Rating Decomposition with Conjoint Analysis and Machine Learning

Florian Volk, Nadine Trüschler, Max Mühlhäuser
Telecooperation Lab
Technische Universität Darmstadt/CASED
Darmstadt, Germany
email: florian.volk@cased.de, max@informatik.tu-darmstadt.de

Abstract—**When customers leave feedback about products, for example, a rating, they often evaluate a product as monolithic unit, neglecting that products are composed of parts with different quality, often delivered by independent suppliers. Manufacturers are more interested in individual ratings for product parts than in an overall rating. With decomposed ratings, manufacturers can improve the product quality, the selection of suppliers, and adapt pricing strategies. In this paper, we present an automated approach to decompose overall product ratings into individual ratings for product parts by the use of the results of a Conjoint analysis and supervised machine learning. Using this approach, individual ratings for product parts can be predicted with a high accuracy.**

Keywords-product ratings; decomposition; conjoint analysis; machine learning; classification; product quality; supplier selection.

I. INTRODUCTION

A recent study has shown that many customers in e-commerce scenarios are strongly influenced by feedback from other customers when making purchase decisions [1].

Customer feedback is a dually user-centric aspect of e-commerce: feedback is created by users and used by users. Product manufacturers are also interested in customer feedback as a source of quality information regarding their products. Based on customer feedback, manufacturers can identify and correct flaws in their products, as well as adapt their products to suit the expectations and requirements of customers better. Moreover, manufacturers can adapt their pricing strategy with regards to issues that are identified using feedback but cannot be corrected.

A. Composite Products

Products are usually created by a manufacturer and sold via some intermediaries to customers. However, manufacturers rarely create their products from scratch but use and assemble parts created and delivered by external third-party manufacturers, called suppliers. For example, the automobile industry heavily relies on suppliers that are independent from the car manufacturers.

The quality of a product depends on the quality of its parts, and thus, on the quality delivered by the suppliers. Manufacturers can improve the quality of their products by selecting the best known suppliers. For this, information on the individual quality of suppliers is needed.

B. Customer View on Products

Customers usually see products as monolithic units, and thus, feedback is targeting the product as a whole, but not a composite of independent parts. Manufacturers, however, are interested in obtaining individual feedback on product parts in order to learn about supplier quality.

For some types of products, customers can clearly differentiate product parts, for example, a tool and the manual found alongside the tool. In this example, low printing quality of the manual can be mentioned in the feedback. A manufacturer receiving such feedback can easily attribute the issue to the supplier of the manual.

In order to evaluate the capabilities of customers to give ratings for product parts, a study with 229 participants was conducted (29.3% female, 70.7% male, 48% university students, 44.9% employees). The participants of the study were confronted with products of increasing complexity (e.g., a nightstand, a bicycle, a hi-fi system) and were asked about their capability to identify the reasons for faults. The self-declared identification capability of the participants decreased with increasing complexity of the products. In a later task of the same study, 35.4% of the participants reported being unable to decompose their feedback into individual ratings. Another 14.8% are unsure. Additionally, the state of warranty influences customer feedback as such that faulty products are returned under warranty without the customers caring about the reason for the fault.

Generally spoken, depending on the type of product and the type of fault, it is hard to obtain decomposed ratings from customer feedback. However, manufacturers are more interested in decomposed ratings rather than in feedback on the product as a whole.

C. Contribution

In this paper, we present an automated approach to decompose overall ratings into individual ratings for suppliers. We combine results from Conjoint analyses with supervised machine learning techniques.

D. Structure

This paper is structured as follows. In Section II, we review scientific work related to our approach. Our method of combining Conjoint analyses with supervised machine learning is presented in Section III. We obtained a data set of 2,544 ratings; this data set is discussed in Section IV.

ISBN: 978-1-61208-440-4

Section V presents and discusses the evaluation of our approach. We conclude in Section VI.

II. Related Work

In [1], Volk et al. present a decomposition-supporting review system aimed at human customers. This form-based review system is intended to assist humans in giving structured and problem-oriented feedback.

