Neural Networks 66 (2015) 36–45

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Diversifying customer review rankings

Ralf Krestel [a,*], Nima Dokoohaki [b]

[a] Department of Computer Science, University of California, Irvine, USA

[b] Decisions, Networks and Analytics (DNA), Swedish Institute of Computer Science (SICS), Stockholm, Sweden

ARTICLE INFO

Article history:
Received 31 August 2013
Received in revised form 27 December 2014
Accepted 18 February 2015
Available online 26 February 2015

Keywords:
Ranking
Topic models
Summarization
Diversification
Review recommendation

ABSTRACT

E-commerce Web sites owe much of their popularity to consumer reviews accompanying product descriptions. On-line customers spend hours and hours going through heaps of textual reviews to decide which products to buy. At the same time, each popular product has thousands of user-generated reviews, making it impossible for a buyer to read everything. Current approaches to display reviews to users or recommend an individual review for a product are based on the recency or helpfulness of each review.

In this paper, we present a framework to rank product reviews by optimizing the coverage of the ranking with respect to sentiment or aspects, or by summarizing all reviews with the top-K reviews in the ranking. To accomplish this, we make use of the assigned star rating for a product as an indicator for a review's sentiment polarity and compare bag-of-words (language model) with topic models (latent Dirichlet allocation) as a mean to represent aspects. Our evaluation on manually annotated review data from a commercial review Web site demonstrates the effectiveness of our approach, outperforming plain recency ranking by 30% and obtaining best results by combining language and topic model representations.

1. Introduction

It has become a routine among on-line and off-line consumers to inform themselves on review platforms before purchasing a certain product. This has given rise to a considerable amount of customer reviews on e-commerce Web sites. With this in mind, potential customers usually browse through a lot of on-line reviews in order to build confidence in a particular item prior to purchasing it. While reviews have become an important factor in helping Web crowds to further assess the quality of products online, the increase in volume of review data has led to an information overload. Popular products have thousands of reviews. While excess of reviews is a growing problem, recommending unbiased and helpful texts is a growing research field. The quality of reviews may vary drastically and might mislead potential buyers. And the humongous amounts of information not only distracts the confidence seeker, it might also hinder the original goal of users in the first place: They will give up buying a certain product. To deal with these problems, review recommendation techniques are proposed. Review recommendation involves implementing machine-learning techniques for analyzing product reviews based on their lexico-semantic features in order to classify the reviews and offer a balanced and useful view of the reviews to the reader.

While review recommender systems aim at automatic classification of reviews, some commercial Web sites such as Amazon and TripAdvisor[1] approach this problem by allowing users to rate the reviews using star ratings to improve the rankings (e.g. *this review was helpful vs. not helpful*). There are two inherent problems to these rankings based on user feedback: First, good objective reviews contain quite likely redundant information and ranking them based on the helpfulness score will not cover all aspects. Second, these Web sites do not take into account the personal bias. Not all reviews are helpful to everybody. Due to the fact that different users put different emphasis on different aspects, (e.g. *I do not care about battery life, but really need lots of memory*), helpfulness can only be used to filter out very badly written reviews. Therefore, researchers are increasingly distinguishing between the task of review recommendation (Aciar, Zhang, Simoff, & Debenham, 2007), and review ranking (Ghose & Ipeirotis, 2007). To improve existing review recommendation techniques and at the same time improve the ranking of reviews, we propose a novel approach to model and rank reviews. The two main components of

[*] Corresponding author.
E-mail address: krestel@L3S.de (R. Krestel).

[1] http://www.amazon.com and http://www.tripadvisor.com.

http://dx.doi.org/10.1016/j.neunet.2015.02.008
0893-6080/© 2015 Elsevier Ltd. All rights reserved.

our system rely on latent Dirichlet allocation to model the reviews and on Kullback–Leibler divergence to generate an adequate ranking. We make use of the assigned star rating for the product as an indicator of the polarity expressed in the review. Our framework covers different ranking strategies based on users' needs and can adapt to various user scenarios. We currently support three strategies: summarizing all existing reviews; focusing on a particular latent topic; or focusing on positive, negative or neutral aspects. We evaluated the system using manually annotated review data gathered from a popular review Web site.

