EVALUATING CUSTOMER SATISFACTION THROUGH ONLINE REVIEWS AND RATINGS

Olivera Grljević, Teaching Assistant[1]
Zita Bošnjak, Full Professor[2]

DOI: https://doi.org/10.31410/tmt.2018.733

Abstract: *Travelers nowadays express their opinions, feelings, or (dis)satisfaction on the Web through reviews and ratings of hotels, restaurants, or other travel-related entities and services. The paper proposes a novel approach to determining the sentiment orientation of reviews with attached average numerical ratings, which usually convey both positive and negative sentiment. It is shown that the analysis and comparison of subsets of reviews with the same attached mark in terms of writing style and vocabulary can significantly reduce the size of reviews with biased sentiment polarity.*

Keywords: *sentiment analysis, opinion mining, hospitality and tourism, sentiment polarity, eWOM, ranking on-line reviews*

1. INTRODUCTION

Travelers nowadays express their opinions, feelings, or (dis)satisfaction on the Web through reviews and ratings of hotels, restaurants, or other travel-related entities and services. This content is referred to as user-generated content. The role of online reviews is twofold [1]. On the one hand, customers buying online are consulting online reviews as a decision support aid. After the purchase, these online buyers generate their own reviews, expressing their opinions. On the other hand, suppliers test their core competitiveness for attracting potential customers and secure online sales by means of an online reputation created mainly by customers' online reviews. The most influential aspect of online reviews is the word-of-mouth spread of good and bad reviews [2]. Negative or poor reviews damage the reputation of the company and the company needs to have the full awareness of online reviews and prompt online management to make appropriate improvements of products or services [3], [4]. To do so, companies should conduct a comprehensive analysis of online reviews to monitor and forecast activities and sentiment of their customers [1].

The amount of travel data grows every day as people continuously add reviews to the sheer amount of existing data. One illustrative example of the vast amount of travel data is given in [5]. The authors reported on more than 60 million members and over 170 million reviews and opinions on TripAdvisor, one of the largest and most visited travel and tourism websites. Since manual approach to extraction of user-generated content has proven to be time consuming, exhausting, and costly [6], while no one is capable of manually reading and processing such a huge information source, the need for automated approaches for processing and summarizing users' relevant information, such as opinion mining technique or sentiment analysis, has emerged.
We can consider one online review as opinion or an expression of sentiment of one person. Opinion is always directed towards a certain entity as a whole (individual, event, topic, etc.) or its particular characteristics or aspects, which are referred to as sentiment targets in [7]. Expres-

[1] University of Novi Sad, Faculty of Economics Subotica, Segedinski put 9-12, 24000 Subotica, Serbia
[2] University of Novi Sad, Faculty of Economics Subotica, Segedinski put 9-12, 24000 Subotica, Serbia

sions of sentiment can be positive, negative, or neutral, and we refer to it as sentiment polarity or sentiment orientation [7]. Sentiment analysis allows companies to analyze collection of opinions, which gives a comprehensive picture of the public opinion on a particular sentiment target [8], as opposed to the individual opinion, which reflects the subjective view of one person. Sentiment analysis 1) identifies opinion in a text, 2) determines the expressed valance (positive or negative) in subjective sentences, 3) conducts its classification based on sentiment polarity, and 4) uses opinion summarization techniques to aggregate overall public opinion and sentiment or to characterize variations in preferences over time [9], [10].

There are different approaches to sentiment analysis, Section 2.1, but all of them require clearly marked examples of either positive and negative overall reviews, or sentiment words. The term *sentiment words* is used to refer to those words that are listed in a predefined polarity dictionary. In the former case, the issue of sentiment polarity determination of learning examples is of high significance, as the whole classification process depends on the accuracy of attached positive/negative sentiment labels to reviews that play a role of learning examples. Unfortunately, labelled sets that one could use for this purpose are quite expensive to obtain and usually are publicly not available, while manual labelling is time consuming and inefficient due to the large number of learning examples that should be provided. In the case of sentiment words usage, dictionaries of words conveying positive/negative sentiment are required for sentiment analysis.

In many research authors used numerical ratings (e.g., star ratings, thumbs up, thumbs down) obtained from the reviewing site to determine sentiment orientation of review [11]. In light of the large impact of ratings on potential consumers, which is even more emphasized in the hospitality and tourism industry that provides services subjective and heterogeneous in nature [2], it is of great importance to understand how accurately the ratings reflect individual consumers' sentiments. Authors of the paper [12] emphasized that numerical ratings typically used in review systems may not be the ideal indicator of customers' perceived service quality and satisfaction. Therefore, the authors imply that the validity of using online reviews or ratings to measure customers' satisfaction levels needs to be explored more closely in the future.

According to the above mentioned, in our research we have focused on the relevance of numerical ratings for automated labeling of reviews for sentiment analysis. The rationale behind our work is that numerical rating scales, ranging from a lower bound (usually 1) to an upper bound (usually 5) corresponds to a reviewers' degree of satisfaction with the graded entity. Consequently, on the one-to-five scale, mark 1 denotes the lowest overall satisfaction, even dissatisfaction, while mark 5 denotes the highest satisfaction of the reviewer. We can straightforwardly connect these boundary values with negative and positive opinions, respectively. Reviews labelled by grading levels next to the extremes, i.e. reviews rated by 2 or 4, predominantly carry a negative or positive sentiment respectively, and rarely convey a mixed sentiment. The main problem is to determine the sentiment polarity of reviews that are marked with the middle grade, grade 3 on the scale from 1 to 5. As in opinion mining one aims to determine whether comments are positive, negative or neutral, mark 3, being in the middle of a numerical scale and therefore equally distant from the expressed dissatisfaction (mark 1) and high satisfaction (mark 5), could be easily mixed up for neutral sentiment. However, reviews labelled with mark 3 often convey a mix of both positive and negative sentiment, and should not be discarded from sentiment analysis as neutral. In our research, we have conducted a thorough analysis and cross-comparison of the reviews with each of these ratings in terms of writing style and vocabulary. Based on the gained insights reviews are labelled for polarity**.** The result of labelling is a dataset with clearly marked examples of positive and negative reviews that is used as the input

for consecutive sentiment analysis. The results of sentiment analysis, presented in the Section 3.3, point to the effectiveness of our approach.

In the sequel of this paper, we firstly described the related work on sentiment analysis, which we have grouped into two categories. The first category refers to different approaches to sentiment analysis and levels of analysis (Section 2.1); the second category includes applications of sentiment analysis as beneficiary for the tourism and hospitality sector (Section 2.2). We introduced the dataset used in this research in Section 3 and give a statistical analysis of the dataset in Section 3.1. Section 3.2 demonstrates a proposed method for determining the actual sentiment polarity of reviews labeled with an average mark of 3, based on the analysis of reviews grouped together according to the attached numerical rank, in terms of writing style and vocabulary used. Application of machine learning algorithms for sentiment analysis and interpretation of results is provided in Section 3.3, while the concluding section gives overview of the conducted research and references for future work.

2. RELATED WORK

Reviewing websites represent the major data source for sentiment analysis and opinion mining. These sites provide numerical ratings and textual reviews. Analyses based on numerical rating system are not sufficient. Only analysis of textual content uncovers nuanced opinions that are generally lost in crude numerical ratings and provides the understanding of consumers opining expressed in reviews [5], while numerical ratings can be used as sentiment indicators [11] - [14].

2.1. Approaches to sentiment analysis

There are two approaches to sentiment analysis [15]: the lexicon-based approach and the machine learning approach. Lexicon-based approach compares individual words from the sentences with the sentiment words listed in lexicons (so called seed words**)** in order to determine whether the words convey sentiment or not. Two branches of this approach are dictionary and corpus-based approach. In dictionary-based approach, initial list of sentiment words is collected manually and sentiment orientation is associated to each word. Bootstrap technique[3] is used to extend the list based on the structure of synonyms and antonyms in WordNet[4] or other online dictionaries. The process of new words inclusion continues until no new words are found in the corpora. *Corpus-based approach* identifies sentiment words specific to a certain domain. It requires a huge corpus to cover domain-specific words and requires labelled training data to produce an accurate sentiment orientation for a word. *Machine learning approach* could be through unsupervised or through supervised learning. The first compares each word of the text with positively or negatively valued word selected for a cluster center. The sentiment orientation of a review is predicted by the average semantic orientation of the words in the review.