The contribution of this paper at hand does not involve user interaction during the decomposition process. The method is fully automated and operates on customer-supplied overall ratings on an ordinal scale.

A. *Natural Language Processing*

Most related approaches to automated feedback decomposition apply natural language processing (NLP) technologies to derive ratings or problem statements from written feedback, i.e., from reviews.

Lin and Hovy locate important sentences within paragraphs in order to create summarized reviews [2]. Instead of a summary, our proposed approach derives individual ratings.

Turney searches for keywords, as, e.g., “good” or “excellent” and then applies unsupervised learning to derive ratings on a five star scale [3] with 74% accuracy. Our approach does not need textual reviews to derive such ratings with higher accuracy.

In order to identify the reviewers’ reasons for leaving feedback, Kim and Hovy assume the existence of a main statement that is either positive or negative in every review [4]. We do not share this assumption and affirm the existence of multiple main statements.

1) *Decomposition by Topic*

The following approaches have in common that they attempt to identify common topics within reviews. Our proposed approach is targeted at product parts, not topics, which may involve multiple product parts at once.

Dave, Lawrence, and Pennock derive (binary) sentiments towards topics [5]. Gamon et al. present a very similar approach in [6].

A summarization approach for multiple reviews is presented by Zhan, Loh, and Liu [7].

2) *Decomposition by Product Features*

Close to our approach is the decomposition of written reviews by product features as presented in the following publications. Product features can often be correlated with certain product parts, and, thus, associated to specific suppliers.

The feature-based summarization (FBS) system by Hu and Liu assigns binary ratings to product features derived from multiple reviews [8]. Liu, Hu, and Cheng extend the FBS system in [9]. Their *Opinion Observer* improves precision and recall over FBS.

Aciar et al. apply an ontology, which is specific to a product, to identify and rate the features of this product as described by the ontology [10]. They also calculate an overall rating. Similar, but without the need for a product-specific ontology, is the approach by Archak, Ghose, and Ipeirotis [11]. Their approach replaces the need to create ontologies first with learning product features from reviews.

By using lexicon-enhanced sentiment classification, Dang, Zhang, and Chen identify sentiments towards a product [12]. These sentiments are rated afterwards.

The closest to our approach is the *Opinion Digger* presented by Moghaddam and Ester [13]. They combine unsupervised property extraction and text mining in order to derive star ratings for product properties.

III. Method

Our approach combines the results of a traditional Conjoint analysis with supervised machine learning, more specific: with random forest classifiers [14]. For this purpose, the Conjoint analysis returns information about the importance of product aspects, or product parts. This information is fed as additional features to the machine learning process together with an overall product rating (see Figure 1).

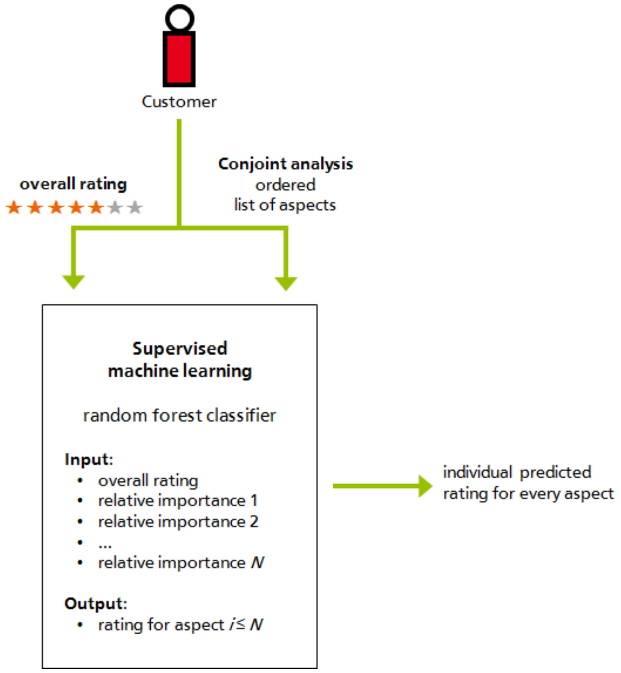


Figure 1. Combining Conjoint analyses and supervised machine learning to automatically decompose ratings.