The main contributions of this paper are: (1) Introducing an algorithm to model reviews using latent topics and user-assigned product ratings. (2) Ranking of reviews to summarize all reviews for a product within the top-K results. (3) Diversification of review rankings based on star ratings and/or latent topics. The remaining of the paper is organized as follows: We present related work in Section 2; Section 3 gives an overview of our framework. Section 4 describes the modeling approach, while Section 5 describes the ranking approach. We present the evaluation in Section 6 and close with conclusions and future work.

2. Overview of the field

State-of-the-art in product review mining can be categorized into two major themes: summarization and ranking.

Summarization of reviews. Existing literature on review summarization techniques have a strong focus on review classification and recommendation (Dave, Lawrence, & Pennock, 2003). While reviews in general have been the focus of the majority of works in this field, a breed of new work focuses on opinion mining while taking online reviews as case studies. Review summaries include structured summaries of review text that provide an organized breakdown by aspects or topics, and various formats of sentiment and sentence summaries. Various summary formats complement each other by providing a different level of understanding. For instance, sentiment prediction on reviews for a product can provide a very generic picture of how users feel about a certain product. While users requiring more specific details can turn to topic-based or sentence/sentiment summaries instead. Two state-of-the-art studies about opinion summarization by Liu (2010) and Pang and Lee (2008) give a broad overview of the field. Both surveys cover previous as well as current work but their focuses vary. According to these surveys, review summarization falls under subjective classification, sentiment analysis, or under traditional text summarization. While researchers differentiate between review summarization methods and classic text summarization techniques (Zhuang, Jing, & Zhu, 2006), the connection is obvious. Both aim at identifying salient information: terms, sentences, or paragraphs. Sentiment analysis techniques try to produce a summarized sentiment consisting of sentences from a source document, a single paragraph (Beineke, Hastie, Manning, & Vaithyanathan, 2003), a structured sentence (Hu & Liu, 2004), attribute-value pairs, or just a sentiment score. To build summaries of sentence list structures, Hu and Liu (2004) introduced a method utilizing word attributes such as frequency of occurrence, part-of-speech tagging and WordNet synsets. Following this approach features are extracted, combined with their contextually close words, and finally used to generate a summary by selecting and re-structuring the sentences following the extracted features. Another approach called Opine (Popescu & Etzioni, 2005) uses relaxation labeling to find the lexico-semantic orientation of words, whereas Pulse (Gamon, Aue, Corston-Oliver, & Ringger, 2005) uses bootstrapping to train a sentiment classifier using features extracted by labeling sentence clusters with respect to their key terms. SumView (Wang, Zhu, & Li, 2013) is a semi-supervised Web application capable of review crawling along with automatic product feature extraction. Users can query features of their interest, which are processed in turn using a sentence selection along with the proposed feature-based weighted non-negative matrix factorization (NNMF) algorithm. Finally, the most characteristic set of sentences are selected to summarize the nominal features of each product. Cambria, Schuller, Xia, and Havasi (2013) shed light on new avenues on sentiment analysis and opinion mining by weighing on the notion of concept level analysis in comparison to topic level analysis. Kim, Ganesan, Sondhi, and Zhai (2011), classify existing approaches under two main categories: aspect oriented summarization and non-aspect oriented summarization. The most common category of opinion summarization technique is aspect-based opinion summarization, which involves generating opinion summaries containing a set of topics (also known as aspects or features). Aspect-based summarization involves three steps: feature identification, sentiment prediction, and summary generation. Non-aspect oriented summarization includes: sentiment summarization (Chaovalit & Zhou, 2005), basic and advanced text summarization (Kim & Zhai, 2009) and entity-based summarization (Stoyanov & Cardie, 2008). Recently hybrid models are also proposed which aim at combining aspect and sentiment models in what authors refer to as joint aspect/sentiment models (Lin & He, 2009; Moghaddam & Ester, 2011).

In our work we mine and summarize reviews by choosing complementing reviews and ranking them according to different strategies. The product ratings serve as an indication of sentiment, and the extracted latent topics ensure topical coverage of relevant aspects.