Authors in [16] described their approach to determining the sentiment polarity of reviews. They regarded an opinion sentence as a positive/negative if there was a majority of positive/negative opinion words in a review. In the case where there was the same number of positive and negative opinion words in the sentence, they predicted the orientation using the average orientation of the closest opinion words for a feature in an opinion sentence (the so-called effective opinions) or the orientation of the previous opinion sentence. In [14], the semantic orientation of a phrase that

[3] Detailed explanation is available on: http://www.stat.rutgers.edu/home/mxie/rcpapers/bootstrap.pdf

[4] Available at https://wordnet.princeton.edu/

contain adjectives or adverbs is calculated as the mutual information between the given phrase and the word „excellent" minus the mutual information between the given phrase and the word „poor". Various supervised machine learning algorithms are used for sentiment classification. In order to train a classification algorithm, it is necessary to provide several thousand examples of labelled text for training. The labelling of text for purposes of sentiment analysis is done for sentiment polarity (whether text conveys positive, negative, or neutral sentiment) and depending on the purpose of the analysis, for some additional labels (such as the sentiment target).

Sentiment analysis is performed at one of three possible levels: document [13], [11], [17], sentence [18], [19], or word-level [19]. The document-level classification aims to classify documents to those that express a positive or negative sentiment [20], [21]. It takes into account the whole document, i.e. review, and starts from the premise that a document discusses only one topic. Having in mind that the feedback mechanism of online reviews provided after service consumption plays critical role in the online sale of the hospitality and tourism industry, [22], the document-level classification often does not provide enough detail about the prevailing consumer opinion on the various aspects of the entity being monitored [9]. Since each tourism product should be evaluated based on its own characteristics, for hospitality and tourism industry aspect-oriented sentiment analysis is more suitable [23]-[25]. An aspect-oriented analysis deals with the classification of sentiment at the sentence [16], [26], [27] or a phrase level according to various aspects of the tourism product [28]. Aspects usually correspond to arbitrary topics considered important or representative of the text that is being analyzed. The third level of sentiment analysis involves classifying a word or phrase according to the polarity of the sentiment [29] - [31].

2.2. Research on sentiment analysis in tourism and hospitality sector and impact on a business

Consumer behavior has changed significantly during the search for useful information about products and services. Customers have replaced offline sources of information with electronic word of mouth marketing (eWOM) [32]. As authors of the paper [2] define it, eWOM refers to all informal communication with consumers, particularly related to the usage or characteristics of goods and services, through Internet-based technology. Generally, in e-commerce customers have shown to believe more in opinion of other people and to trust them more than promotional campaigns of a company [33], [34], and they put equal trust in online reviews and personal recommendations of friends [35], [36]. With expansion of reviewing sites, online reviews are third influencing factor on purchase decisions after coupons and discounts [37]. The similar behavior is observed within tourism and hospitality sector. Customers have more trust in websites with reviews than in professional guides and travel agencies and perceive blogs as more credible and trustworthy than traditional marketing communications [38]. With such a strong impact on purchasing decisions online reviews affect online sales as well, and they should be treated as strategic tool in hospitality and tourism management, particularly in promotion, online sales, and management of online reputation [22], [38].

The positive relationship between customer rating (star rating and customer rating) and online sales of hotels is noted in the study conducted in [39]. Authors analyzed correlation between ratings from online travel agency and Booking.com and hotels in Paris and London. Contrary to expectations, ratings expressed through stars did not influence the sales, while according to the study a higher customer rating significantly increase online sales of hotels. A 1% increase in online customer rating increased sales per room up to 2.68% in Paris hotels and up to 2.62% in London hotels. Higher customer ratings also result in higher prices of the hotels and that the

prices of high star hotels were more sensitive to online customer ratings. Also, reviewers tend to give a higher percentage of negative comments to more expensive restaurants than to the less expensive ones as noted in [40].

Management of online reputation implies activities aimed at achieving the goals of strengthening the consumers' trust in the company, improving the visibility and recognition of brand, and the strengthening the impact on the consumer preferences [4]. Given the strong impact online reviews have on purchasing decision, as highlighted in the previous section, monitoring of consumer opinion should be one of the activities in achieving this goal. Negative reviews with low ratings are more likely to reflect real problems, as stated in [41]. Depending on how management monitors and responds to customer comments and problems recognized in them, online comments and reviews can destroy a restaurant or help secure the business's longevity [41]. Restaurant managers who respond successfully to comments in electronic forums can turn an unsatisfied customer to a loyal one.

3. APPLICATION OF SENTIMENT ANALYSIS ON TOURISM AND HOSPITALITY REVIEWS: A CASE STUDY

The growing number of sites providing consumers' ratings and feedback lead to the problem of excess of information. It is impossible to read all the data and difficult to find the relevant information for one to get an overall image. Some sites only provide a rating system (by stars or numbers) or text reviews; others also provide a combination of text reviews and ratings. Both approaches have their drawbacks: a simple number on a rating system provides insufficient information, while in a long review, users express complex opinions about more features.

Content available on reviewing sites reveal consumers emotions, excitements, critics, and concrete reasons for their (dis)satisfaction, as illustrated with the following examples taken from TripAdvisor website:

a) The hotel is in a central location, ***unfortunately*** this is where the positive comments end. The room was ***dark*** and ***needed updating***, so ***not*** that ***impressed*** initially, but the late-night club ***noise*** was ***awful***. I live on a main road, so I'm used to noise, but this was ***ridiculous***; ***banging*** techno, ***shouting*** and ***heavy*** traffic until 6.00am. I can't believe that I paid £109 for the privilege.
b) Just returned for a 3-day break at the Wordsworth Hotel with the family. All I can say is ***even better*** than last year, and that was ***excellent***. Food, rooms, facilities and service were ***brilliant***. ***Well done*** all staff and management, you made a family getaway a long weekend to remember.

Analysis of user-generated content identifies the public's affection on tourism-related entities. Insights gained are useful for examining the impact of public sentiment on the accommodation choices. The good and the bad accommodation characteristics can be aggregated, as well as sentiment focused on the kindness of staff, bed comfort, noise intensity, breakfast, or other aspects of business and used to improve problematic services in a given accommodation. Sentiment analysis is used to conduct such analysis. It requires a dataset with clearly marked examples of reviews with negative and positive sentiment. Such a dataset is used to train the sentiment analysis model.

In our research we used a dataset comprised of hotel reviews collected from the website TripAdvisor [42]. In addition to textual reviews of hotels, the dataset contains a numerical rating

for each review. Reviewers provide these ratings when submitting a review that reflects his/her overall satisfaction with the property. Numerical ratings can take the value from 1 (the least satisfied) to 5 (the most satisfied). Similar to work of authors [11], [13], [14], we used these ratings to label reviews for sentiment.

Presuming that low numerical ratings (such as 1 and 2 on the 1-5 scale) predominantly convey negative sentiment, while high numerical ratings (4 and 5 on the 1-5 scale) predominantly convey positive sentiment, our focus is on the subset of reviews marked by the average grade 3, for which it is most difficult to automatically determine the sentiment polarity. Such reviews should not be mistaken for a neutral sentiment. In our case the number of such reviews in the starting dataset is 2184 (10.66%), Table 2.

Table 1 illustrates some hotel reviews, collected from TripAdvisor, with attached numerical ratings and sentiment polarity based on the review's semantics (not its rating). It can be seen that reviews marked with ratings 2 and 3 carry mixed sentiment - part of the review text refers to a positive impression, and part of the review text refers to a negative one.