A. *Conjoint Analysis*

Conjoint analyses are a set of techniques commonly used in market research. As an instance of regression analysis, Conjoint analyses identify the order of importance of product parts [15][16]. Thereby, Conjoint analyses return an importance weight for every aspect of the tested product. Conjoint analyses are often used during product design to identify the most important aspects of the product from a customer perspective. This information is used to optimally customize a product for the target market.

There are three main types of Conjoint analyses:

- **Traditional Conjoint analysis**. Participants of a study are asked to rate multiple instances of the same product with minimal differences for multiple aspects. Usually, only a subset of the possible combinations of stimuli (characteristics of a product aspect) is shown to the participants.
- **Adaptive Conjoint analysis**. A test system generates new combinations of stimuli based on the previous ratings by participants. This is done to focus on the product aspects that are most important for specific participants.
- **Choice-based Conjoint analysis**. The participants are presented with two products in each iteration of the analysis. Instead of rating these products, the participants only state which product is the better one according to their beliefs.

Instead of asking study participants directly for their impression of the importance for all product aspects, Conjoint analyses derive these from a multitude of products to reveal hidden relations between aspects, which the participants are not aware of.

B. Supervised Machine Learning with Random Forests

According to Witten et al. [14] and van Leeuwen [17], machine learning is the automated construction of algorithms that “learn” from data.

In supervised machine learning, a model is trained on a set of sample input data and the desired output data. The calculated model (in our case: the trained classifier) is a generalization of the input and generates its output values accordingly.

Trained models are evaluated with a second data set that only includes input data, but lacks the desired output. Comparing the generated output with the desired output (which is known to the evaluator) allows to measure the performance of the trained model.

We have compared the performance of several classifiers, namely naïve Bayes, SVM, Cart tree, and random forest classifiers. The random forest classifiers [14] (see chapter 8.3) perform best on our data set.

IV. DATA SET

Our approach was evaluated on a data set obtained from 212 human raters. Every rater participated in a traditional Conjoint analysis for two product types: digital cameras and smartphones.

The two data sets each contain 2.544 rating samples (212 raters and 12 products).

A. Participants

In a first step, the raters were confronted with different products in the form of a traditional Conjoint analysis. In this step, importance weights for the product features (assignable to product parts) were obtained. As a traditional Conjoint analysis was performed instead of a choice-based Conjoint analysis, this first step also returned overall product ratings for every tested combination of features and stimuli.

Afterwards, the raters were asked to rate all individual parts of the products. This second step generated the ratings for individual features. These ratings are used to train and evaluate the random forest classifiers.

B. Rating Scale

All products and aspects are rated on a seven-star scale ranging from one star (worst rating) to seven stars (best rating).

The seven-star scale was used instead of the five-star scale known from common internet shopping portals to enable a higher level of detail in the ratings. For the study design, we assumed that a scale of nine or more stars overstrains the differentiation capabilities of most human raters, but. We, however, wanted to achieve a higher level of granularity than a five-star scale. Furthermore, an odd number of available ratings guarantees the existence of a neutral rating.

C. Products and Stimuli

The stimuli were taken from real-world products and are given in Table I. Table I already orders the product features by their importance weights as retrieved from the Conjoint analysis.

TABLE I. PRODUCT FEATURES AND STIMULI IN THE DATA SETS

Data set	Product Feature	Stimuli
Digital cameras	Display size [inch] Zoom Resolution [megapixels] Price [€]	2.7, 2.8, 3.0 3.6x, 7.1x, 40x 12, 14.2, 20 310, 419, 566
Smartphones	Display size [inch] App store Talking time [hours] Price [€]	4.0, 5.0, 5.7 Google, Apple, Microsoft 7, 8, 9 460, 535, 777

The analysis was conducted in English and in German language. English-language participants were shown the respective real world prices in US-$ to account for influences from different tax schemes and the seller’s market strategies.

