Towards Opinion Mining from Reviews for the Prediction of Product Rankings

Wiltrud Kessler, Roman Klinger, and Jonas Kuhn
Institute for Natural Language Processing
University of Stuttgart
70569 Stuttgart, Germany
{wiltrud.kessler,roman.klinger,jonas.kuhn}@ims.uni-stuttgart.de

Abstract

Opinion mining aims at summarizing the content of reviews for a specific brand, product, or manufacturer. However, the actual desire of a user is often one step further: Produce a ranking corresponding to specific needs such that a selection process is supported. In this work, we aim towards closing this gap. We present the task to rank products based on sentiment information and discuss necessary steps towards addressing this task. This includes, on the one hand, the identification of gold rankings as a fundament for an objective function and evaluation and, on the other hand, methods to rank products based on review information. To demonstrate early results on that task, we employ real world examples of rankings as gold standard that are of interest to potential customers as well as product managers, in our case the sales ranking provided by Amazon.com and the quality ranking by Snapsort.com. As baseline methods, we use the average star ratings and review frequencies. Our best text-based approximation of the sales ranking achieves a Spearman's correlation coefficient of $\rho = 0.23$. On the Snapsort data, a ranking based on extracting comparisons leads to $\rho = 0.51$. In addition, we show that aspect-specific rankings can be used to measure the impact of specific aspects on the ranking.

1 Introduction

Opinion mining (often referred to as sentiment analysis) is the task of identifying opinions about specific entities, products or persons in text. Reviews for products, for instance from Amazon.com, are a typical resource for opinions. Often, opinion mining is approached as a text classification task in which snippets (like sentences, paragraphs, or phrases) are categorized into being objective or subjective and in the latter case positive, negative, or neutral (Liu, 2015; Täckström and McDonald, 2011; Sayeed et al., 2012; Pang and Lee, 2004). More differentiated results can be obtained by methods that additionally identify the target of the opinion, specific mentions of product characteristics usually called aspects (Choi et al., 2010; Johansson and Moschitti, 2011; Yang and Cardie, 2012; Hu and Liu, 2004; Li et al., 2010; Popescu and Etzioni, 2005; Jakob and Gurevych, 2010; Klinger and Cimiano, 2013).

It has been proposed to use the extracted information for summarizing specific information about a product (Hu and Liu, 2004). The main advantage of such result is that a star rating is not only associated to the whole product but separated for specific aspects. This is helpful when a user aims at getting an overview of the content of reviews but it might still be leading to an overwhelming amount of information.

In this work, we propose to aim at generating a ranked list of products and hypothesize that such a ranking would be more helpful for the typical task of a user to select a product based on specific needs than the exact and isolated value. We therefore discuss two main prerequisites to be able to reach that goal: Firstly, we discuss the need for gold ranking information, which is the fundament for evaluation. Such ranking can in addition be used for data-driven optimization of methods to automatically generate such rankings based on structured or textual review (and therefore opinion-mining based) information. In this work, we utilize two external gold standards,

Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 51–57,
Lisboa, Portugal, 17 September, 2015.

namely the Amazon.com sales ranking of products of a specific category, and the quality ranking by product aspects available at the website Snapsort.com (a service that collects detailed information about cameras and provides comparisons between them).

Secondly, we discuss different approaches to use (target-oriented) opinion mining methods to produce a ranking of products. We focus on fine-grained methods which associate opinion expressions with different aspects. This enables us to create aspect-specific rankings by using only those expressions that refer to a specific aspect of the product. A ranking from a combination of selected aspects can be used to create specific, personalized rankings. Aspect-specific rankings can also be used to determine the influence of an aspect on the overall ranking.

Previous work in this area is comparatively limited. Ganesan and Zhai (2012) enhance information retrieval models by splitting the query into separate parts for the product's aspects and use a dictionary-based query expansion of opinion words. Tkachenko and Lauw (2014) extract statements comparing products from review texts and generate ranked pairs from these comparisons. They perform two types of evaluations. On the one hand, they compare their system output to the ranking retrieved from a gold standard (annotated by crowdsourcing). On the other hand, they generate a gold standard of product quality for specific predefined characteristics (for instance that smaller is better for cameras). In contrast, our work aims at ranking the products themselves and handles the influence of the aspects as a latent variable without predefining them. Further, we use external sources for evaluation.