Review text	**Rating**	**Corresponding sentiment polarity**
Though the room was labeled no smoking it stunk of room deodorizer. The room to our door was about an inch and a half short. The carpet was vacuumed but was worn and dirty. Temperature control on the room AC didn't seem to function.	1	Negative
Hotel rooms are clean (bonus) Glacier Rock restaurant downstairs is very nice and bar is open late. 5th floor breakfast was horrible. Need a new location for breakfast. Hot no a/c, food was mostly prepackaged and only 4 tables.	2	Mixed
The free breakfast in the morning was about what you would expect for a hot breakfast. The seating area is VERY limited. The room was clean but the beds were not very comfortable and the toilet wasn't working properly. We didn't have to call maintenance as my boyfriend was able to "unstick" the valve. Overall, the front desk staff was very friendly and the bar attached to the hotel was nice enough.	3	Mixed
Room was clean & comfortable. The scrambled eggs at breakfast were very good...best I've had for a free breakfast. Staff were very friendly. The hot tub was perfect. We ate at the Glacier Rock 2 nights...super yummy. French onion soup was super. 26 beers on tap. Not walking distance to attractions.	4	Positive
Comfortable beds, hot shower, AC kept the room cool and no noise to wake us up. Better breakfast than you find at most hotels. Both people I spoke to at the front desk were very nice. Thank you, Mark, for taking care of us.	5	Positive

Table 1: Hotel reviews with ratings and sentiment polarity
Source: Authors based on the publicly available data on TripAdvisor website

In this paper, we present the case study in which both text reviews and numerical ratings provided by customers are taken as input data. Reviews ranked with the same rating are grouped together. In the sequel, we will refer to reviews associated with marks 1 or 2 as the highly negative or negative ones respectively, and to reviews with associated marks 4 or 5 as the positive or highly positive ones. The first two will be grouped into the negative category or negative subset of the original dataset, while the latter two will be grouped into the positive category or the positive subset. Reviews ranked by the middle grade have the most unclear sentiment polarity, so they are subject of the analysis. In the sequel of this paper, we will refer to these reviews as mixed reviews, due to the mixture of both positive and negative emotions conveyed in majority of such reviews.

The sentiment polarity of reviews ranked by 3 is determined based on the analysis and comparison of review groups in terms of writing style and vocabulary used. The approach described in the sequel enables to determine the sentiment polarity of reviews that convey both negative and positive sentiment more accurately. The empirical study showed that the number of reviews where the sentiment polarity could not be determined by our approach is significantly reduced and is of manageable size for manual classification.

The dataset we have used for our research comprises of 20491 reviews of hotels collected from TripAdvisor website[5] [42]. Besides the textual reviews, the dataset contains ratings on one-to-five scale that the authors of the reviews provide to reflect their overall satisfaction. The dataset is pre-processed in terms of removal of stop words (words common in language such as *a*, *the*), while prior the sentiment analysis we conducted tokenization (segmentation of sentences on words) and stemming (reducing inflections and reducing words to their base). Figure 1 illustrates common pre-processing steps for sentiment analysis.

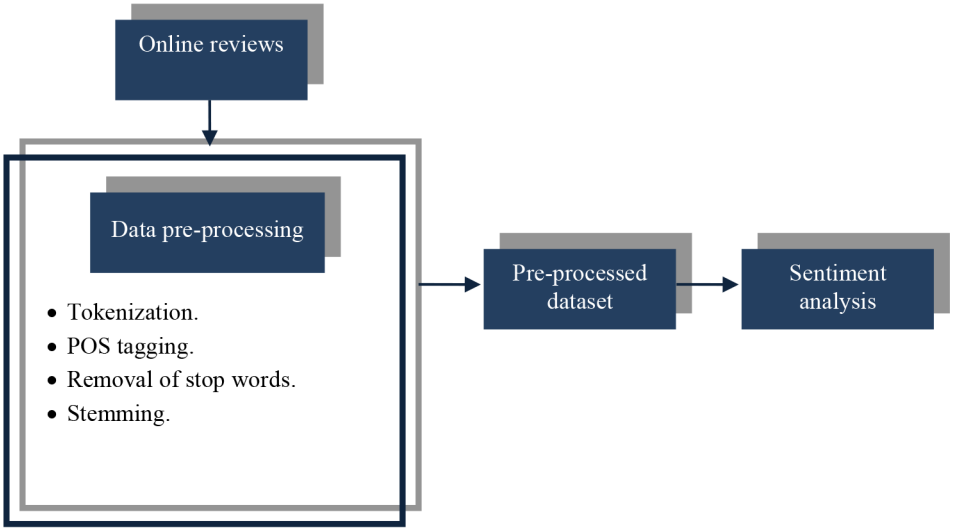

Figure 1: Data pre-processing
Source: Authors

3.1. Corpus analysis

A distribution of reviews according to ratings is presented in Table 2. A tendency to remember pleasant moments more accurately than unpleasant ones is referred to as Pollyanna Principle and it is observed in most online reviews [43]-[45], [12]. According to the authors, this tendency is associated with the fact that online reviews are predominantly positive. Distribution of reviews presented in Table 2 shows the higher frequency of ratings marked with ratings 4 and 5. This points to the applicability of Pollyanna Principle to our dataset. The imbalanced distribution of positive and negative reviews can also be attributed to euphemism and political correctness [46].

1	2	3	4	5
1421 (6.93%)	1793 (8.75%)	2184 (10.66%)	6039 (29.47%)	9054 (44.19%)

Table 2: Distribution of reviews according to ratings
Source: Authors

5 Available at https://zenodo.org/record/1219899/#.W9CbtXszapp

	1	2	3	4	5
Min	14	12	11	11	9
Max	956	1111	960	1076	1198
Median	84	97	88	80	71
Avrg	114	129	115	108	95
Std	99.80	108.23	98.51	98.48	85.29
Total	162513	230575	251623	653475	862453

Table 3: Corpus analysis according to review rating and the number of words
Source: Authors

Table 3 presents the basic statistics of the dataset. If we observe the average number of words in reviews, we can see that the lengthiest reviews are marked with lower ratings (2, 3, and 1). This tendency can be described through Negativity Bias, which is based on the idea that negative events have a stronger effect on the psychological state and human behavior [45], [47] that could lead to wordiness when restating negative events. The authors in [12] also observed that the word count in the reviews is negatively associated with the overall rating, i.e. consumers who are extremely dissatisfied tend to spend more time and energy to extensively critique the property, its amenities, and services. From the statistics provided in Table 3 we can see that reviews with ratings 4 and 5 comprise on average of less words (108 and 95, respectively) than other reviews.

Figure 2: Word cloud for mixed reviews
Source: Authors based on visualization in Voyant tool

We have analyzed the top 100 most frequent words used throughout the dataset and compared their usage between reviews with different ratings. Regardless of the review rating we could conclude that certain words are present in each category, such as *hotel*, *room*, *breakfast*, *location*, *did not*, *not*, *good*, *great*, etc. This occurrence of words among different categories of reviews is not surprising, since reviews are focused on hotels and such content is only natural. Further analysis was conducted on reviews marked with rating 3, as we indicated that this category is the most unclear in terms of sentiment polarity. Figure 2 illustrates the relevant word cloud for mixed reviews. The more prominent the word the more frequent it is in the analyzed reviews. We can see that predominantly words bear positive sentiment (*great*, *good*, *nice*, *comfortable*, etc.), while sporadically we can identify some mildly negative words (*average*, *small*, *bit*, *just*, *old*, etc.). This supports our intention to additionally explore the group of mixed reviews for sentiment polarity. Among the top 100 words that are present in mixed reviews, we

have identified 5 words which are present only in the top 100 words in this particular category of reviews: *average*, *beds*, *fine*, *things*, and *try*. We can see that these words are bearing neither excitement nor disappointment. Among the top 100 words used in mixed reviews, we have identified 11 words which are not present in the top 100 words in highly negative reviews (reviews with rating 1) while they are present in the top 100 words in negative reviews (rating 2), positive reviews (rating 4), and highly positive reviews (rating 5): *walk*, *friendly*, *price*, *quite*, *restaurant*, *view*, *right*, *best*, *buffet*, *beautiful*, *reviews*. This knowledge can be used to single out mixed reviews with the tendency to convey positive sentiment. The word *problem* is present in the top 100 words for all categories except in highly positive reviews, while word *ok* is only present in the top 100 words derived from negative and mixed reviews.