V. EVALUATION

Multiple performance metrics are used to evaluate the classification performance of the trained random forest classifiers.

A. Leave-One-Out Cross Validation

A full Leave-One-Out cross validation (LpO cross validation with p=1) was conducted.

In total, 2,544 classifier models are trained on all rating samples but one. The remaining one rating sample is used to test the classifier. This way, every sample is once used to test and used 2,543 times to train the classifier. All possible combinations of samples in the data sets are used. Thus, performing an additional 10-fold cross validation was omitted.

The Leave-One-Out cross validation yields a prediction accuracy of 88%. For comparison, random guessing on a seven-star scale achieves only 14% accuracy.

B. *Monte-Carlo Simulation*

Full Leave-One-Out cross validation generates the most reliable performance measurements. However, training classifiers for every product feature with all available ratings (but one) is a computationally expensive task.

In order to evaluate the performance of our approach in a more realistic model, we performed a 1,000 round Monte-Carlo simulation. In every round, the classifiers are trained with just 100 randomly chosen samples from our data sets and tested on the remaining 2,444 samples. A lower number of samples reduces the computational effort for training the classifier model. The results are shown in Figure 2.

1) *Compared Approaches*

a) The approach proposed in this paper.

b) A baseline approach, which uses the same supervised machine learning technique, but the training set omits the importance weights derived from the Conjoint analysis.

c) Randomly guessing a rating on a seven-star scale.

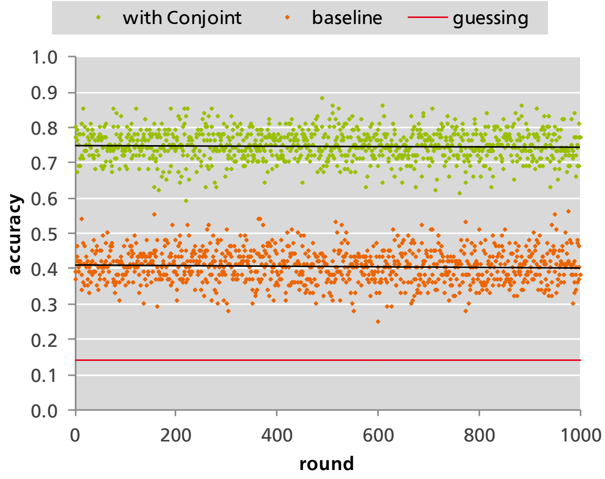


Figure 2. Comparison of the contributed approach with a baseline approach and guessing.

As can be seen, our proposed approach outperforms the other two considerably. With only 100 samples in the training set, an average accuracy of 75% is achieved. The baseline approach on average achieves slightly more than 40% accuracy. Guessing only selects the correct rating in 14% of the cases.

2) *Detailed Performance Metrics*

Figure 3 shows the measures performance of the trained random forest classifiers in detail. The performance is measured with standard measures for categorical data, e.g., [18].

As can be seen, the best results are achieved for the price of the smartphone and the smartphone display size. Worst performance is measured for the smartphone app store and the smartphone talk time.

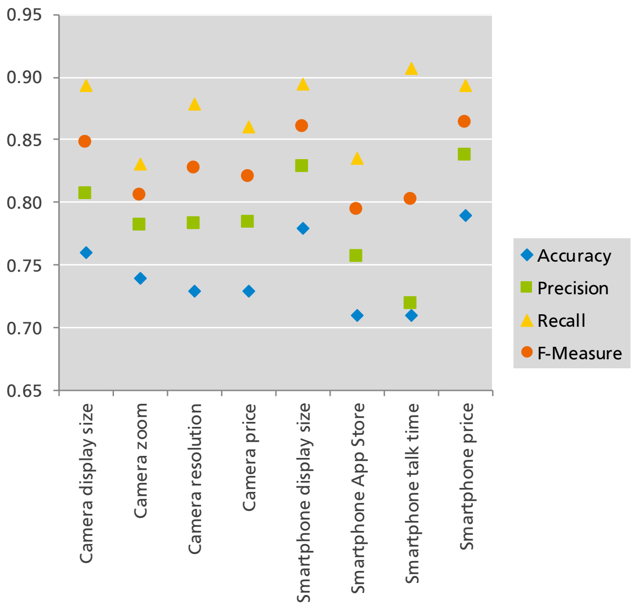


Figure 3. Detailed performance metrics of the trained random forest classifiers (average values).