We provide the following main contributions:

- We discuss the task of predicting a full ranking of products in addition to isolated prediction of ratings.
- We demonstrate how methods for target-oriented and comparison-based opinion mining can be used to predict product rankings. As real-world examples of such rankings, we use the sales ranking from Amazon.com and the quality ranking from Snapsort.com.
- We show that fine-grained opinion mining methods achieve a substantial performance in predicting these rankings from textual information.
- We present aspect-specific rankings that allow for an understanding of the impact of each aspect on the external ranking.

2 Towards Aspect-based Ranking of Products

Most opinion-mining approaches tackle the task of extracting evaluations of products and aspects (targets of opinion) as the result of the mining process. This leaves the interpretation of the ratings of different aspects to the end user. However, the underlying assumption is that this end user is able to combine the information in a way that it can be utilized for making specific decisions. This utility of the information from opinion mining systems is clearly depending on the use cases and subjective needs. Therefore, important characteristics of a ranking of products are:

- The ranking supports specific needs of an individual or of a downstream task.
- The ranking can be purely subjective or inter-subjective.
- A user can be aware of the factors influencing the preferences leading to a ranking or not.

One instance of a ranking which is directly available from structured meta-data is the *sales ranking* of a category of products from an online shop (in this work, we use the sales ranking of Amazon.com). This ranking addresses for instance the needs of a product manager to maximize the popularity of a product. This ranking is inter-subjective and the user is typically not fully aware of all factors influencing the rank. Such factors are the price of the product, the quality, price-performance ratio, advertisements, etc. Therefore, taking into account information generated by fine-grained opinion-mining methods can shed light on the impact of these aspects on this ranking. If reviews and sales ranking come from the same source, the number of reviews being available for a product can be presumed to correlate (or at least interact) with the number sold. Reviews play an important role for a buying decision, so the interaction will also work in the other direction, when a product has many reviews and most of them are positive, chances go up that people will buy it.

Another instance of an available source of information is an *expert ranking* in which a domain expert compares different products and aspects of them and put them into an order. A common source for such ranking are domain specific magazines or websites with the aim of providing users with a condensed source of information supporting their purchase decision. This ranking is typically purely subjective, however, different factors are taken into account, which might be disclosed or not. In this work, we employ the information made available from Snapsort.com. It is a service that collects detailed information about cameras and provides comparisons between them. Their score incorporates aspects from technical specifications like shutter, viewfinder size, whether image stabilization is available, as well as popularity (how many times the camera has been viewed on the website) or number of lenses available. Such a ranking has been used in recently published previous work by Tkachenko and Lauw (2014) who use a partial expert rating in their gold standard when they specify predefined characteristics for their product (for instance that smaller is better for cameras) and evaluate against these aspect-specific rankings.

Both sales ranking and expert ranking are attempting to combine opinions from or for a set of users. However, a ranking of products might be highly subjective. Therefore, we propose that an actual ranking should be based on crowdsourcing without predefining the aspects taken into account to make a decision. As common in annotation tasks for ranking, requesting a full ranking of a list of products from annotators is a cumbersome challenge. Therefore, we propose that such crowdsourcing task should be set up in a learning-to-rank setting, in which annotators are asked to define a preference to a pair of products. Such annotations can then later be used for compiling an inter-subjective ranking as well as a personalized ranking. This approach is not performed in this paper but constitutes relevant future work. From such rankings, a personalized preference function can be learnt which weights different aspects against each other, even if the user is not aware of these factors.

Related to this proposal is the work by Tkachenko and Lauw (2014) who created a gold standard of textual comparison mentions with crowdsourcing. Ganesan and Zhai (2012) use information from semi-structured reviews in which users provide scores for different aspects.

3 Methods

Our goal is to create a ranked list of products based on sentiment information. To rank products in this work, we compare three methods for textual analysis and two baselines.

Two approaches are based on counting words or phrases with a positive and negative polarity. The first assigns these polarities based on a dictionary in which the respective class is explicitly stated. The polarity score $\mathrm{score}(p)$ for a product p is then calculated as the number of all positive words (pos) in all reviews for this product minus the number of all negative words (neg):

$$\mathrm{score}_{\text{dict}}(p) = \mathrm{pos}(p) - \mathrm{neg}(p)\,. \qquad (1)$$

To account for the impact of longer reviews, we normalize these numbers by the number of tokens in all reviews for the specific product all_p:

$$\overline{\mathrm{score}}_{\text{dict}}(p) = \frac{\mathrm{score(p)}}{\mathrm{all}_p}\,. \qquad (2)$$

The ranked list of products is then created by sorting according to this score. We refer to the two variations of this method as DICT and DICT-NORM.