Table 4 illustrates the similarities between mixed reviews and the positive category (ratings 4 and 5), and the negative category (ratings 1 and 2). We can observe clear differences in the most common words. Words that indicate some pleasant emotion or satisfaction, such as *helpful*, *comfortable*, *excellent*, *close*, *free*, are shared between the mixed and the positive category of reviews, while words that indicate certain unpleasantries, such as *bad*, *nothing*, *old*, are shared between the mixed and the negative category of reviews.

Frequent words	Positive reviews (4 and 5)	Negative reviews (1 and 2)
bit	yes	no
free	yes	no
helpful	yes	no
street	yes	no
comfortable	yes	no
excellent	yes	no
close	yes	no
large	yes	no
airport	yes	no
big	yes	no
bad	no	yes
nothing	no	yes
old	no	yes
star	no	yes
shower	no	yes
told	no	yes

* the full list of words shared among mixed and negative reviews is provided

Table 4: Frequent words shared between mixed reviews and positive or negative reviews
Source: Authors

Observation of common occurrence of single words is not very informative since they cannot denote the context. We have introduced a context through analysis of bigrams and trigrams, which represent two or three adjoin words derived from an analyzed text. Table 5 illustrates the percentage of shared bigrams, or trigrams, between mixed reviews and other categories of reviews. Mixed reviews share 96% of identified bigrams with positive reviews, 90.56% with highly positive reviews, and more than 80% of identified trigrams are shared within both categories of positive reviews. As can be seen, these percentages are significantly smaller when we consider the negative category of reviews. Based on the provided statistics we can conclude that mixed reviews have more in common with positive and highly positive reviews compared to the similarity to negative and highly negative reviews.

	highly negative reviews (rating 1)	negative reviews (rating 2)	positive reviews (rating 4)	highly positive reviews (rating 5)
% of shared bigrams	33.11%	56.26%	96.00%	90.56%
% of shared trigrams	19.63%	37.38%	83.18%	81.31%

Table 5: Percentage of shared n-grams between mixed reviews and other categories of reviews
Source: Authors

We have identified 1525 different bigrams that occur at least 10 times in the mixed reviews and analyzed their relation with bigrams identified in other categories of reviews. Out of 1525 bigrams, 441 are present in all categories of reviews, such as: *great location*, *staff friendly*, *good location*, *room clean*, *walking distance*. These bigrams describe common aspects evaluated by travelers. Through similarity check, we have identified 32 bigrams that occurred only in mixed reviews. They indicate lack of excitement or average impression and sentiment, such as: *location average*, *room hot*, *bad place*, *rooms ok*, *overall not*. Similarities between negative reviews and mixed reviews (7 shared bigrams) are reflected through word choice which indicates criticism but with lack of enthusiasm and can be illustrated through the following bigrams: *ok hotel*, *better days*, *lot desired*, *bed hard*, *not allowed*, *not pleasant*, *not fun*. Mixed reviews exhibit more similarities with positive reviews (55 shared bigrams) than with negative reviews which can be illustrated through following bigrams: *average hotel*, *decent hotel*, *breakfast ok*, *basic clean*, *just average*, *little better*, *rooms average*, *rooms basic*, *mixed fillings*, *little disappointed*, *no bar*, *quite noisy*, *not luxurious*, *small bed*, etc. All these bigrams indicate basic satisfaction without any signs of excitement. The number of bigrams never occur in highly negative reviews, but they occur more than 10 times in all other categories of reviews is 387 (such as *good value*, *location great*, *friendly helpful*, etc.), while there are 20 bigrams which never occur in highly positive reviews (such as *not return*, *room ok*, *room tiny*, etc.), but occur in other categories. Mixed reviews share 537 different bigrams with the positive category of reviews (ratings 4 and 5). These bigrams have never occurred in the negative category of reviews (ratings 1 and 2). On the other hand, mixed reviews share only 8 bigrams with the negative category of reviews (rating 1 and 2) which have never occurred in the positive category. Table 6 illustrates 10 most frequent bigrams shared with the positive category of reviews, and 8 bigrams shared with the negative category.

We have identified 108 different trigrams that occur more than 10 times in the mixed reviews and analyzed their relation with trigrams identified in other categories of reviews. Out of 108 trigrams, 16 are present in all categories of reviews, such as: *king size bad*, *did not work*, *did not know*, *did not want*, *room not ready*. We have identified 11 trigrams that are found only in mixed reviews. As with bigrams, they indicate lack of excitement or average impression and sentiment, such as: *ok nothing special*, *did not help*, *not bad place*, *not bad hotel*. Only two trigrams are shared between mixed reviews and negative reviews. They indicate that advertised rating of hotel is not in line with the services. Four trigrams are shared between mixed reviews and positive reviews and they indicate that visitors are satisfied with the location of the hotel. Trigrams which never occurred in highly negative reviews, but do occur in other categories of reviews indicate clear positive sentiment (*hotel great location*, *staff friendly helpful*, *good value money*, *hotel good location*, *10-minute walk*) or mild negativity (*no big deal*, *did not bother*, *don't speak English*). Mixed reviews share 50 different trigrams with the positive category of reviews (ratings 4 and 5) which have never occurred in the negative category (ratings 1 and 2). They indicate that visitors value the most a location and friendliness of staff. On the other hand, mixed reviews share only

2 trigrams with the negative category of reviews (rating 1 and 2) which have never occurred in the positive category. They indicate dissatisfaction with lack of hot water and air conditioning system. Table 7 illustrates the 10 most frequent trigrams shared with the positive category of reviews, and 3 trigrams shared with the negative category. The overall conclusion is that mixed reviews exhibit more similarities with the positive than with the negative category of reviews.

Frequent bigrams	Positive reviews (4 and 5)	Negative reviews (1 and 2)
clean comfortable	yes	no
great view	yes	no
no complaints	yes	no
small clean	yes	no
overall good	yes	no
really enjoyed	yes	no
helpful friendly	yes	no
clean hotel	yes	no
nice little	yes	no
hotel quite	yes	no
no help	no	yes
smelled like	no	yes
toilet paper	no	yes
better places	no	yes
food poisoning	no	yes
poor quality	no	yes
water shower	no	yes
positive note	no	yes

Table 6: Shared bigrams between mixed reviews and positive or negative reviews
Source: Authors

Frequent trigrams	Positive reviews (4 and 5)	Negative reviews (1 and 2)
flat screen TV	yes	no
hotel staff friendly	yes	no
5-minute walk	yes	no
good not great	yes	no
free Internet access	yes	no
ocean view room	yes	no
clean staff friendly	yes	no
15-minute walk	yes	no
easy walking distance	yes	no
location hotel great	yes	no
no hot water	no	yes
no air conditioning	no	yes

Table 7: Shared trigrams between mixed reviews and positive or negative reviews
Source: Authors

3.2. Review labelling

In the presented approach, we have labelled highly negative and negative reviews (ranked 1 and 2) as examples of hotel reviews with negative sentiment, rule 1 in Table 8. Highly positive and positive reviews (ranked 5 and 4) are labelled as examples of hotel reviews with positive sentiment, rule 2 in Table 8. Labelling of mixed reviews (ranked 3) with sentiment polarity is

conducted based on the sentiment score, explained in the following of this section, and the following rules:

1. If sentiment score is greater than zero sentiment polarity is positive (rule 3 in Table 8). This indicates that a mixed review is more similar to the positive category than to the negative category.
2. If sentiment score is less than zero sentiment polarity is negative (rule 4 in Table 8). This indicates that a mixed review is more similar to the negative category than to the positive category.
3. If sentiment score is equal to zero, indicating that a mixed review shares attributes with both positive and the negative category, it should be manually inspected to determine its' sentiment polarity (rule 5 in Table 8).