Diversification of reviews and their rankings. The problem of personalized ordering of results has been subject to research in both classic retrieval of documents as well as within recommender systems research. A first approach by Carbonell and Goldstein (1998) based on maximum marginal relevance (MMR) was used as a ranking metric which balances relevance as the similarity between query and search results with diversity as the dissimilarity among search results. Ziegler, McNee, Konstan, and Lausen (2005) take into account a user's full range of interests through diversifying generated recommendation lists and by doing so they minimize redundancy among the recommended items.

While most existing work focuses on the task of diversification of search results, there is also some recent work on review mining. Yu, Zha, Wang, and Chua (2011) look at ranking aspects of reviews. The aspect ranking algorithm identifies important aspects by taking into account the aspect frequency and influence of consumers opinions given to each aspect. When evaluating sentiment classification and aspect rating, they report better Kullback–Leibler divergence (KLD) compared to Hu and Liu (2004). Similar to our work, Xu, Meng, and Cheng (2011) state that two requirements should be taken into account while generating a good summary: representativeness and diversity, in addition to aspect-relevance and sentiment intensity. They present an aspect-based summarization method for online reviews, that incorporates an aspect-sensitive Markov random walk model to satisfy the representativeness requirement, as well as a greedy redundancy to meet the diversity requirement.

We propose a greedy algorithm to minimize the Kullback–Leibler divergence between the topic models of the top-K ranked reviews and all reviews for a product. In addition, we diversify review rankings based on latent topics and language models (LM) to get an optimal coverage for all topics within the top-K results.

3. How to rank reviews?

In contrast to Web search results, reviews for a product cannot be ranked based mainly on relevance since all reviews are supposed to be equally relevant for the product the review is about.[2] As discussed in Section 2, review recommendation or classification is a well-studied problem. However, most approaches do not optimize a ranking of reviews but evaluate the reviews individually. Our goal is not to find the best or most helpful individual review for a product but to *find the top-K reviews which provide the user with a good summary* of the opinions about a product. To this end, we model reviews using language models (LM) (Ponte & Croft, 1998) and latent topics extracted with latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003). We combine these models with the user-assigned star ratings associated with each review. The reviews are ranked in a greedy fashion to minimize the Kullback–Leibler divergence (KLD) between the reviews already ranked and all reviews for a product. This ensures that the final ranking covers all popular topics and reflects the overall rating for the product. Our framework also allows to set a different objective to optimize the ranking, e.g. cover all positive aspects of a product, or cover all sentiments associated with one particular aspect of the product. In the following section we describe the conceptional architecture of our framework in more detail.

3.1. System overview

Our framework consists of two main components to reflect the two conceptual steps of (1) modeling the reviews and (2) creating the rankings:

1. An LDA/LM component generates the topic and/or language models of the review data;
2. A ranking component optimizes the ranking based on a user-defined strategy.

A graphical overview of the framework can be seen in Fig. 1. To model the reviews of a product, the LDA/LM component takes all reviews written for this product, together with the associated user ratings assigned by the review authors. Our hypothesis is that users who assign five stars (on a five point Likert-scale) mainly talk about positive experience with the product or its features, whereas a review accompanied by a 1-star rating indicates a review with rather negative points.[3] Since we do not want to exclude the possibility that also in a 5-star review a minor negative point could be expressed, we smooth the assignment of reviews to rating classes and assign probabilities for each document belonging to each rating class based on a right stochastic matrix. Especially a review with a 3-star rating can contain negative as well as positive opinions which can be modeled using this matrix by assigning this 3-star review with 50% to the positive class and with 50% to the negative class.

Based on the topic model for each review we rank the reviews by minimizing the Kullback–Leibler divergence between the top-K aggregated reviews of the ranking and all reviews to summarize the topics of all reviews for a product. In this case, our target distribution that we try to approximate is the aggregated topic distribution of all reviews. We also investigate two other target distributions to optimize the ranking with respect to coverage of one topic and coverage of either all positive or negative aspects. After discussing the preprocessing steps in the next section, we describe the modeling approach in Section 4 and the ranking in Section 5.

