Sentiment Analysis of Reviews:
Should we analyze writer intentions or reader perceptions?

Isa Maks and Piek Vossen
Vu University, Faculty of Arts
De Boelelaan 1105, 1081 HV Amsterdam
e.maks@vu.nl, p.vossen@vu.nl

Abstract

Many sentiment-analysis methods for the classification of reviews use training and test-data based on star ratings provided by reviewers. However, when reading reviews it appears that the reviewers' ratings do not always give an accurate measure of the sentiment of the review. We performed an annotation study which showed that reader perceptions can also be expressed in ratings in a reliable way and that they are closer to the text than the reviewer ratings. Moreover, we applied two common sentiment-analysis techniques and evaluated them on both reader and reviewer ratings. We come to the conclusion that it would be better to train models on reader ratings, rather than on reviewer ratings (as is usually done).

1 Introduction

There is a growing volume of product reviews on the web which help customers to make decisions when planning to travel or buying a product. Sentiment-analysis tools try to discover user opinions in these reviews by converting the text to numerical ratings. Building these tools requires a large set of annotated data to train the classifiers. Most developers compile a training and test corpus by collecting reviews from web sites on which customers post their reviews and give a star rating. They test and train their tools against these reviewer ratings assuming that they are an accurate measure of the sentiment of the review.

However, when reading reviews and comparing them with the reviewer ratings there does not always seem to be a clear and consistent relation between these ratings and the text (cf. also Carrillo de Albornoz et al., 2011). That is, from a reader's perspective, there is a discrepancy between what the reviewer expresses with the numerical rating and what is expressed in text. For example, the following hotel review was rated '7' (weakly positive), whereas possible guests probably would not go to the hotel after having read the review.

The hotel seems rather outdated. The breakfast room is just not big enough to cope with the Sunday-morning crowds.

This mismatch between the reviewer's rating and the review's sentiment may lead to problems. For example, reviews are often ranked according to their reviewer's ratings from highly positive to highly negative. If the review text is not in accordance with its ranking, the rankings may become ineffective. In the area of sentiment analysis and opinion mining the mismatch may lead to methodological problems. Testing and training of sentiment-analysis tools on reviewer ratings may lead to the wrong results if the mismatch between the ratings and the text proves to be a common phenomenon.

We assume that one of the most important sources of this mismatch is the fact that the reviewer writes the review and, separately, rates the experience (i.e., with the book he read, with the hotel he stayed at, with a product he bought). Of course, both text and rating are based on the same experience but they do not

Proceedings of Recent Advances in Natural Language Processing, pages 415–419,
Hissar, Bulgaria, 7-13 September 2013.

necessarily express the same aspects of it. If we have a closer look at the hotel review above, the reviewer probably rates the hotel with a '7', because there may be some positive aspects which he does not mention in his review.
We hypothesize that reader ratings which express the reader's perceptions of the sentiment of a text are a good alternative. As the reader's judgment is based solely on the text of the review, we assume that its rating is closer to the sentiment of the text than the reviewer's rating.

In this study we investigate whether the observed mismatch between reviewer rating and the sentiment of the review is a common phenomenon and whether reader ratings could be a more reliable measure of this sentiment than reviewer ratings.

The next section presents related work. In section 3, the reliability of reviewer and reader ratings as a measure of a review's sentiment is further investigated by performing an annotation study. In section 4, we study the effect of the different types of ratings on the performance of two widely used sentiment-analysis techniques. Finally, we conclude with a discussion of our findings.

2 Related Work

There is a large body of work concerning sentiment analysis of customer reviews (Liu, 2012). Most of these studies regard sentiment analysis as a classification problem and apply supervised learning methods where the positive and negative classes are determined by reviewer ratings. Studies propose additional annotations only when focusing on novel information which is not reflected in the user ratings (Toprak et al., 2010, Ando and Ishizaki, 2012). The issue of a possible mismatch between reviewer ratings and review text is usually not addressed.

Much attention is paid to the customer's (or reader's) perspective in studies in the area of business and social science. Mahony et al. (2010) and Ghose et al. (2012) study product reviews in relation to customer behavior. Their aim is to identify reviews which are considered helpful to customers and to know what kind of reviews affect sales. Their work is similar to ours because of the focus on the effect of the review text on the customer/reader, but they also include other types of information such as transaction data, consumer browser behavior and customer preferences. However, none of these studies focus on the relationship between reviewer rating and review text.

As far as we know, Carrillo de Albornoz et al. (2011) is the only study which mentions the mismatch between rating and text. They ignore reviewer ratings and employ a new set of ratings for the training and testing of their system. From their work, however, it is neither clear to what extent the new ratings differ from the user ratings as they do not report inter-annotator agreement scores nor what the effect is of the different ratings on classifier performance.