No.	Antecedent	Consequent
1	a review is rated with 1 or 2 by reviewer	set the polarity of the sentiment of the review as negative
2	a review is rated with 4 or 5 by reviewer	set the polarity of the sentiment of the review as positive
3	a review is rated with 3 and sentiment score is greater than zero	set the polarity of the sentiment of the review as positive
4	a review is rated with 3 and sentiment score is less than zero	set the polarity of the sentiment of the review as negative
5	a review is rated with 3 and sentiment score is equal to zero	manually label a review with sentiment polarity

Table 8: Rules for labelling reviews for sentiment polarity based on structured ratings
Source: Authors

3.2.1 Sentiment score calculation

Based on the observations provided in Section 3.1 on the identified single words, bigrams, and trigrams shared among mixed reviews and the positive/negative category of reviews, we calculated the sentiment score for each mixed review. In the calculation we could not simply use the sum of all occurrences of a common frequent word, bigram, and trigram between a mixed review and the positive/negative category as a similarity measure, as the obtained sums were not comparable with each other due to the different number of frequent words, bigrams and trigrams characteristic for the positive/negative category of reviews.

Therefore, we normalized the number of common positive/negative trigrams, bigrams and words for each review using the z-score. [48]. A z-score is the difference, expressed in standard deviations (SDs), between the value of a data point and the value of the mean of a population of data to which that data point is being compared. The formula for calculating z-scores is:

$$z(x) = \frac{(x - \mu)}{\sigma} \quad (1)$$

where x represents individual data in a set, μ is the sample mean, and σ is the sample standard deviation.
Analogously to the work of [49], [50] in financial sector for evaluating insolvency based on z-score of the sum of the equity-to-asset ratio and the return on asset ratio, we applied z-score to calculate the sentiment score of a review. In order to explain the calculation of z-scores in the context of our analysis, we refer to Table 9. We compared each review (denoted r_i in the table)

to single words, bigrams and trigrams that frequently occur in positive and negative category of reviews, resulting in six similarity values for review r_i (denoted sp_i, bp_i, tp_i, sn_i, bn_i, and tn_i respectively). If the corresponding z-score for value sp_i is calculated ($x=sp_i$), the sample set is $\{sp_1, sp_2, ... ,sp_k\}$. Z-scores for values bp_i, tp_i, sn_i, bn_i, and tn_i are calculated in a same way.

The sentiment score of a review was calculated as the sum of z-scores for sp_i, bp_i, and tp_i multiplied by 1, and the sum of z-scores for sn_i, bn_i, and tn_i multiplied by -1. This value was treated as the sentiment score used in rules 3-5 in Table 8.

Reviews	Nº of common occurrences with the positive category			Nº of common occurrence with the negative category		
	Single words	Bigrams	Trigrams	Single words	Bigrams	Trigrams
r_1	sp_1	bp_1	tp_1	sn_1	bn_1	tn_1
r_2	sp_2	bp_2	tp_2	sn_2	bn_2	tn_2
...	...	...	...	...	...	...
r_k	sp_k	bp_k	tp_k	sn_k	bn_k	tn_k

Table 9: Number of shared words, bigrams and trigrams between mixed reviews and positive or negative category of reviews
Source: Authors

Using z-scores permits a more objective and consistent overview of data. If the numbers of common words, bigrams and trigrams are observed for a mixed review and the positive/negative category of reviews, where the list of frequent words in the positive category is not of the same length as the list of frequent words for the negative category, it would be difficult to decide which number is really larger than the other. Using z-scores rather than the raw data relieves the analyst of the burden of adjusting the perception of what is the exact meaning of the number of common words, bigrams and trigrams, according to the different scales.

3.2.2 Future work on review labelling

Following the rules presented above, we have labelled mixed reviews (ranked 3) and the resulting dataset is used for sentiment analysis (Section 3.3). Although we achieved satisfactory results of sentiment analysis models by implementation of a simplest approach to labelling of reviews, certain modifications to the proposed approach could be made and we propose them as future work.

In our analysis, we used a simplest calculation of the number of common words/bigrams/trigrams between a mixed review and the positive/negative category of reviews, by counting each common word/bigram/trigram as one occurrence, and summing up all such occurrences. Each time a common word/ bigram/trigram is observed between the mixed review and the highly positive/highly negative review, it is considered as an indicator of the presence of positive/negative sentiment expression in a mixed review. Having in mind that reviews ranked by mark 2, although negative, bear some slight positive attitude (otherwise the rank would be the lowest one, namely 1), and they can express both positive and negative sentiment, the existence of a common word or bigram/trigram between a mixed review and a negative one (but not a highly negative) is a less certain indicator of negative sentiment than it is the case with highly negative reviews. By analogy, the same holds for reviews marked by 4. Therefore, the improved version of calculation of similarity degree between a mixed review and the positive/negative

category of reviews would take the above mentioned into consideration, counting each shared word/bigram/trigram between the mixed and the highly positive/negative reviews as one point, and each shared word and bi-/tri-gram between a mixed and the positive/negative reviews as less than 1 point (for e.g. as 0.75).

If some word/bigram/trigram is very frequent in the positive/negative category of reviews, it can be considered as a typical expression for conveying positive/negative sentiment. On the contrary, words/bigrams/trigrams rarely used in the positive/negative category of reviews can be considered atypical for conveying positive/negative sentiment. Having this in mind, a modification could be made in the calculation of the number of words a mixed review has in common with the positive/negative category. Instead of counting each shared word/bigram/trigram equally (adding 1 to the sum for each occurrence), a more frequent (and more typical) common word/bigram/trigram should add higher weight to the sum than a less frequent (and less typical) common word/bigram/trigram. As the most frequent words in the positive/negative categories of reviews change for each dataset, along with their frequencies, this modification would require the weights, linearly dependent on the frequencies, to be calculated for each dataset. Additionally, separate calculation could be done for highly positive reviews, for the group of positive/negative ones, and for the highly negative reviews.

Number of common words/bigrams/trigrams between mixed reviews and the positive/negative category could be calculated in a way that there is an additional reward for each common infrequent word/bigram/trigram, such that it never occurred in an opposite category of reviews (negative/positive).

3.3. Application of sentiment analysis

The task of sentiment analysis is basically a classification task: user-generated content is automatically classified based on expressed feelings, usually into the positive or the negative class [51]. In the sequel of this section, we described the results of experiments over the dataset that was previously pre-processed and labelled as described in sections 3.1 and 3.2. Experiments were conducted with different classification algorithms and approaches to sentiment analysis. Performance of each developed classification model was evaluated in terms of accuracy, precision and recall.

In order to train classification models to predict the sentiment polarity of a review, we have divided the pre-processed and labelled dataset into a training set (75% of reviews) and a test set (25% of reviews). Feature engineering, as the following step, refers to the process of creating features which machine learning algorithms use for training. We used the domain knowledge of training data to create the indexed vocabulary and then to represent each review by a histogram vector that counts the number of appearances of each word in the review [52]. This approach converts reviews from the training dataset to a matrix of word counts and is referred to as Bag-of-Words model commonly used in natural language processing (NLP) to create vectors out of text.