C. *Accuracy in Ordinal Classification*

Even though the performance of our approach is considerably better than the compared approaches, it is even better than what the chosen performance metrics indicate.

Accuracy, precision, recall, and f-measure are calculated based on the amount of true positives, true negatives, false positives, and false negatives. In binary classification tasks, predicted ratings are classified according to the classes shown in Table II.

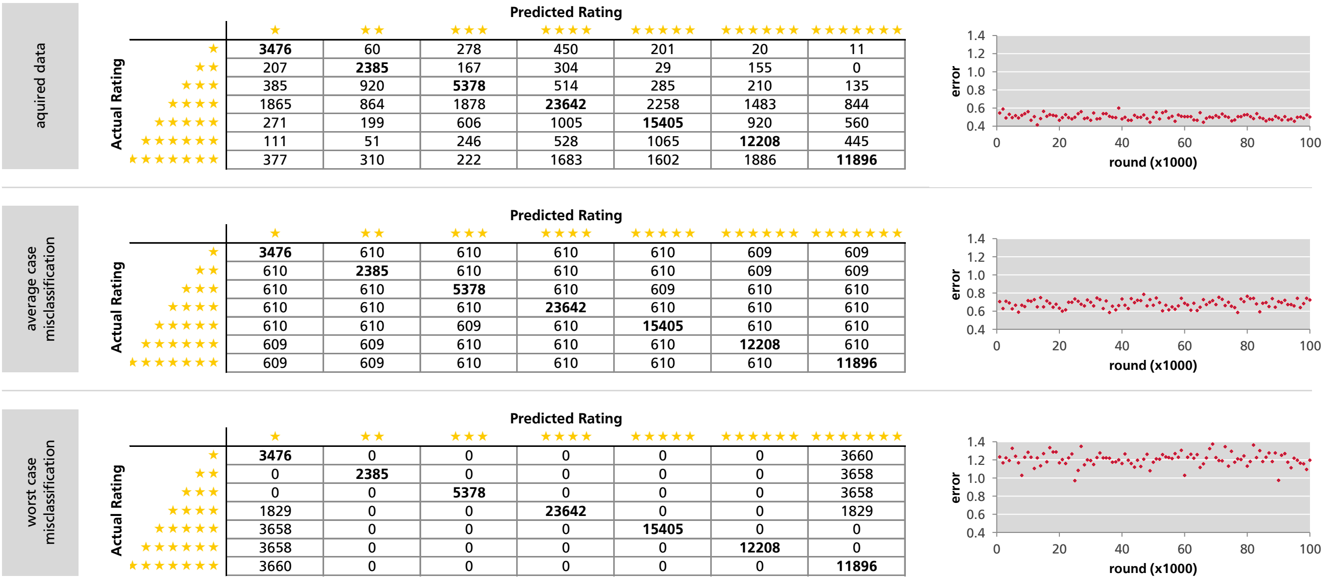
TABLE II. CONFUSION MATRIX FOR BINARY CLASSIFICATION TASKS

		prediction	
		true	*false*
observation	*true*	true positive	false negative
	false	false positive	true negative

However, predicting ratings is an ordinal classification task. As such, there are multiple options for a misclassification with different severity.

For example, predicting a rating of 4 stars rating instead of a (correct) 5 stars rating is less severe than predicting a 1 star rating. Nevertheless, both misclassifications affect the standard measures the same way.