3 Reviewer and reader annotations

To get a better understanding of the relationship between reviewer ratings, review text and reader ratings, we perform an annotation study which allows us to answer the following research questions: (1) To what extent are mismatches between reviewers' ratings and sentiments common? And (2) Can reader ratings be employed to measure review sentiment more reliably?

3.1 Hotel review corpus

For the annotation study we compiled a corpus of Dutch hotel reviews. The corpus consists of 1,171 reviews extracted from four different booking sites during the period 2010-2012. The reviews have been collected in such a way that they are evenly distributed among the following categories:

- They are collected from different booking site like Tripadvisor.com, zoover.com, hotelnl.com and booking.com
- They include most frequent text formats: pro-con (where boxes are provided for positive and negative remarks) and free text.

- They include reviews on hotels from all over the world (although the majority is Dutch).
- They include reviewer's ratings ranging from strong negative to strong positive

Each review contains the following information:

- Reviewer rating: a user rating given by the reviewer translated to a scale ranging from 0 to 10 (very negative to very positive) describing the overall opinion of the hotel customer.
- Review text: a brief text describing the reviewer's opinion of the hotel.
- Reader ratings: ratings of two readers on a scale ranging from 0 to 10. These ratings are described in more detail in the next section.

3.2 Reader ratings and agreement scores

Two annotators (R1 and R2), both native speakers of Dutch and with no linguistic background, added a reader rating to each review. They were asked to read the review and rate the text on a scale from 1-10 (very negative to very positive), answering the question whether the reviewer would advise them to choose the hotel, or not. They were asked to ignore their own preferences as much as possible.

We measured the Pearson Correlation Coefficient (*r*) between the 10-point numerical rating scales of each annotator pair (R1, R2 and reviewer), regarding the reviewer (REV) also as an annotator. As correlation can be high without necessarily high agreement on absolute values, we also performed evaluations on categorical values. A 2-class evaluation was performed by translating 1 to 5 ratings to 'positive' and 6 to10 ratings to 'negative'; a 4-class evaluation is performed by translating 1-3 ratings to 'strong negative', 4 to 5 ratings to 'weak negative', 6 to 7 ratings to 'weak positive' and 8 to 10 to 'strong positive'. Agreement was measured between each annotator pair in terms of percentage of agreement (%) and kappa agreement (κ).

raters	1/10	2-class		4-class	
REV-R1	0.82 *r*	0.81 κ	0.90%	0.51 κ	0.63%
REV-R2	0.83 *r*	0.82 κ	0.91%	0.53 κ	0.65%
R1-R2	0.92 *r*	0.92 κ	0.96%	0.71 κ	0.78%

Table 1. Inter-annotator agreement.

Table (1) shows that inter-annotator agreement is quite high between all raters, both when correlation is measured on the 10-point-scale ($r >= 0.82$) and when agreement is measured with the 2-class annotation sets ($\kappa >= 0.81$). Agreement on the 4 class annotations is much lower ($\kappa >= 0.51$) showing that polarity strength is difficult to annotate. However, given the purpose of this study, we are not interested in agreement as such. Our focus is on the differences in agreement between readers and reviewers. From that perspective it is interesting to note that, according to all measures, the reviewer is an outlier. Agreement between each individual reader and the reviewer (REV-R1 and REV-R2, respectively) is consistently lower than agreement between both readers (R1-R2). The differences already become important when measuring agreement on 2-class annotations, but even more prominent when measuring agreement on 4-class annotations. All observed differences ranging from 5 up to 15%, are statistically significant ($p < 0.01$).

On the basis of these results, we can answer our research questions (cf. section 3). We infer that the observed mismatch between the sentiment of the review and reviewer rating is a relatively common phenomenon. With respect to at least 10% (cf. table 1, row 2, column 4) of the reviews (when reviews are categorized in 2 categories) up to approx. 37% (cf. table 1, row 1, column 6) of the reviews (when reviews are categorized in more fine-grained categories) readers do not agree with the reviewer. Secondly, the fact that readers have higher agreement with each other than with the reviewer confirms our hypothesis that reader ratings are a more accurate measure of the review's sentiment than reviewer ratings.

4 Implications for sentiment analysis

We investigated how automated sentiment analysis methods perform with the different sets of annotations by applying two widely used approaches to document-level sentiment classification. Classifier accuracy is measured against the three sets of ratings (R1, R2 and REV) we described in the previous section.

4.1 The lexicon-based approach

The first method is a lexicon-based approach which starts from a text which is lemmatized with the Dutch Alpino-parser[1].The approach is similar to the "vote-flip-algorithm" proposed by Choi and Cardie (2008). The intuition about this algorithm is simple: for each review the number of matched positive and negative words from the sentiment lexicon are counted. If polar words are preceded by a negator, their polarity is flipped; if polar words are preceded by an intensifier, their polarity is doubled. We then assign the majority polarity to the review. In the case of a tie (being zero or higher than zero), we assign neutral polarity. The sentiment lexicon used in this approach is an automatically derived general language sentiment lexicon obtained by WordNet propagation (Maks and Vossen, 2011).