For our experiments, we have selected five classification algorithms, commonly used in NLP [53], [9], [54]: *Logistic Regression* (LR), *Naïve Bayes* (NB), *Support Vector Machines (SVM)*, *Random Forest* (RF), and *Xtreme Gradiant Boosting* (XGB). We applied each classification algorithm to the pre-processed dataset, using the training set in the learning phase, and the test set to measure the performance of the model. There are several measures that can be used to evaluate

performance of classification models. We used Accuracy, Precision, Recall, and F1 measures, and calculated their values on the test dataset [55], [56]. Classification *Accuracy* refers to the overall success rate and it represents the number of samples (reviews) that are correctly classified over the total number of classifications [57], formula (2). *Precision* shows the relation of correctly classified samples (reviews) to the majority (positive) class against the total number of entities classified in this class by the model, formula (3). *Recall* shows which percentage of positive samples (reviews) is correctly classified (as positive), formula (4). Since *Precision* and *Recall* cannot be directly compared, *F1-Measure* is calculated as their harmonic mean, formula (5).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1-measure = \frac{2\,x\,Precision\,x\,Recall}{Precision + Recall} \tag{5}$$

where *TP* refers to the number of true positive reviews (which are correctly assigned to a positive category), *TN* refers to the number of true negative reviews (which are correctly not assigned to a negative category), *FP* is the number of false positive reviews (which are incorrectly assigned to a positive category), and *FN* is the number of false negative reviews (which are incorrectly classified to the negative category).

Table 10 illustrates the performance of each classifier over the test set in terms of *Accuracy*. We can observe that SVM classifier classifies reviews to positive and negative class with the smallest overall accuracy, 66.52% of reviews were correctly classified, while NB classifier exhibits the best overall performance, 88.54% of reviews were correctly classified. The main drawback of this measure is that we cannot conclude if the classifier exhibits worse performance over a positive or a negative class. For this purpose, we analyze *Precision*, *Recall*, and *F1-Measures* as they point to the performance of each classifier over the particular class, Table 11.

	Logistic Regression	**Naive Bayes**	**Support Vector Machines**	**Random Forest**	**Xtreme Gradient Boosting**
Accuracy	87.26%	88.54%	66.52%	80.98%	85.17%

Table 10: Overall accuracy of classifiers measured on test data
Source: Authors

Analysis of the data presented in Table 11 indicates that SVM classifier has extremely low value for Recall on the negative class (0.11) meaning that of all reviews that are truly negative; this classifier successfully identifies only 11%. All classifiers exhibit difficulties with identification of new reviews with negative sentiment. This is indicated by smaller values for Recall for the negative class than for the respective Precision. The difference among the two evaluation measures are lower for some classifiers, and more significant for the others. For e.g., with NB classification the Precision of 0.89, and the Recall of 0.78 is achieved for the negative class, while for the LR classification these values are almost the same, 0.83 and 0.82 respectively. If we want to generate an overall conclusion, we should observe the F1-measure. According to its values, the NB and LR models are the best classifiers, as they most successfully classify both positive reviews (91% and 90% respectively) and negative reviews (83% was successfully identified).

Classification algorithm	Precision		Recall		F1-measure	
	Positive	Negative	Positive	Negative	Positive	Negative
Logistic Regression	0.9	0.83	0.9	0.82	0.9	0.83
Naive Bayes	0.88	0.89	0.94	0.78	0.91	0.83
Support Vector Machines	0.66	0.86	0.99	0.11	0.79	0.19
Random Forest	0.84	0.75	0.86	0.72	0.85	0.73
Xtreme Gradient Boosting	0.86	0.83	0.91	0.75	0.89	0.79

Table 11: Evaluation measures for classifiers
Source: Authors

For additional evaluation of classifiers, it is common to use the Receiver Operating Characteristic (ROC) curve. ROC curve plots the false positive rate against the true positive rate for a number of different candidate thresholds. ROC takes values between 0 and 1 and describes how good the model is at predicting the positive class when the outcome is actually positive. However, research indicate that the use of ROC curve is not feasible in case of imbalanced data, when it shows overly optimistic view of algorithm's performance [58], or it offers deceptive visual interpretation of classification performance [59]. The main reason for optimistic representation of performance is that ROC curve uses the false positive rate. As an alternative to ROC curve, authors of [58] - [60] suggest Precision-Recall curve (PR) in which false positive rate is carefully avoided. Unlike a ROC curve, a PR curve is not necessarily monotonic across all thresholds because an increase in the threshold can decrease TP or FP [61]. The use of PR curve is particularly advised in case of largely skewed class distribution, which is the case with our dataset. Positive reviews are more frequent in the dataset than negative ones – 73.66% of reviews belong to the positive category, while 15.68% of reviews belong to the negative category. Figure 3 illustrates the PR curves for all classifiers and shows that SVM and RF classifiers have the worst performances, while LR model would be more suitable choice than NB model since the PR curve for LR classifier is mostly above other PR curves[6].

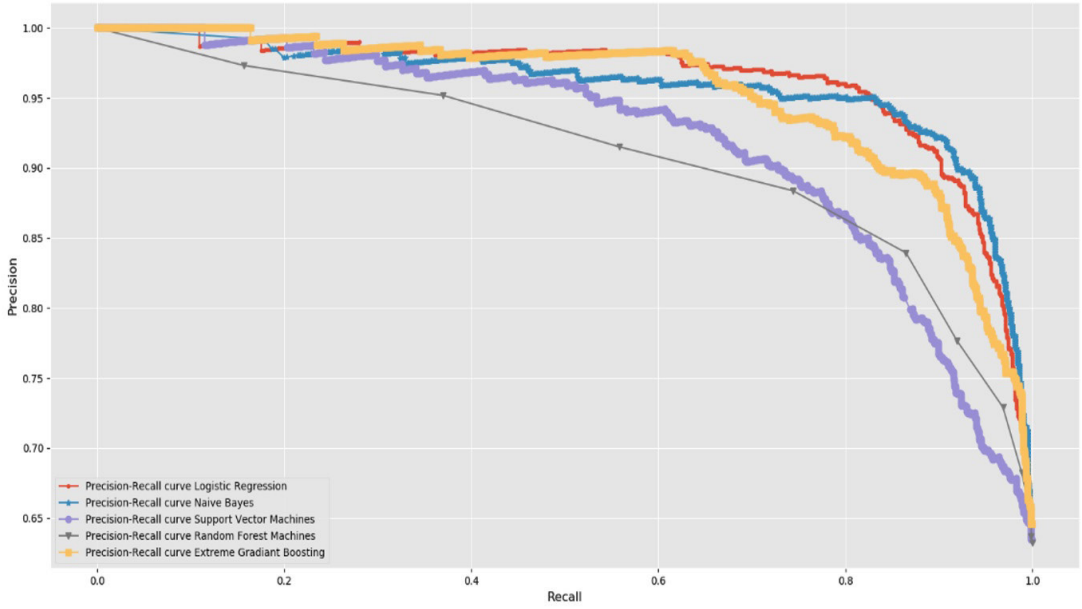

Figure 3: Precision-Recall curves
Source: Authors based on visualization in Python

6 Introduction to the precision-recall plot https://classeval.wordpress.com/introduction/introduction-to-the-precision-recall-plot/

The results of our experiments point to the possibility of successful implementation of classification algorithms on online reviews in tourism and hospitality sector even in the scenario when pre-labelled data completely lack. Our approach to labelling reviews (section 3.2) has proven to be successful given the fact that almost all chosen classifiers, except SVM, trained on the labelled data, had high performance indicators. They had successfully predicted the class a test review belonged to, from 79% to 91% of cases (see the values of F1-measure in Table 11). However, we must note that great deal of analytical skills is required for proper implementation of sentiment analysis and heavy experimentation with different data sources, classification models, parameters, and performance measures. As the results indicate, application of a different classification algorithm over the same dataset with the same pre-processing steps can result in different precision of classification, and an analyst must be skillful to make good estimates and select the appropriate classification algorithm.

