standard error

above the best model. This approach corresponds to common machine learning practise to ensure parsimony and generalizability of the resulting model (Friedman, Hastie, & Tibshirani, 2010). The resulting coefficients β_j are the weights of terms in the created dictionary.

In our study, samples of varying sizes are drawn from the training data. For each training size, several dictionaries are created at different levels of sparsity removal. Thereby, the different performances can be compared and further inferences on the possibilities with this method and the requirements for a suitable dictionary concluded. Ultimately, the validation data have to be used for an independent performance evaluation with regards to several statistical measures: correlation, root-mean-squared-error (RMSE) and mean absolute error (MAE).

Results

Overall, 25 118 reviews are scraped from an online booking platform, namely booking.com, for various large-sized hotels located in different cities in Germany to avoid influences of local language and characteristics. The language of the statements is German. All reviews were written during the period from January 2013 to January 2018.

Each review consists of a positive and a negative part in which customers can express their opinions with respect to both sentiments separately. Further, a total rating score is assigned scaling from one to ten where a higher score equals a better evaluation and therefore a better experience by the customer. The positive and negative feedbacks are not equally distributed as some authors tend to only give either one or the other. In total, the review corpus contains 20 714 positive and 18 037 negative evaluations. The average rating has a score of 7.9.

Performance of various Dictionaries on the Validation Set

The total training corpus makes up 79.7% of the reviews; the remaining 20.3% build the validation corpus for assessing the predictive power of the developed sentiment model regarding the score. As mentioned, the training

set is further split up into samples containing varying amounts of reviews:

sample size = {200, 500, 1000, 2000, 5000, 10 000, 20 000}

Beyond, different sparsity levels are tested to restrict the number of possible term candidates which can be offered to the dictionary creation and consequently passively limiting its size, too. We test the following levels:

sparsity = {0.98, 0.99, 0.995, 0.999}

Hence, a word that appears in less than 2%, 1% etc. of the online reviews is not considered as a regressor. The following table shows the effect of sparsity-based term pre-filtering for the largest training sample of 20 000 reviews:

Table 1. Influence of different sparsity levels on number of terms in Document-Term-Matrix

sparsity	100% (all terms)	99.9%	99.5%	99%	98%
terms	32,541	1775	473	229	92

As one can observe the major part of terms occurs in less the 0.1% of the reviews. Further, note that each term corresponds to one

regression coefficient and therefore impacts on the degrees of freedom of the model. These considerations emphasize the importance of term pre-filtering using sparsity.

We conclude that the sparsity value should not be chosen too small: unless given a very small training data set, the best performance is observed for a sparsity of 0.999.

As depicted in Figure 2, with an increasing size of the training set the dictionary's predictive performance improves and it becomes more robust towards new data – the correlation for the training set aligns with the one of the validation set. In particular, at a sample size of 10 000 reviews, the dataset can be considered large enough to avoid overfitting. Thus as a result of this study, 10 000 reviews are sufficient to create a robust sentiment dictionary. Nonetheless, it is still possible to improve the correlation further by adding even more reviews. Table 2 gives a general overview of the performances of the best-generated dictionaries for each training set regarding correlation, mean absolute error and root-mean-squared error.

Final Sentiment Dictionary

The final dictionary is generated based on the largest training set of 20 000 reviews using a

Figure 2. Correlation between expressions and scores for training and validation set

Table 2. *Performance of created dictionaries on validation set evaluated on Cor, MAE, RMSE*

Criterion	Size of Training Dataset						
	200	500	1000	2000	5000	10 000	20 000
correlation	0.421	0.479	0.516	0.558	0.599	0.612	0.625
mean absolute error	1.128	1.100	1.071	1.032	0.988	0.978	0.963
root-mean-squared error	1.432	1.385	1.348	1.302	1.256	1.242	1.225