While there are some metrics, to the best of our knowledge, no standard evaluation method for ordinal classification has been established, yet [19][20].



aquired data

Predicted Rating

Actual Rating	★	★★	★★★	★★★★	★★★★★	★★★★★★	★★★★★★★
★	**3476**	60	278	450	201	20	11
★★	207	**2385**	167	304	29	155	0
★★★	385	920	**5378**	514	285	210	135
★★★★	1865	864	1878	**23642**	2258	1483	844
★★★★★	271	199	606	1005	**15405**	920	560
★★★★★★	111	51	246	528	1065	**12208**	445
★★★★★★★	377	310	222	1683	1602	1886	**11896**

average case misclassification

Predicted Rating

Actual Rating	★	★★	★★★	★★★★	★★★★★	★★★★★★	★★★★★★★
★	**3476**	610	610	610	610	609	609
★★	610	**2385**	610	610	610	609	609
★★★	610	610	**5378**	610	609	610	610
★★★★	610	610	610	**23642**	610	610	610
★★★★★	610	610	609	610	**15405**	610	610
★★★★★★	609	609	610	610	610	**12208**	610
★★★★★★★	609	609	610	610	610	610	**11896**

worst case misclassification

Predicted Rating

Actual Rating	★	★★	★★★	★★★★	★★★★★	★★★★★★	★★★★★★★
★	**3476**	0	0	0	0	0	3660
★★	0	**2385**	0	0	0	0	3658
★★★	0	0	**5378**	0	0	0	3658
★★★★	1829	0	0	**23642**	0	0	1829
★★★★★	3658	0	0	0	**15405**	0	0
★★★★★★	3658	0	0	0	0	**12208**	0
★★★★★★★	3660	0	0	0	0	0	**11896**

Figure 4. Confusion matrices with the same accuracy, precision, and recall, but with different prediction performance (shown as absolute prediction error).

On the example of accuracy, precision, recall, and one specific random forest classifier from the Monte-Carlo simulation, Figure 4 illustrates how our approach performs better than indicated.

- On top, the ratings predicted by the trained classifier are shown alongside with the absolute prediction error |actual rating – predicted rating|. As can be seen, the typical absolute prediction error is 0.5 stars.
- In the middle, the confusion matrix of the average misclassification case is shown, i.e. correct classifications are as returned by the trained classifier, but the misclassifications are evenly distributed over all classes. The error is about 0.7 stars. However, for this confusion matrix, the same accuracy, precision and recall are calculated as for the first case, which has predicts more correct ratings.
- The worst case misclassification case is shown in the bottom row of Figure 4. Again, all correct classification are as returned by the trained classifier. All misclassifications are maximally wrong, i.e., |actual rating – predicted rating| is maximized. Here, the error is about 1.2 stars, still yielding the same accuracy, precision and recall as all the other two cases. Additionally, the standard deviation of the prediction error (0.074 stars) is more than twice as high as for the first confusion matrix (0.032 stars).

VI. Conclusion

We have presented an automated approach for rating decomposition. This approach enables manufacturers to break down overall product ratings (usually given by customers) into individual ratings for product parts. Under the assumption that the product parts can be related to contributions made by external suppliers, the manufacturer is able to track the performance of its suppliers. With this information available, manufacturers can improve their choice of suppliers, and, thus, improve the quality of the products given to customers.

By supplying the results of a Conjoint analysis as additional features to a supervised machine learner, a classification model is trained, which predicts ratings for product parts based on an overall rating and importance weights for product parts.

Our evaluation shows that the proposed approach outperforms the baseline approach, which omits the results of the Conjoint analysis, as well as naïve guessing. In a leave-one-out cross validation, our approach achieves 88% prediction accuracy, i.e., the correct rating (on a seven star scale) is predicted in 88% of the cases.

A previous study has shown that many human customers are in need of assistance when giving decomposed reviews. Our approach is independent from user interaction, as only the overall rating is retrieved from the customer and automatically decomposed into individual ratings.

A. *Future work*

Our data set is specific to two product types and a seven-star rating scale. In order to evaluate the performance of the proposed approach in a general fashion, more tests are useful. This relates to both a wider set of products and to different rating scales.

When training the classifier, the leave-one-out cross validation requires the model to be trained on all available rating samples (but one), which implies extensive computational effort. The performed Monte-Carlo simulation only used 100 samples (3.9%) and achieved a lower, but still noticeably high accuracy of about 75%. It is subject to future work to find an optimal balance between training set size and prediction performance.

