What Can We Learn From Review Data?

Julia Neidhardt, Nataliia Pobiedina, and Hannes Werthner

Institute of Software Technology and Interactive Systems
Vienna University of Technology
{neidhardt,pobiedina,werthner}@ec.tuwien.ac.at

Abstract

Online reviews of tourism services provide valuable resources of knowledge not only for travelers but also for companies. Tourism operators are more and more aware that user related data should be seen as an important asset. This work-in-progress analyzes free text reviews as well as numerical ratings of group tours with the aim to characterize their relations. This is done with the help of statistical models. On the one hand, these models comprise textual attributes and sentiment scores of the reviews, based on text mining techniques and sentiment analysis respectively. On the other hand, non-textual attributes such as meta data about the tours and user related factors are included. First results imply a very moderate relationship between sentiment scores and ratings; the non-textual attributes appear to have a higher impact.

Keywords: Travel Reviews; Data Quality; Text Mining; Sentiment Analysis; User Satisfaction

1 Introduction

The tourism landscape has profoundly been affected by the Web, giving rise to new directions of research in eTourism (Werthner et al., 2014). Online communities serve as platforms for people to communicate and to interact. As a consequence, the amount of available data and user generated content has exploded. Thus, this quantity of data is a valuable resource for research because it enables to study the behavior of people as well as their interactions. Furthermore, the huge amount of data has become an important asset of tourism companies. The advantages of properly handling data are manifold: from improving customer relationship management, both in terms of attracting new travelers and maintaining existing ones, till identifying points for improvement in the business. However, new challenges arise: how to manage data and ensure its quality, how to preserve privacy, and how to mine knowledge from it.

With this development a number of techniques to analyze huge amounts of textual and relational data have emerged. Text mining methods help to analyze available content and facilitate decision making processes. Another focus of text mining is to analyze and to discover interesting patterns in texts, including trends and outliers (Aggarwal & Zhai, 2012). The term sentiment analysis refers to approaches that aim at extracting subjectivity from text either to decide whether a text is objective or subjective or whether a subjective text is positive or negative (Taboada et al., 2011).

In this work-in-progress paper we study travel related reviews with the objective to characterize the relation of the textual content of the reviews to their numerical ratings. We start with the problem statements: 1) How to compare the content of data to meta data? 2) How to relate different factors to the overall satisfaction?

We apply sentiment analysis to characterize whether a review has a positive or a negative orientation. For this, a lexical-based approach is chosen. Also in (Gräbner et al., 2012) a lexicon-based approach is applied to relate tourism related reviews to their numerical rating. However, there focus is on the construction of the lexicon, whereas we make use of an available one. In (Schmuck et al., 2013) statements about product properties of hotel reviews are extracted. Then it is tested whether those statements are subjective and, as a consequence, positive or negative. It is shown that for subjectivity recognition a lexical based approach (based on an already available wordlist of positive and negative words) performs better than machine learning techniques. In (Garcia et al., 2012) an approach is introduced that makes use of lexical data bases to calculate sentiment scores for tourism related reviews.

2 The Data

The work is done within a project with a partner company (Due to contractual commitments we do not disclose the name). This company is an online marketplace where group tours to over 200 countries world-wide can be compared and booked. On the platform, users can engage with co-travelers in so-called meets before the tour. The messages in the meets are usually short and are often written in moments when users are excited. After the tour, a traveler can leave a review, containing free text and five-star ratings for the categories guide, transportation, accommodation, meals, value for money. The text of a review can be left empty but the ratings have to be chosen: "5" is the maximum possible rating and "1" is the minimum.

Before starting the analysis of the data, we uncovered several data quality issues. After removing duplicates, empty entries, reviews that were not submitted directly via the platform and reviews with missing ratings, only 3912 reviews from the original 25265 are left. 2145 reviews in the final sample have 5 stars in all categories. On the other hand, only 155 reviews (4.0%) have no 5 star at all. For a review we know the creation date, the resp. user, and the date when the tour started. Based on the dates of submission and travel, we see how soon a review was done after the tour had finished. We consider reviews, which are submitted within 11 days as *soon*, and the rest as *late* reviews. We also know how active users are overall on the platform. We measure their activities by the amount of messages they write in the meets.