Table 3. *Most influential negative and positive terms in sentiment dictionary*

Negative Terms	Weight	Positive Terms	Weight
RENOVIERUNGSBEDURFT	-1.187	phantast	0.745
DRECKIG	-1.08	ZUGEWIES	0.617
JUGENDHERBERG	-1.026	auf	0.548
UNSAUB	-1.001	genial	0.494
ENTGEG	-0.868	traumhaft	0.49
SCHIMMEL	-0.86	VIELLEICHT	0.438
SAUBER	-0.822	perfekt	0.431
HERUNTERGEKOMM	-0.818	best	0.43
SCHMUTZIG	-0.803	PERFEKT	0.423
KOMFORT	-0.745	gepasst	0.415

sparsity of 0.999. It consists of 778 terms out of which 428 are connoted with a negative weight and respectively 350 with a positive one, underlying the ability of the LASSO to perform an implicit variable selection. Table 3 summarizes the ten strongest positive and negative terms represented, highlighting that such with a negative connotation have a higher influence on the total rating. For instance, if the term “RENOVIERUNGSBEDURFT” (meaning, that an object needs to be renovated) appears in a review, the predicted score already decreases by negative 1.187 points on average. As the term is taken from the negative category of a review, it has been capitalised during the initial text mining steps in order to differentiate it from the positive one.

As counterintuitive seem terms such as “JUGENDHERBERG” (Eng. youth hostel), “SAUBER” (Eng. clean) or “KOMFORT” (Eng. convenience). Taking a look at the composition of the dataset can give a

reasonable explanation – the reviews originate from evaluations for well-known hotels which is why a resemblance to a youth hostel can be regarded as a deficit. Additionally, the terms in question are capitalized and thus part of the negative category in a review which means, that even positive words such as “clean” or “convenience” can have a negative connotation if given in a negative review. This result shows the benefit of the chosen approach to separate different feedback categories prior to further text mining and regression.

Another aspect worth noticing are the positively associated terms belonging to the negative category of a review such as “VIELLEICHT” (Eng. maybe) which can be explained by terms having the function of relativizing negative feedback. Finally, one can notice that the terms “perfekt” and “PERFEKT” (Eng. perfect) have quite similar effects independent of the context (positive or negative) in which it is used.

Figure 3. Word Cloud representing positive connoted terms as green, negative connoted terms as red. Higher weight symbolized by thicker letters in each category respectively.

Some of the positively connoted terms in Figure 3 underline an occurring difficulty when applying stemming during text mining, particularly within the German language – sometimes, the remaining word stem cannot be traced back to its origin. “auf” as an example, can be part of many expressions and now all of those which are stemmed to this term can have the same weight association. A manual inspection of corresponding documents might help to overcome this issue during dictionary creation.

Finally, there are some terms directly linking back to relevant factors of the hotels influencing the evaluation such as “SCHIMMEL” (Eng. mildew). This can give decision-makers the necessary know-how on which aspects they need to focus on in order to improve the evaluations of their hotels and thereby also customer satisfaction.

Despite all the mentioned restrictions and issues that may arise when trying to draw conclusions from the created Word Cloud and Sentiment Dictionary, the two tools still are able to provide a solid overview on positively and negatively perceived elements of a business.

Especially the coloured Word Cloud allows the beholder to quickly assess the analysed object, as it immediately categorises the mentioned aspects into negative and positive by making use of a colour code to support the given connotations visually.

The owner of a tourism entity, e.g. a hotel, can thus gain a valuable outside view on his business. It allows the hotel owner to retrieve information from hundreds of independent reviews, which provide important feedback, in a compromised and compact manner. They can then see which actions have to be taken in order to improve the perception of their business. This can include very practical issues, such as “SCHIMMEL” (Eng. mildew) which should initiate a renovation or a reduction of humidity. It can also include issues more related to soft skills (of the staff) such as “UNFREUND” (Eng. unfriend(ly)liness). Here, actions of another kind such as the training of staff have to be taken. These rather soft skills related reviews are of great importance to the owner of a tourism entity, e.g. a hotel, since they are difficult to assess from the inside.