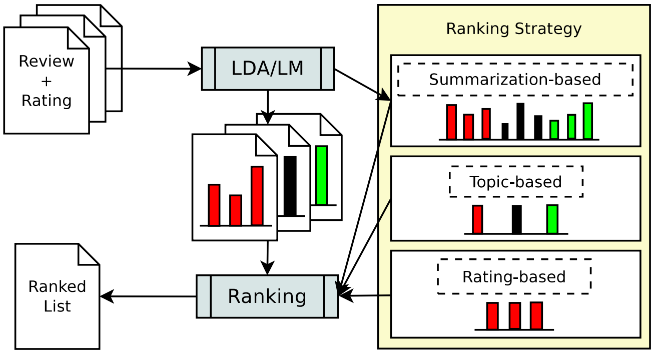


Fig. 1. Overview of the review ranking system: reviews together with ratings are used to extract topic distributions using LDA or word distributions using LM. Rankings are computed minimizing KL-Divergence with task-specific target distributions.

3.2. Preprocessing

Since reviews are user-generated, they may contain grammatical errors, sloppy language, and spelling errors. Therefore, preprocessing the raw data is an important step we briefly want to elaborate on. We used the Stanford POS Tagger (Toutanova, Klein, Manning, & Singer, 2003) for tokenization and part-of-speech tagging. Then WordNet (Fellbaum, 1998) was used to get the lemmas of the terms, and finally, all terms that are not verbs, adverbs, nouns, or adjectives are removed, thus we are getting rid of most stop words.

Since unigrams might not give an accurate picture of what a review is about we extract *n*-grams of variable lengths in the next step. Especially in the context of product reviews, multi-term phrases are important to model the data, e.g. "Microsoft Windows 7 Professional", "not recommended", or "graphic card" denote meaningful *n*-grams that should be considered as single semantic units. Therefore, we partition our data into meaningful *n*-grams first. Based on the work of Deligne and Bimbot (1995), we compute multigram models for the documents in our corpus the following way: Each sentence is considered a sequence of *n*-grams with variable length. The likelihood of a sentence is computed by summing up the individual likelihoods of the *n*-grams corresponding to each possible segmentation of the sentence. This is done using a Viterbi-like algorithm to find the maximum likelihood segmentation. In an iterative fashion, we re-estimate and update the probabilities until convergence.

As a result, all documents in our corpus of product reviews are segmented into variable-length *n*-grams and the latent topics can now be based on *n*-grams or phrases instead of fixed sized units or single terms. Table 1 shows a snippet of an original product review together with its preprocessed version without stop words but with part-of-speech information and multigrams.

4. Modeling reviews

To model the review data we make use of language models and topic models. We combine this information with the assigned star ratings for the reviews to cover positive and negative statements associated with a particular latent topic.

4.1. Finding latent topics

A product review usually covers different aspects or features of a product. For example, users have an opinion about the price of a product or the service of a company. Instead of a fine-grained extraction of features and sentiment, as done for instance by Bross

[2] We do not consider in our setting obvious spam or fake reviews, but rely on the review platform to take care of those.

[3] In Section 6.2 we back up this hypothesis by analyzing user-assigned ratings and review content.

Table 1
Review snippet: original (left); segmented and POS-tagged (right).

Review snippet	Bag of words
"…Wish burger—also know as a veggie burger (no meat) Ketchup and Mustard are actually available at In and Out.. just ask, its really easy Double Meat is la double double with no cheese Flying Dutchman …"	wish.n burger.n also.r know.v veggie.n_burger.n meat.n actually.r available.a ketchup.n mustard.n just.r ask.v really.r easy.a double.r_meat.n double.a_double.a cheese.n flying.n_dutchman.n

Table 2
Top terms composing the latent topics "ticket" (left) and "waiting" (right) for America West Airlines.