4. CONCLUSION

User-generated contents on the Web, in a form of reviews and/or ratings, convey consumers' opinions and feelings towards products, services or other entities. Prospective on-line customers use this freely provided data in their decision-making process, so e-word-of-mouth spread of good and bad reviews can influence the business significantly, ranging from attracting new customers (provided the expressed sentiment in reviews is positive), to ruining the reputation of a company (for a negative sentiment). Consequently, companies should monitor the on-line activities and sentiment of their customers and promptly respond to customer comments and problems. This is even more emphasized in the hospitality and tourism industry that provides subjective services, heterogeneous in nature, where no "try before you buy" or "return in case not satisfied" features exist so the perceived risk for the consumers is even higher.

Subjects of reviews and/or ratings in the domain of tourism are travel-related entities and services. It is important to apply semantic analysis and sentiment classification techniques to study the textual content of the reviews and identify various dimensions on which consumers evaluate tourism-related entities and services. However, the number of reviews and/or ratings grows every day as people continuously generate new contents and it is impossible to process and summarize relevant information from reviews manually, but companies should use automated approaches (opinion mining technique or sentiment analysis), for which either clearly marked examples of positive and negative overall reviews, or sentiment words should be provided in advance. As this prerequisite is hard to meet due to the lack of freely available datasets marked for sentiment polarity, some authors used numerical ratings (e.g., star ratings, thumbs up, thumbs down) obtained from the reviewing site to determine the sentiment orientation of a review. However, there is evidence in literature that numerical ratings may not be the ideal measure of customers' (dis)satisfaction. In the paper we described a simple and yet effective method for determining the sentiment polarity of online reviews with attached numerical ratings.

The proposed approach is based on the similarity check in terms of vocabulary and writing style among three categories of reviews: the positive category, comprising of reviews associated with high numerical ranks (4 and 5), the negative category, comprising of reviews associated with low numerical ranks (1 and 2), and the mixed category consisting of reviews that convey both positive and negative sentiment (with the associated rank of 3).

The statistical analysis of the dataset revealed that the overall distribution of the reviews in TripAdvisor tends to be heavily skewed towards the positive ratings, while negative reviews are typically associated with greater word count than positive reviews. It is useful to focus specifically on these negative reviews since they are most likely to spread negative WOM about the tourism facility. The study also explored word choices of reviewers scoring tourism-related entities at lower ratings versus higher ratings. The finding can be helpful to service providers in uncovering patterns of deficiencies in the service standards and delivery and to act accordingly. Therefore, using the dimensions uncovered through analysis of textual contents, both positive and negative, service providers can better position themselves and target the needs and preferences of their customers.

The insights gained through corpus analysis were formulated in a form of rules for sentiment polarity determination. The result of the labelling following the devised rules (section 3.2) was a dataset with clearly marked examples of positive and negative reviews for sentiment analysis.

In order to investigate whether the results of labelling can be used as valuable help in data preparation for sentiment analysis, i.e. for building a successful classification model for sentiment analysis or not, several experiments have been conducted by five classification algorithms, commonly used in NLP: *Logistic Regression*, *Naïve Bayes*, *Support Vector Machines*, *Random Forest* and *Xtreme Gradient Boosting*. The best results were obtained with the NB and LR, as they exhibit high classification performance on both positive and negative reviews (above 90% and 80% correctly classified reviews, respectively). The results of our experiments described in section 3.3 point to the possibility of successful implementation of classification algorithms on online reviews in tourism and hospitality sector even in the scenario when pre-labeled data completely lack.

Once a satisfactory classification model of sentiment analysis has been developed it can have various business applications, and be beneficial for both customers and hotel managers [25], [51]. Aside from saving a lot of time and easing the decision process for consumers, such a system would also help hotel managers on improving their services based on feedback information how their hotel is seen by customers, what services they liked or disliked [25]. As pointed in [51] automated sentiment analysis models can aggregate overall satisfaction or dissatisfaction of customers by summarizing positive and negative online comments. They represent a useful tool for benchmark and analysis of public opinion towards the key competition through analysis of online reputation of competition, and public stance towards their key products, brands, or services. Furthermore, aspect-based sentiment analysis offers detailed insight into the public opinion, as it can pinpoint positive and negative aspects of products, services, or their individual characteristics. Although both positive and negative reviews enhance consumer awareness of hotels, positive reviews tend to improve overall attitude towards the hotels, and this affect is considerable for lesser-known hotels. By implementing sentiment analysis models, any business can monitor variations in public opinion through time. Gaining a better understanding of the associations within the various attributes of the properties and traveler reviews may lead to an improvement of the services provided and a decrease in the postings of negative reviews.

Although textual comments provide fine-grained information about a service provider's reputation that is likely to engender a buyer's trust in the service provider's competence and credibility, there is another interesting aspect that is revealed in analysis of on-line reviews [12]. Even when consumers have given very low ratings to a certain property, they have not completely given up on the properties, so they are willing to return to it in the future. The number of such

consumers counts for some 9% in the low reviews, and suggests that consumers are willing to give a second chance if service providers are willing to take the negative feedback into consideration and ensure that service delivery is significantly improved. Consequently, managers must develop strategies that improve consumers' perceptions of their responsiveness, i.e. willingness to take under consideration the reviews, both positive and negative, and enhance their service.

Although the paper clearly stated the importance of the application of modern data analysis tools, such as content analysis and sentiment classification techniques, and further demonstrated one possible approach in the domain of hospitality and tourism through a case study and the empirical research, there is a limitation to relying on the data from a single Web site, TripAdvisor. Reviews available on one site may not be the true representative sample of consumer opinion and also the service providers' perceived quality. To enhance the capability of generalization of the findings, similar data analysis should be applied to reviews from various sites and services. This would provide deeper insight into the issues consumers are mostly concerned with in reviews in the tourism and hospitality industry and their behavioral patterns.

REFERENCES

[1] Bing, P., Yang, Y. *Monitoring and Forecasting Tourist Activities with Big Data*. In: *Management Science in Hospitality and Tourism. Theory, Practice, and Applications.*, New York, Apple Academic Press, 2016.
[2] Litvin, S., Goldsmith, R.E., Pan, B. (2007) Electronic word-of-mouth in hospitality and tourism management. *Tourism Management,* 29 (2008), 458–468.
[3] Tripp, T. M., Grégoire, Y. (2011). When Unhappy Customers Strike Back on the Internet. *MIT Sloan Management Review, 52* (3), 37-44.
[4] Dijkmans, C., Kerkhof, P., Beukeboom, C. J. (2015). A stage to engage: Social media use and corporate reputation. *Tourism Management, 47*, 58-67.
[5] Puri, C. A., Kush, G., Kumar, N. (2017) Opinion Ensembling for Improving Economic Growth through Tourism. *Procedia Computer Science* 122, 237-244.
[6] Akehurst, G. (2009). User Generated Content: the use of Blogs for Tourism Organisations and Tourism Consumers, *Service Business* 3, 51-61. doi: 10.1007/s11628-008-0054-2.
[7] Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers.
[8] Pang, B., Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval, 2* (1-2), 1-135.
[9] Medhat, W., Hassan, A., Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*, 1093-1113.
[10] Grljevic, O., Bosnjak, Z., Svilengacin, G., Kovacevic, A. (submitted). The linguistic construction of sentiment expressions in student opinionated content: a corpus-based study.
[11] Pang, B., Lee, L., Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques*. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Volume 10. Association for Computational Linguistics, pp. 79–86.
[12] Racherla, P., Connolly, D.J., Christodoulidou, N. (2013). What Determines Consumers' Ratings of Service Providers? An Exploratory Study of Online Traveler Reviews. *Journal of Hospitality Marketing and Management*, 22(2), 135-161, DOI: 10.1080/19368623.2011.645187
[13] Dave, K., Lawrence, S., Pennock, D. M. (2003). *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*. In WWW ,03 Proceedings of

the 12th international conference on World Wide Web. ACM New York, NY, USA ©2003, pp. 519-528.

[14] Turney, P.D. (2002). 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 (ACL'02), Philadelphia, Pennsylvania, USA.

[15] Aye, Y. M., Aung, S.S. (2018). Senti-Lexicon and Analysis for Restaurant Reviews of Myanmar Text. *International Journal of Advanced Engineering, Management and Science (IJAEMS)*, 4(5). ISSN:2454-1311.