Therefore, a Sentiment Analysis and the entailing Word Cloud can help a business owner to see through the eyes of their clients and to perceive their business the way his clients do. This insight allows for problems to be assessed and appropriate actions to be taken in order to improve the perception of the business. Overall, we conclude from our results that accurately predicting the score of online hotel reviews is possible by means of an automatically created domain-specific sentiment dictionary based on text mining and LASSO regression.

Discussion

Our research follows up on the one done by Pröllochs *et al.* (2015) in so far as we implement LASSO regression on a different domain to create a domain-specific dictionary. However, our aim is to be able to predict review rating scores and by doing so inferring on possible factors driving such reviews which may be disguised as terms within our dictionary. Similarly to their research, LASSO regression proved to be a successful method for finding relevant terms in an unstructured big dataset.

As previously mentioned, the limited number of studies that exist within the domain of hotel reviews in the context of sentiment analyses mainly focus on binary or multiclass classification but not on the prediction of ratings on a numeric scale which adhere more detailed information. Therefore, the research is an extension to just classifying the reviews into a positive or negative category. It allows an insight into the relationship between words used and customer individual rating scores. Compared to studies using a lexicon with pre-selected words the proposed methodology holds the advantage of being able to adapt to the specific domain as well as to the words used in the online reviews and as a result creates a non-static dictionary which can be easily updated.

In contrast to Gräbner *et al.* (2012), the proposed methodology in our work applies a statistical learning approach in order to optimize weights for each term as opposed to using equal static predefined weights.

The size of 20 000 reviews used for training in this study is comparatively large. The conclusion from systematically testing sentiment models trained on smaller data indicate that the amount of data required for stable dictionary creation of about 10 000 reviews has not been reached by all mentioned studies. As Shi and Li (2011) have concluded within their area of research, working with term frequency-inverse document frequency instead of term frequency to measure the importance of words prior to performing the machine learning method can lead to improved results. The results are obtained using support vector machines (which do not perform an implicit variable selection and interpretable results as LASSO does). Therefore, this conclusion may not be directly transferable to our proposed methodology but it is worth noticing and might be a direction for future studies.

Conclusion

Hotels and other hospitality businesses are permanently assessed and evaluated by their customers via digital channels. It is crucial for the businesses to know how to understand and learn from those reviews in order to improve and adapt to the customers' wishes (Rihova *et al.*, 2015).

This paper shows that it is not necessary to scan through every single entry, but that it is possible to assess these entries with the help of an automatically created sentiment dictionary. A dictionary of this manner does not only allow to evaluate which terms and in a sense, therefore, features of a business are connoted positively or negatively, but it also assigns these written entries a numeric score which is based on the expressions used and feelings conveyed by the authors. As its main contribution, this research provides a practical approach for creating such a sentiment dictionary for German organisations in the hospitality business based on a Bayesian learning method. Moreover, the necessary size is investigated, i.e. the number of reviews of the training dataset for a stable model, deducing that 10 000 reviews are adequate.

The results of this paper are limited insofar as the dataset consists only of reviews for large hotels in Germany. Besides that, the reviews

are scraped from one specific online platform. The large sample size compared to other studies and its diversity are encouraging that it is possible to transfer the sentiment model and to predict the underlying sentiment of online reviews also for other hotels and platforms which might be an investigation topic for future research, particularly for smaller and medium-sized businesses.

In addition, future research may concern the comparison of the performance while using different modelling methodology such as SVMs, random forests, boosting or neural networks and further investigate the trade-off between interpretability and performance of the model.