Term	Prob.	Term	Prob.
ticket.n	0.038	concourse.n	0.015
voucher.n	0.027	miss.v	0.015
clerk.n	0.016	take.v_off.r	0.015
care.v	0.011	hour.n_late.r	0.012
availability.n	0.008	change.n	0.009
complain.v	0.008	delay.n	0.009
look.v	0.008	flight.n_attendant.n	0.009
nightmare.n	0.008	meeting.n	0.009
suggest.v	0.008	not.r	0.009
america.n_worst.r	0.009	reggie.n	0.009

and Ehrig (2010), we rely on a statistical approach to find features or aspects. We therefore make use of latent Dirichlet allocation, which models each review as a mixture of latent topics.[4] In a training phase, LDA learns word clusters (topics) by examining co-occurrence of words in documents. This probabilistic assignment of different topics to a single review allows later to identify topically similar reviews. In case no suitable training corpus is available, methods such as commonsense-based topic modeling (Rajagopal, Olsher, Cambria, & Kwok, 2013) could be adapted for the product review domain to provide topic assignments for reviews. Given the huge amount of review data on the Web, we can use LDA to find latent topics. Note that our approach does not depend on a specific topic modeling algorithm and LDA could be replaced by commonsense-based topic modeling, latent semantic indexing (Hofmann, 1999) or random indexing (Kanerva, Kristoferson, & Holst, 2000) using a bag-of-concepts model.

LDA identifies a given number of $|Z|$ topics within a corpus of $|D|$ documents. Each term w in a review with N_d terms is associated with a topic z. Being the most important parameter for LDA, the number of latent topics $|Z|$ determines the granularity of the resulting topic model. In order to find the latent topics, LDA relies on stochastic modeling. The modeling process of LDA can be described as determining a mixture of topics for each document in the corpus, i.e., $P(z \mid d)$, with each topic described by multigrams following another probability distribution, i.e., $P(w \mid z)$. This can be formalized as:

$$P(w_i \mid d) = \sum_{j=1}^{|Z|} P(w_i \mid z_j) P(z_j \mid d),$$

where $P(w_i \mid d)$ is the probability of the ith multigram for a given document d and z_i is the latent topic. $P(w_i \mid z_j)$ is the probability of w_i within topic z_j. $P(z_j \mid d)$ is the probability of picking a term from topic z_j in the document.

With LDA at hand, we are able to represent latent topics as a list of multigrams with a probability for each multigram indicating the membership degree within the topic. Furthermore, for each document in our corpus (reviews in our case) we can determine to which topics it belongs, also associated with a degree of membership (topic probability $P(z_j \mid d_i)$).

An example for two extracted latent topics represented by the top 10 terms is shown in Table 2. Beside the terms, also the probability for the terms belonging to the topic are shown. For this example we used $|Z| = 50$ latent topics.

4.2. Building language models

In its simplest form, a language model for a document d with words w_i can be computed using a maximum likelihood estimate:

$$P(w \mid d) = \frac{c(w, d)}{\sum_{w_i \in d} c(w_i, d)}$$

where $c(w, d)$ is the count of word w in document d. To prevent the probability $P_{lm}(w \mid d)$ from being zero in case word w does not appear in document d, various smoothing methods have been introduced and compared, e.g. by Zhai and Lafferty (2001). The un-smoothed model using a maximum likelihood estimate can be complemented by different components, with the Laplace smoothing being the simplest, adding 1 to each count:

$$P(w \mid d) = \frac{c(w, d) + 1}{\sum_{w_i \in d} c(w_i, d) + 1}.$$

Apart from this simple language model based on the bag-of-words assumption, more sophisticated language models have been proposed in the past, e.g. based on n-grams (Brown, deSouza, Mercer, Pietra, & Lai, 1992). A possible extension to our approach could be to consider phrases instead of fixed multi-grams as a starting point in our analysis.

With LDA and LM, each review d can now be described as a multinomial distribution either over latent topics $P(z_i \mid d)$ (LDA) or over terms $P(w_i \mid d)$ (LM). In the next section we describe how to incorporate the user ratings into these probability distributions.

4.3. Combining models and star ratings

Instead of performing sentiment analysis on a word, sentence, or document level, we use the user-supplied star rating as an indicator for the sentiment expressed in a review. Thus we know which aspects or topics a review is about and we know the overall sentiment of a user towards the reviewed product.