[16] Hu, M., Liu, B. (2004). Mining Opinion Features in Customer Reviews. *Proceedings of the 19th International Conference on Artificial Intelligence AAAI'04*, pp. 755-760.

[17] Bibi, M. (2017). *Sentiment Analysis at Document Level.* Retrieved from https://www.researchgate.net/publication/320729882_Sentiment_Analysis_at_Document_Level, uploaded on 31 October 2017.

[18] Dos Santos, C.N., Gatti, M. (2014). *Deep convolutional neural networks for sentiment analysis of short texts.* In COLING, pp. 69–78.

[19] Kim, S.M., Hovy, E. (2004). *Determining the Sentiment of Opinions*. In COLING ,04 Proceedings of the 20th international conference on Computational Linguistic.

[20] Tang, D., Qin, B., Liu, T. (2015). *Learning semantic representations of users and products for document level sentiment classification*. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Volume 1, pp. 1014-1023.

[21] Xu, J., Chen, D., Qiu, X., Huang, X. (2016). Cached long short-term memory neural networks for document-level sentiment classification. arXiv preprint arXiv:1610.04989.

[22] Schuckert, M.L., Law, X.R. (2014): Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel and Tourism Marketing*, 32(5), 608-621, doi:10.1080/10548408.2014.933154.

[23] Tjahyanto, A., Sisephaputra, B. (2017). The Utilization of Filter on Object-Based Opinion Mining in Tourism Product Reviews, *Procedia Computer Science* 124, 38-45.

[24] Marrese-Taylor, E., Velásquez, J.D., Bravo-Marquez, F., Matsuo, Y. (2013). Identifying Customer Preferences about Tourism Products using an Aspect-Based Opinion Mining Approach, *Procedia Computer Science* 22, 182-191, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013.

[25] Bucur, C. (2015). *Using Opinion Mining Techniques in Tourism*, 2nd Global Conference on Business, Economics, Management and Tourism, Prague, Czech Republic, *Procedia Economics and Finance*, 23, 1666-1673.

[26] Broß, J. (2013). *Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques,* Dissertation at Freie Universität Berlin, Berlin.

[27] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A. S., Hoste, V. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), pp. 19-30.

[28] Li, X., Bing, L., Li, P., Lam, W., Yang, Z. (2018). *Aspect Term Extraction with History Attention and Selective Transformation*. IJCAI 2018, Computation and Language, arXiv:1805.00760

[29] Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., Ma, S. (2014). *Explicit factor models for explainable recommendation based on phrase-level sentiment analysis*. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 83-92. ACM.

[30] Agarwal, A., Biadsy, F., Mckeown, K. R. (2009). *Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams*. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 24-32. Association for Computational Linguistics.
[31] Ikeda, D., Takamura, H., Ratinov, L.-A., Okumura, M. (2008). Learning to Shift the Polarity of Words for Sentiment Classification. Transactions of the Japanese Society for Artificial Intelligence, 25(1), 50-57.
[32] Gruen, T.W., Osmonbekov, T., Czaplewski, A.J. (2006). eWOM: The impact of customer-to-customer online know-how exchange on customer value and loyalty. *Journal of Business Research, 59*(4), 449-456.
[33] Berthon, P.R., Pitt, L.F., Plangger, K., Shapiro, D. (2012). Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy. *Business Horizons, 55*(3), 261-271.
[34] Pitt, L.F., Berthon, P.R., Watson, R.T., Zinkhan, G.M. (2002). The Internet and the birth of real consumer power. *Business Horizons, 45* (4), 7-14.
[35] Gligorijevic, B., Luck, E. (2012). *Engaging Social Customers – Influencing New Marketing Strategies for Social Media Information Sources*. In Contemporary Research on E-business Technology and Strategy, pp. 25-40. Springer Berlin Heidelberg.
[36] Park, D.-H., Lee, J., Han, I. (2007). The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement. *International Journal of Electronic Commerce, 11*(4), 125-148.
[37] Yang, C.-S., Chen, C.-H., Chang, P.-C. (2015). Harnessing consumer reviews for marketing intelligence: a domain-adapted sentiment classification approach. *Information Systems and e-Business Management, 13*(3), 403-419.
[38] Leung, D., Law, R., van Hoof, H., Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel and Tourism Marketing, 30*(1-2), 3–22.
[39] Öğüt, H., Onur, T.B.K. (2012). The influence of internet customer reviews on the online sales and prices in hotel industry, *The Service Industries Journal*, 32(2), 197 – 214.
[40] Lei, S., Law, R. (2015). Content Analysis of TripAdvisor Reviews on Restaurants: A Case Study of Macau. Journal of tourism, 16(1), 17-28.
[41] Savvas Pantelidis, I. (2010). Electronic Meal Experience: A Content Analysis of Online Restaurant Comments, *Cornell Hospitality Quarterly* 51(4), pp. 483-491, DOI: 10.1177/1938965510378574
[42] Alam, M.H., Ryu, W.-J., Lee, S., (2016). Joint multi-grain topic senti-ment: modeling semantic aspects for online reviews. *Information Sciences*, 339, 206–223.
[43] Boucher, J., Osgood, C.E. (1969). The Pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behaviour*, 8, 1-8. doi: 10.1016/S0022-5371(69)80002-2.
[44] Huang, T.H., Ho-Cheng Y., Chen, H.H. (2012). *Modeling Polyanna phenomena in Chinese sentiment analysis*. In Proceedings of COLING 2012: Demonstration papers, pp. 231-38. Mumbai, India
[45] Taboada, M. (2016) Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2, 325-347.
[46] Jing-Schmidt Z. (2007). Negativity bias in language: A cognitive-affective model of emotive intensifiers. *Cognitive Linguistics*, 18, 417-43
[47] Rozin P, Royzman EB. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5, 296-320
[48] Robert, A., Warner, M.D. (2016). *Chapter 2 - Using Z Scores for the Display and Analysis of Data*. In *Optimizing the Display and Interpretation of Data*, pp. 7-51. doi: 10.1016/B978-0-12-804513-8.00002-X

[49] Roy, A.D. (1952) Safety first and the holding of assets. *Econometrica*, 20 (3) (1952), 431-449.
[50] Mare, D.S., Moreira, F., Rossi, R. (2017). Nonstationary Z-Score measures. *European Journal of Operational Research*, 260(1), 348-358.
[51] Grljević, O., Bošnjak, Z. (2018). Sentiment analysis of customer data. *International Journal of Strategic Management and Decision Support Systems in Strategic Management*, 23(3), 38-49.
[52] Passalis, N., Tefas, A. (2018). Learning bag-of-embedded-words representations for textual information retrieval. *Pattern Recognition*, 81, 254-267.
[53] Jurafsky, D., Martin, J.H. (2018). Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Draft of September 23, 2018.
[54] Yang, Y., Loog, M. (2018). A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 83, 401-415.
[55] Baeza-Yates, R. Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison-Wesley Professional.
[56] Wiebe, J., Wilson, T., Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39*(2-3), 165-210.
[57] Hammam M. A., Elmahdy, A.N., Halawa, A.A., Youness, H.A. (2018). Improve the automatic classification accuracy for Arabic tweets using ensemble methods. *Journal of Electrical Systems and Information Technology*. Available online https://www.sciencedirect.com/science/article/pii/S2314717218300266. doi: 10.1016/j.jesit.2018.03.001.
[58] Davis, J., Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves*. ICML ,06 Proceedings of the 23rd international conference on Machine learning, p.p. 233-240. doi: 10.1145/1143844.1143874.
[59] Saito, T., Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*, 10(3): e0118432.
[60] Fawcett, T. (2004). ROC Graphs: Notes and Practical Considerations for Researchers. Kluwer Academic Publishers. Printed in the Netherlands.
[61] Ozenne, B., Subtil, F., Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, 68(8), 855-859.