4.2 The machine-learning approach

The second method is a machine learning approach that also starts from a text that is lemmatized by the Dutch Alpino-parser. After lemmatization the text is transformed to a word-vector representation by applying Weka's StringToWord Vector with frequency representation (instead of binary). We used Weka's NaiveBayesMultinominal (NBM) classifier to classify the reviews. The NBM was chosen because our review texts are rather short (with an average of 68 words) and, according to Wang and Manning (2012), NBM classifiers perform well on short snippets of text. Results reported are average of ten-fold-cross-validation-accuracies using R1, R2 and REV ratings as training and test data.

4.3 Results on different types of ratings

Results are evaluated against the whole set of 1,172 reviews (cf. table 2 'all'). As many approaches to sentiment analysis do not use the class of weak sentiment (Liu, 2012), we also evaluated against a subset of strong negative (ratings 1 to 3) and strong positive (ratings 8 to 10) reviews (cf. table 2, 'strong'). Table (2) shows the classification results in terms of accuracy, obtained by the lexicon-based approach (LBA, row 1, 2, 3) and the machine-learning approach (NBM, row 4, 5, 6).

	name	ratings	all	strong
1	LBA	REV	78.3	85.0
2	LBA	R1	80.5	88.1
3	LBA	R2	80.7	88.1
4	NBM	REV	83.6	86.4
5	NBM	R1	86.9	92.2
6	NBM	R2	86.7	92.2

Table 2. Results of sentiment analysis.

The results show that both approaches perform well against all ratings. Classification of the strong sentiment reviews seems considerably easier than classification of the whole review set. Interestingly, both sentiment analysis approaches appear to perform better on reader ratings than on reviewer ratings. The better performance holds across both selections of reviews and with both approaches. Differences are statistically significant (chi-square test, $p<0.05$) in all cases but the LBA approach on the whole dataset which is almost statistically significant.

5 Discussion and Conclusions

We performed an annotation study that showed that the observed mismatch between reviewer ratings and review's sentiment is a rather frequent phenomenon. Considerable part of the reviews (ranging from 9 to 37% depending on the granularity of the classification) is classi-

[1] http://www.let.rug.nl/ vannoord/alp/Alpino/

fied by the reviewer in the wrong sentiment class.

The annotation study also showed that reader ratings are a more accurate measure. We already expected reader ratings to be closer to the text because they are exclusively based on it. In addition, the annotation study shows that readers agree in their ratings and that the review's sentiment can be reliably annotated by readers.

Our experiments in section 4 show that sentiment-analysis tools perform better with reader ratings than with reviewer ratings. This should probably not surprise us as sentiment analysis behaves like a reader whose only source of information is the review text. As such, this is a promising result. However, since reviewer ratings are widely available and come for free with the text, they will often be used to evaluate the tools. Likewise, training and fine-tuning will be done with reviewer ratings rather than with reader ratings.

We think that researchers and system developers should be aware of the differences between reviewer and reader ratings and their effects on the system they develop. Recently, many sentiment analysis tools perform a more in-depth analysis identifying aspects of products (and services) and their sentiments (Liu, 2012). Again, reviewer ratings are used to train and test these systems. In view of our findings, it seems advisable that researchers and system developers make the effort to collect a set of reader ratings and train and test their tools with them. The additional value of sentiment analysis should be sought in finding the sentiment of the text rather than in finding the sentiment of its writer.

Acknowledgements

This research is supported by the European Unions 7th Framework Programme via the OpeNER (Open polarity enhanced Named Entity Recognition) Project (ICT-296451).

References

Ando, M. and S. Ishizaki (2012) Analysis of travel review data from Reader's point of View. In *Proceedings of WASSA-2012*. Jeju, South-Korea.

Carrillo de Albornoz, J., L. Plaza, P. Gervás and A. Diaz (2011). A joint model for feature mining and sentiment analysis for product review rating. In *Proceedings of ECIR-2011*. Dublin, Ireland.

Choi, Y. and C. Cardie (2008). Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of EMNLP '08*. Hawaii, USA.

Ghose, A., G. Ipeirotis and B. Li (2012). Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*, Vol. 31.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, USA.

Mahony, M., P. Cunningham and B. Smyth (2010). An assessment of machine learning techniques for review recommendation. In *Proceedings of AICS*.

Maks, I., and P. Vossen (2011) Different Approaches to Automatic Polarity Annotation at Synset Level. In: *Proceedings of the First International Workshop on Lexical Resources*, WoLeR 2011, Ljubljana.

Toprak, C., N. Jakob and I. Gurevych. (2010) Sentence and Expression Level Annotation of Opinions in User-Generated Discourse. In *ACL 2010*. Uppsala, Sweden.

Wang, S. and C. Manning. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In *Proceedings of ACL-2012*.