The information about tours encompasses such attributes as tour length, location, tour operator, maximum possible group size, and preferable age of the participants. A tour may span across several continents, a variety of countries and numerous cities. Out of 855 tours in our sample, there are only two tours including three continents, 26 tours across two continents, and the rest are done on one continent. To differentiate tours by their length, we group them into categories: *short* tours up to 3 days, *medium* length tours from 4 up to 11 days, *long* tours last 12-20 days, and *very long* tours with a program for more than 20 days. For the maximum group size we introduce the following categories: *small* groups with 4-15 participants, *medium* groups with 16-20 persons, and *large* groups which allow more than 20 persons. There is also a considerable amount of reviews for which maximum group size is *not indicated*.

3 The Text

In Section 2 we mention that the ratings of the tours are extremely positive. To get a more comprehensive picture, we incorporate the content into our analysis. Thereby, we design a pre-processing procedure which includes decoding html/xml codes, removing html hashtags and hyperlinks, replacing contractions and omitting non-English reviews. We also design a procedure to identify the most popular emoticons.

To obtain sentiment scores for reviews, we perform tokenization and part-of-speech (POS) tagging; then we apply SentiWordNet (Baccianella et al., 2010). However, there a word can have different positive and negative scores depending on the context. To resolve this issue, we use the average of all scores (Taboada et al., 2011). Once a negation is encountered, we swap positive and negative scores for the rest of the tokens in the sentence (Miller et al., 2011). Also, emoticons are taken into account. We assign a sentiment score of 1.0 to positive and -1.0 to negative emoticons.

Finally, we get several textual attributes for the reviews, e.g., the number of words, the number of emoticons, the number of nouns, adjectives or sentiment scores. For the extracted sentiment information in the reviews, we calculate an *effective sentiment score* as a difference of positive and negative scores per each word and sum up the calculated scores per each review. In Figure 1 we observe that most of the reviews are either neutral or slightly positive. According to the user specified ratings 87% reviews have five stars; such positivity is not reflected in the effective sentiment scores.

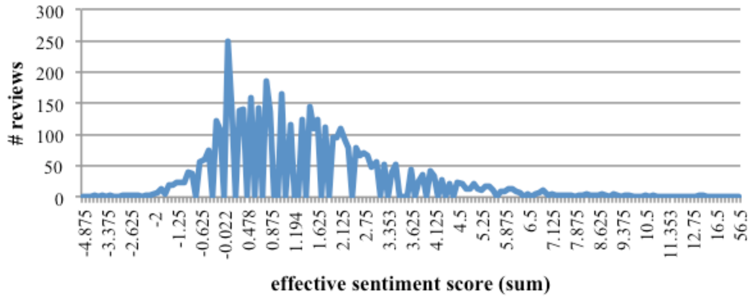

Fig. 1. Distribution of reviews according to sentiment scores

4 Factors of User Satisfaction

We use statistical models to define the dependencies between the ratings and the sentiment scores as well as other textual, tour and user factors. The goal is to identify those factors relating most to user satisfaction. Since the distribution of the ratings is skewed, we divide reviews into two subgroups: those with top rating in all categories; those where at least one rating is less than five. There are 2145 reviews (54.8%) in the first and 1767 in the second. We develop two binary logistic regression models.

In the first model, predictor variables are the grouped length of a tour; the group size (cf. Section 2); the number of countries the tour passed and the continent. We keep one tour operator (*Tour Operator B*) as a control variable. The second model uses all variables. Here, we include whether a user has posted the review *soon* or *late* (cf. Section 2), the sum and the variance of the sentiment scores (cf. Section 3). The number of words in a review is used as a control variable. We also include a binary variable “Comment in Meets by User” for indicating whether or not the composer of the review has participated within a meet. The models are tested on a subsample of 3910 reviews, where we randomly select 850 reviews belonging to group 1 (“5 stars only”) and 850 belonging to the other group. Table 1 shows the results.

Table 1. Logistic Regression Models

Logistic Regression	**Model 1**	**Model 2**
(Intercept)	-1.11 (0.29)***	-0.94 (0.30)**
Tour Length Medium	0.31 (0.15)*	0.36 (0.16)*
Tour Length Short	0.78 (0.24)**	0.75 (0.25)**
Tour Length Very Long	0.26 (0.36) .	0.18 (0.16)
Group Size Large	0.52 (0.18)**	0.69 (0.19)
Group Size Medium	0.31 (0.17) .	0.44 (0.18)*
Group Size Small	-0.15 (0.17)	-0.04 (0.17)***
Number of Countries	-0.05 (0.03)*	-0.05 (0.03).
Continent Asia	0.94 (0.27)***	1.13 (0.28)***
Continent Australia	-0.03 (0.30)**	0.23 (0.31)
Continent Europe	0.79 (0.28)**	0.74 (0.28)**
Continent North America	1.23 (0.32)***	1.24 (0.33)***
Continent South America	0.77 (0.36)*	0.83 (0.37)*
Tour Operator B	3.34 (0.76)***	3.02 (0.76)***
Time of Posting by User Soon		-0.74 (0.11)***
Sentiment Scores Sum		0.02 (0.03)
Sentiment Scores Variance		3.11 (1.43)*
Review Length		-0.001 (0.00)
Comment in Meets by User		-0.30 (0.12)**
R^2	0.081	0.129
Number of Complete Cases	1700	1687
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $.p < 0.1$		