References

- Barclays. (2016). The feedback economy – a £3.2bn opportunity for the UK hospitality and leisure industry. Barclays. URL: https://newsroom.barclays.com/r/3355/the_feedback_economy___a__3_2bn_opportunity_for_the_uk (Accessed on 25.07.2018).
- Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: Text Mining of Online Hotel Reviews. *Journal of Hospitality Marketing & Management*, 25(1), 1–24.
- Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, 35(6), 953–975.
- Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, 26(7), 675–693.
- Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4),
- eMarketer. (2013). With online reviews critical to travelers, marketers adjust their approach. eMarketer. URL: <https://www.emarketer.com/Article/With-Online-Reviews-Critical-Travelers-Marketers-Adjust-Their-Approach/1010013#gUmzAddkks6CbHsO.01> (Accessed on 25.07.2018).
- Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25(5), 1–54.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).
- Govindarajan, M. (2014). Sentiment analysis of restaurant reviews using hybrid classification method. *International Journal of Soft Computing and Artificial Intelligence*, 2(1), 17–23.
- Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. In Fuchs, M., Ricci, F., & Cantoni, L. (eds.) (2012). *Information and communication technologies in tourism 2012*. Wien: Springer, 460–470.
- Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. In O'Connor, P., Gretzel, U., & Höpken, W. (eds.) (2008). *Information and Communication Technologies in Tourism 2008*. Wien: Springer, 35–46.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (Second edition, corrected at 12th printing 2017). *Springer series in statistics*. New York, NY: Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R* (Corrected at 8th printing). *Springer texts in statistics*. New York, Heidelberg, Dordrecht, London: Springer.
- Karsi, R., Zaim, M., & El Alami, J. (2017). Impact of corpus domain for sentiment classification: An evaluation study using supervised machine learning techniques. *Journal of Physics: Conference Series*, 870(1), 12005.
- Kasper, W., & Vela, M. (2011). *Sentiment analysis for hotel reviews*. Paper presented at the Computational Linguistics-Applications Conference, October 17-19, 2011, Jachranka, Poland, 45–52.
- Kitingan, A. (2016). The importance of online reviews. TourismTattler. URL: <https://www.tourismtattler.com/articles/marketing/importance-online-reviews/63565> (Accessed on 25.07.2018).

- Manning, C. D., & Schütze, H. (2005). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.
- Pröllochs, N., Feuerriegel, S., & Neumann, D. (2015). Generating domain-specific dictionaries using Bayesian learning. *ECIS 2015 Completed Research Papers*, (Paper 144).
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria. URL: <https://www.R-project.org/> (Accessed on 25.07.2018).
- Rihova, I., Buhalis, D., Moital, M., & Gouthro, M.-B. (2015). Conceptualising customer-to-customer value co-creation in tourism. *International Journal of Tourism Research*, 17(4), 356–363.
- Sarvabhotla, K., Pingali, P., & Varma, V. (2010). Supervised learning approaches for rating customer reviews. *Journal of Intelligent Systems*, 19(1), 79–94.
- Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing*, 32(5), 608–621.
- Shi, H.-X., & Li, X.-J. (2011). *A sentiment analysis model for hotel reviews based on supervised learning*. Paper presented at the 2011 International Conference on Machine Learning and Cybernetics, July 10-13, 2012, Guilin, China, Vol. 3, 950–954.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.
- Wickham, H. (2016). *rvest: Easily harvest (Scrape) Web Pages -- R package version 0.3.2*. URL: <https://CRAN.R-project.org/package=rvest> (Accessed on 25.07.2018).
- Wing-ki Leung, C., Chi-fai Chan, S., & Chung, F.-I. (2006). *Integrating collaborative filtering and sentiment analysis: A rating inference approach*. Paper presented at the ECAI 2006 Workshop on Recommender Systems: in conjunction with the 17th European Conference on Artificial Intelligence, August 28-29, 2006, Riva del Garda, Italy, 62–66.
- Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1), 180–182.
- 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), 634–639.
- Zhang, X., & Yu, Q. (2017). *Hotel reviews sentiment analysis based on word vector clustering*. Paper presented at the 2nd IEEE International Conference on Computational Intelligence and Applications (ICCI), September 08-11, 2017, Beijing, 260–264.