This first dictionary-based method is easy to implement and to use. However, it might not take into account context specific formulations of polarity expressions. As a second method, we therefore opt for a machine learning-based detection of subjective phrases with their polarities in context, specifically we use JFSA (Joint Fine-Grained Sentiment Analysis Tool, Klinger and Cimiano (2013)[1]). Calculating the product score and ranking is performed analogously to the dictionary-based approach. We refer to the two variations of this method as JFSA and JFSA-NORM.

As our goal is to ultimately generate a ranked list of products, it is a straight-forward idea to exploit textual comparison expressions, as in this example:

$$\underbrace{\text{It}}_{\substack{\text{entity}\\\text{(preferred)}}}\ \text{has a}\ \underbrace{\text{better}}_{\text{predicate}}\ \underbrace{\text{lens}}_{\text{aspect}}\ \text{than}\ \underbrace{\text{the T3i}}_{\substack{\text{entity}\\\text{(not preferred)}}}\ .$$

To extract such comparisons, we employ CSRL (Comparison Semantic Role-Labeler, Kessler and

[1] https://bitbucket.org/rklinger/jfsa

Kuhn (2013)). The system identifies comparative predicates ("better"), the two entities that are involved ("It" and "the T3i"), which one is preferred ("It"), and the compared aspect ("lens"). To identify the products that are referenced, we associate a mentioned entity to the product name (or names) with the minimal cosine similarity on token level. In the example, "the T3i" would be associated with the camera "Canon EOS Rebel T3i". The pronoun "It" is mapped to the reviewed product.

The score for a product is calculated based on the number of times it occurs as a preferred product (pref) minus the number of times it occurs as a non-preferred product (npref):

$$\mathrm{score}_{\mathrm{CSRL}}(p) = \mathrm{pref}(p) - \mathrm{npref}(p)\,. \quad (3)$$

The resulting score for a product is used for sorting analogously to the previous approaches. We refer to this method as CSRL.

We use two baselines that do not take the textual information of a review into account: The first method sorts products by their average star rating (from one to five stars, as assigned by the author of a review) of all reviews for the respective product (STAR). The second method sorts the products by the number of reviews it has received (from none to many, NUMREVIEWS). The intuition is that products which are sold more often gather more reviews.

Two of our methods, JFSA and CSRL recognize aspects of products together with a subjective phrase or comparison, respectively. Besides creating one ranking that is a combined measure of all aspects of the product, we have the option to use only evaluations regarding specific aspects which results in an aspect-specific ranking. As one aspect can be referred to with several expressions, a normalization of the aspect mentions is needed for this filtering. In the experiments in this paper, we use manually compiled lists of textual variations for the most frequent aspects in our dataset[2]. In the target-specific version of a method, subjective phrases or entity mentions are only counted towards the score of a product if there is a token overlap between the recognized aspect and a textual variation of the target aspect.

[2]The lists for aspect mention normalization are available as supplementary material. For instance, *video* contains "video", "videos", "film", "films", "movie", "movies", "record", "records", "recording".

Method	Amazon	Snapsort
STARS	-0.027	$\mathbf{0.436}*$
NUMREVIEWS	$\mathbf{0.331}*$	0.095
DICT-NORM (GI)	$0.125*$	-0.148
DICT-NORM (MPQA)	$0.142*$	-0.145
DICT (GI)	$0.219*$	$0.426*$
DICT (MPQA)	$0.222*$	$0.441*$
JFSA-NORM	$0.151*$	-0.230
JFSA	$\mathbf{0.234}*$	$0.404*$
CSRL	$0.183*$	$\mathbf{0.511}*$

Table 1: Results (Spearman's ρ) of the target-agnostic methods for predicting the sales ranking of Amazon and the Snapsort quality ranking. Significance over random is marked with * ($p < 0.05$). The best baseline and the best text-based method are marked in bold.

4 Experiments

4.1 Experimental setting

For evaluation, we use camera reviews retrieved from Amazon with the search terms "camera" and "camera" in conjunction with "fuji", "fujifilm", "canon", "panasonic", "olympus", "nikon", "sigma", "hasselblad", "leica", "pentax", "rollei", "samsung", "sony", "olympus". As the first gold ranking, we extract the Amazon sales rank from the product descriptions ("Amazon Best Sellers Rank" in the "Camera & Photo" category) as retrieved between April 14th and 18th, 2015 and include only products for which a rank is provided. The resulting list contains 920 products with a total of 71,409 reviews. Product names are extracted from the title of the page and shortened to the first six tokens to remove additional descriptions.