In Model 1, the most significant predictors for the top rated reviews are whether the tour is offered by *Tour Operator B* and the continent of the tour. The baseline of the factor continent is *Africa*. We see that tours in *Asia*, *Europe*, *North America* and *South America* are significantly more likely to be top rated than tours in *Africa*. As opposed to this, a tour in *Australia* is slightly more likely to be rated worse. The number of countries of a tour is also significantly related to the outcome – the more countries the worse the rating. Another significant predictor is the tour length. Compared to *long* tours (i.e., the baseline category) a tour of any other length (i.e., *medium*, *short* or *very long*) is likely to get a better rating. If the group size is *large*, the tour is significantly more likely to get top rated compared to the baseline (i.e., the group size is not indicated). Surprisingly, tours with *smaller* group size are more likely to be rated badly. This points at psychological factors independent of tour characteristics.

The results for Model 2 are displayed in Table 1 on the right side. We see that the tour related variables of the model hardly change. Among the new variables, the time of the posting is strongly significant. If a review is posted *soon*, than the user is more likely to be not satisfied (when users are not satisfied, they tell it immediately). If a user was engaged in a meet, she/he may be more critical, as the likelihood for a top rating decreases. As for the sentiment scores within a review, only the variance is significant. Thus, “5 stars only” reviews contain a higher variety of emotion.

To find out which of the predictor variables of Model 2 has the highest impact, an analysis of variance (ANOVA) is done. It shows that the time of the posting is the best predictor, then continent, group size, Tour Operator B, the length of the tour, user participation in a meet, the variance of the sentiment scores, the number of countries, the review length and the average sentiment score. Thus, the impacts of the tour and user related variables are higher than those of text. Whether or not a user was engaged in a meet can be considered as a good predictor. However, the user information is incomplete. We only know for 1268 reviews (31.6%) that the user was engaged in a meet. Thus, if the data quality improves, the results might change.

5 Conclusions

While analyzing reviews of group tours from an online platform, we discovered that the reviews are predominantly positive, making it difficult to differentiate them. To tackle this problem, we applied sentiment analysis. However, this did not help us to understand the data. We found out that the resulting sentiment scores of the reviews were mainly neutral. Especially for short reviews the sentiment scores lead to wrong results. We learnt that meta data associated with the reviews, e.g. tour and user related factors, predicted better the overall user satisfaction.

The moderate association between sentiment scores and ratings gives room for further research. Thus, in a next step we will look at the text in more detail. We will apply statistical models based on the unigram-, bigram- and trigram-frequency distributions including pointwise mutual information and likelihood ratio (Manning, 1999).

References

Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC Vol. 10, pp. 2200-2204.

Garcia, A., Gaines, S., & Linaza, M. T. (2012). A Lexicon Based Sentiment Analysis Retrieval System for Tourism Domain. e-Review of Tourism Research (eRTR), 10, 35-38.

Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. In: Information and Communication Technologies in Tourism, Springer, Vienna, New York: 460-470.

Manning, C. D. (1999). Foundations of statistical natural language processing. H. Schütze (Ed.). MIT press.

Miller, M., Sathi, C., Wiesenthal, D., Leskovec, J., & Potts, C. (2011). Sentiment Flow Through Hyperlink Networks. In ICWSM.

Schmunk, S., Höpken, W., Fuchs, M., & Lexhagen, M. (2013). Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC. In Information and Communication Technologies in Tourism 2014 (pp. 253-265).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Werthner, H., Cantoni, L., Dickinger, A., Gretzel, U., Jannach, D., Pröll, B., Ricci, F., Scaliogne, M., Stangl, B., Stock, O., & Zanker, M. (2014). Future research issues in IT and tourism. A Manifesto as a result of the JITT workshop in June 2014, Vienna. To be published in Information Technology & Tourism.