ACKNOWLEDGMENT

Some of the results discussed here will also be available in “Florian Volk, Detailing Reviews and Ratings for Trust-Enhanced Composition, PhD thesis, Technische Universität Darmstadt, 2015” *(to appear).*

This work has been co-funded by the DFG as part of project S1 within the CRC 1119 CROSSING.

REFERENCES

[1] F. Volk, J. Pitzschel, and M. Mühlhäuser “Making the Most of Customer Product Reviews,” in 7th International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC), 2014, pp. 1–6.

[2] C.-Y. Lin and E. Hovy, “Identifying Topics by Position,” in Proceedings of the fifth conference on Applied Natural Language Processing, 1997, pp. 283–290. Association for Computational Linguistics.

[3] P. Turney, “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), ACL ’02, 2002, pp. 417–424. Association for Computational Linguistics.

[4] S.-M. Kim and E. Hovy, “Automatic Identification of Pro and Con Reasons in Online Reviews,” in Proceedings of the COLING-ACL ’06, 2006, pp. 483–490. Association for Computational Linguistics.

[5] K. Dave, S. Lawrence, and D. Pennock, “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews,” in Proceedings of the 12th International Conference on World Wide Web (WWW), 2003, pp. 519–528. ACM.

[6] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, “Pulse: Mining Customer Opinions from Free Text,” in Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis (IDA), IDA’05, 2005, pp. 121–132. Springer-Verlag.

[7] J. Zhan, H. T. Loh, and Y. Liu, “Gather customer concerns from online product reviews — A text summarization approach,” in Expert Systems with Applications 36(2), 2009, pp. 2107– 2115.

[8] M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), KDD ’04, 2004, pp. 168–177. ACM.

[9] B. Liu, M. Hu, and J. Cheng, “Opinion Observer: Analyzing and Comparing Opinions on the Web,” in Proceedings of the 14th International Conference on World Wide Web (WWW), WWW ’05, 2055, pp. 342–351. ACM.

[10] S. Aciar, D. Zhang, S. Simoff, and J. Debenham, “Informed recommender: Basing recommendations on consumer product reviews,” in IEEE Intelligent Systems 22(3), 2007, pp. 39–47.

[11] N. Archak, A. Ghose, and P. G. Ipeirotis, “Show me the Money! Deriving the Pricing Power of Product Features by Mining Consumer Reviews,” in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), KDD ’07, 2007, pp. 56–65. ACM.

[12] Y. Dang, Y. Zhang, and H. Chen, “A Lexicon Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews,” in IEEE Intelligent Systems, 25(4), 2010, pp. 46–53.

[13] S. Moghaddam and M. Ester, “Opinion Digger: An Unsupervised Opinion Miner from Unstructured Product Reviews, “ in Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), CIKM ’10, 2010, pp. 1825–1828. ACM.

[14] I. H. Witten, E. Frank, and M. A. Hall, „Data Mining: Practical Machine Learning Tools and Techniques, Third Edition,” Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

[15] J. Huber, “A History of Choice-Based Conjoint,” in 2001 Sawtooth Software Conference Proceedings, 2001, pp. 213–223, Sequiem, WA, USA. Sawtooth Software.

[16] J. J. Louviere and G. Woodworth, “Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data,“ in Journal of Marketing Research 10, 1983, pp. 350–367.

[17] J. van Leeuwen, “Approaches in Machine Learning,” in Philips Symposium on Intelligent Algorithms (SOIA’02), SOIA. Philips Research Laboratories, 2002.

[18] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” in Biometrics 33(1), 1977, pp. 159–174.

[19] L. Gaudette and N. Japkowicz, “Evaluation Methods for Ordinal Classification,” in Advances in Artificial Intelligence, LNCS vol. 5549, 2009, pp. 207-210.

[20] E. Frank and M. Hall, “A Simple Approach to Ordinal Classification,” in Machine Learning (ECML), LNCS vol. 2167, 2001, pp. 145-156.