As a second external gold ranking, we use the quality ranking provided by Snapsort. From the top 150 products in the Amazon sales ranking, 56 are found on Snapsort. We use the rank in the category "best overall" of "all digital cameras announced in the last 48 month" as retrieved on June 12th, 2015.[3]

JFSA is trained on the camera data set by Kessler et al. (2010). CSRL is trained on the camera data by Kessler and Kuhn (2014). For the methods DICT and DICT-NORM, we try two different sources of opinion words, the general

[3]The full list of products with their names and the rankings are available in the supplementary material.

Aspect	#	ρ	σ
performance	637	0.301	0.009
video	600	0.278	0.013
size	513	0.218	0.017
pictures	790	0.213	0.003
battery	541	0.208	0.012
price	625	0.198	0.008
zoom	514	0.196	0.013
shutter	410	0.191	0.016
features	629	0.190	0.009
autofocus	403	0.175	0.013
screen	501	0.136	0.012
lens	457	0.099	0.012
flash	591	0.093	0.011

Table 2: Results (Spearman's ρ and standard deviation σ) of JFSA for predicting the Amazon sales ranking when only the subjective phrases are taken into account which refer to the specified target aspect. The number of products for which at least one evaluation of the target aspect is found is shown in column #.

inquirer dictionary (Stone et al., 1996)[4] and the MPQA subjectivity clues (Wilson et al., 2005)[5].

To measure the correlation of the rankings generated by our different methods with the gold ranking, we calculate Spearman's rank correlation coefficient ρ (Spearman, 1904). We test for significance with the Steiger test (Steiger, 1980).

4.2 Results

As described in Section 2, we take into account two different rankings for evaluation: The Amazon.com sales ranking contains 920 products and is an example for a ranking that may be useful for sales managers or product designers. The second is the expert ranking by Snapsort.com which contains 56 products. These two rankings are conceptually different. There is no correlation between the two rankings ($\rho = -0.04$).

Table 1 shows the results for the baselines and the target-agnostic methods on the gold rankings. There is a pronounced difference between the results for the two gold rankings.

The best result on Amazon (significantly outperforming all other methods) is achieved by counting the reviews ($\rho = 0.33$, NUMREVIEWS). For Snapsort, however, NUMREVIEWS leads to only $\rho = 0.1$. One factor that explains this difference in performance is the fact that in case of Amazon the reviews and the ranking come from the same source and it is unclear whether the popularity of a product leads to many reviews or a high number of reviews leads to higher sales. Though "popularity" is one aspect that influences the Snapsort rating, it is not as prominent.

The performance of the STARS baseline is not significantly different from random for Amazon. This is partly explained by the fact that among the products with a 5.0 star rating many have only very few reviews (less than 10). This is less of a problem in the Snapsort ranking. Also, we would expect that what is contained in the reviews are quality judgements that are more closely aligned with what Snapsort uses for ranking than what influences sales.

The dictionary-based polarity assignment based ranking (DICT) approximates the sales ranking with $\rho = 0.22$, for both MPQA and GI. Normalization of the polarity scores reduces the correlation. The similarity of the results obtained by the two different dictionaries is reflected in the very high correlation of the resulting rankings (without normalization: $\rho = 0.99$; with normalization: $\rho = 0.8$). However, the non-normalized rankings are not correlated with the normalized rankings of the same dictionary (GI $\rho = -0.16$, MPQA $\rho = -0.14$).

The dictionary-based ranking is slightly outperformed by JFSA with $\rho = 0.23$. Normalization by token number (and therefore implicitly the review count) decreases the performance to $\rho = 0.15$. The difference of JFSA to DICT-NORM (GI) and DICT (MPQA and GI) is significant ($p < 0.05$). For Snapsort, normalization has a strong negative impact.

On Amazon, the ranking achieved with CSRL is mediocre in comparison to the other methods. CSRL suffers more clearly from data sparseness (the highest number of subjective phrases for a product found by JFSA is over 9000, while the highest number of comparisons that mention a given product is 662 for CSRL). On the Snapsort ranking however, CSRL leads to the best result of all experiments with $\rho = 0.51$.

In comparison to using all information extracted from reviews to generate a ranking, the aspect-specific results allow for an understanding of the

[4] 3518 entries; 1775 positive, 1743 negative using the categories from Choi and Cardie (2008).

[5] 6456 entries; 2304 positive, 4152 negative.

impact of each aspect on the gold ranking. Aspect-specific rankings for important aspects are highly correlated with the gold ranking, while those for completely irrelevant aspects have a correlation near random. The results for the Amazon sales ranking and JFSA are shown in Table 2. Due to data sparseness, a substantial amount of products receive a score of 0. To eliminate the resulting artificial inflation of ρ while enabling a comparison between methods with different numbers of scored products, we add the zero-scoring products in random order and average over 100 different ranked lists. We omit the results for CSRL and the results on Snapsort which are all close to random.

For the ranking created with JFSA, the aspect *performance* contributes most to approximating the sales ranking ($\rho = 0.30$) followed by *video* ($\rho = 0.28$). Both results outperform the target-agnostic ranking of JFSA ($\rho = 0.23$) (significant for *performance*).

5 Conclusion and Future Work

We have presented the task of predicting a ranking of products and introduced three potential sources for gold rankings: A sales ranking and expert based ranking have been used in the experiments in this paper. In addition, we discussed how to set up a crowdsourcing-based annotation of rankings. We demonstrated early results how to use different opinion mining methods (dictionary-based, machine learning, comparison-based) to predict such rankings. In addition, we have presented experiments on how aspect-specific rankings can be used to measure the impact of that specific information on the ranking.

The methods discussed here show a limited performance, however, these results of approximating a real world ranking are promising and encouraging for further research. Though the correlation scores are comparatively low, they allow for an analysis of the influence of a specific aspect on the ranking as shown for the Amazon sales ranking.

The best result for the Amazon sales ranking is achieved based on the number of reviews (NUMREVIEWS). This might be seen as an instance of the chicken-egg dilemma, and it may be the case that there are many reviews *because* the product has been sold many times. The same effect cannot be observed on Snapsort. It is further worth noting that the average star rating (STARS) is not informative towards Amazon sales ranking, but gives good results on Snapsort.

The methods which take into account the polarity of phrases lead to the second best performance (JFSA and DICT) for Amazon. For Snapsort, the comparison-based CSRL is outperforming all other methods and shows the highest performance of all experiments in this paper ($\rho = 0.51$).

For future work, we plan to formulate the problem in a learning-to-rank setting with data generated in a crowdsourcing paradigm to combine the different measures discussed in this paper and allow for a straight-forward adaptation to different rankings.

Acknowledgments

We thank Christian Scheible for fruitful discussions. Wiltrud Kessler has been supported by a Nuance Foundation grant.

References

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *EMNLP*, pages 793–801. ACL.

Yoonjung Choi, Seongchan Kim, and Sung-Hyon Myaeng. 2010. Detecting Opinions and their Opinion Targets in NTCIR-8. In *NTCIR-8 Workshop Meeting*, pages 249–254.

Kavita Ganesan and ChengXiang Zhai. 2012. Opinion-based entity ranking. *Information Retrieval*, 15(2):116–150.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *ACM SIGKDD*, pages 168–177.

Niklas Jakob and Iryna Gurevych. 2010. Using anaphora resolution to improve opinion target identification in movie reviews. In *ACL*, pages 263–268.

Richard Johansson and Alessandro Moschitti. 2011. Extracting opinion expressions and their polarities: exploration of pipelines and joint models. In *ACL-HLT*, pages 101–106.

Wiltrud Kessler and Jonas Kuhn. 2013. Detection of product comparisons - How far does an out-of-the-box semantic role labeling system take you? In *EMNLP*, pages 1892–1897. ACL.

Wiltrud Kessler and Jonas Kuhn. 2014. A corpus of comparisons in product reviews. In *LREC*, pages 2242–2248.

Jason S. Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. 2010. The 2010 icwsm jdpa sentment corpus for the automotive domain. In

4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010).

Roman Klinger and Philipp Cimiano. 2013. Bi-directional inter-dependencies of subjective expressions and targets and their value for a joint model. In *ACL*, pages 848–854.

Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Sentiment analysis with global topics and local dependency. In *AAAI*, pages 1371–1376.

Bing Liu. 2015. *Sentiment Analysis*. Cambridge University Press.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, pages 271–278.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *HLT-EMNLP*, pages 339–346. ACL.

Asad Sayeed, Jordan Boyd-Graber, Bryan Rusk, and Amy Weinberg. 2012. Grammatical structures for word-level sentiment detection. In *NAACL-HLT*, pages 667–676. ACL.

Charles Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1996. *General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Oscar Täckström and Ryan McDonald. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *ACL-HLT*, pages 569–574.

Maksim Tkachenko and Hady W. Lauw. 2014. Generative modeling of entity comparisons in text. In *CIKM*, pages 859–868.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT*, pages 347–354. ACL.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *EMNLP-CoNLL*, pages 1335–1345. ACL.