For LDA, we start with the Z original topics and are interested in the relationship of the topics and the assigned ratings $r \in R = \{1, \ldots, 5\}$. Thus, each topic is associated with each possible rating and the probability $P(z, r \mid D)$ illustrates how likely it is that a particular topic occurs with a particular rating in the whole corpus D. For one document d the likelihood, assuming independence between the topic and the rating, can be computed as:

$$P(z, r \mid d) = P(z \mid d) P(r \mid d)$$

with $P(z \mid d)$ being the topic distribution for document d and $P(r \mid d)$ being a generalized Bernoulli distribution $M(|R|, \Phi_d)$ over ratings that maps the user-provided hard rating for review d into probability space. When we put all Φ_d together, we get a right stochastic matrix with D rows and R columns. If the dataset contains user information – next to the review and the rating – then this matrix can be used to account for different user behavior,

[4] We use the LDA implementation by McCallum (2002), which makes use of Gibbs sampling to compute the latent topics.

e.g. users who tend to rate everything with the most negative rating. Since we do not correct for these individual user biases, we only map the ratings based on the originally assigned rating class. This leaves us with a five by five right stochastic matrix. In practice, the values of this matrix depends on the dataset and the typical user rating behavior on a review platform.

As a baseline, we combine language models with the rating information and compute a similar likelihood based on the words instead of the topics:

$$P(w, r \mid d) = P(w \mid d)P(r \mid d).$$

This represents the language model of a document for each rating class. Similarly we have a topic model representation of each document for each rating class. This means each review is modeled as either a topic or a word mixture depending on its rating with some overlap between the classes according to the Φs. In the following we will describe only the topic-rating model case; for the language model representation, the process is analogous.

5. Ranking reviews

The ranking of the reviews can follow different strategies depending on the user's information need. We propose these three ranking strategies:

1. Summary-focused Ranking (Section 5.1).
2. Sentiment-focused Ranking (Section 5.2).
3. Topic-focused Ranking (Section 5.3).

Each strategy is associated with a different target distribution and the objective function that we want to minimize is the Kullback–Leibler divergence of the top-K entries of the ranking and the given target distribution Φ_t.

Kullback–Leibler divergence estimates the number of additional bits needed to encode a distribution U, using an optimal code for Q, and having a combined vocabulary size of V:

$$KLD(U \parallel Q) = H(U, Q) - H(U) = \sum_{i=1}^{|V|} u_i * \log_2 \left(\frac{u_i}{q_i} \right).$$

In our setting, distribution Q is the combined topic-rating model of the top-K reviews and Q is the target distribution Φ_t. Thus $KLD(U \parallel Q)$ can be directly used to measure the similarity of the top-K entries of the ranking and the target distribution. ($H(U, Q)$ is the cross entropy of U and Q and $H(U)$ is the entropy of U). The greedy algorithm used to compute the rankings based on a strategy with target distribution Φ_t is shown in Algorithm 1.

```
Input: A Set of customer reviews $D_p$ for product $p$ and a target distribution $\Phi_t$
       based on the ranking strategy
Output: A Ranking $C_p = \{d_1, d_2, \dots, c_k\}$ of customer reviews $d_i \in D_p$ for
        product $p$ with $rank(d_1) < rank(d_2) < \dots < rank(d_r)$
$C_p \leftarrow \emptyset$
for $i \leftarrow 1$ to $k$ do
    $d_{min} \leftarrow d_i$
    for $j \leftarrow 1$ to $|D_p| - i$ do
        if $KLD(P(d_j \cup C_p)||\Phi_t) < KLD(P(min \cup C_p)||\Phi_t)$ then
            $d_{min} \leftarrow d_j$
        end
    end
    $C_p \leftarrow C_p \cup d_{min}$
end
return $C_p$
```

Algorithm 1: Greedy algorithm for minimizing Kullback–Leibler divergence with respect to target distribution Φ_t

This algorithm can also be used to simultaneously optimize the ranking based on topic model distance and language model distance. Using LM + LDA we ensure to not only cover the relevant abstract topics but also the relevant terms. The function to minimize in this case is:

$$KLD(P(z, r|D_K) \parallel \Phi_t^z) + KLD(P(w, r|D_K) \parallel \Phi_t^w)$$

with Φ_t^z being the target distribution in topic space and Φ_t^w being the target distribution in word space.

5.1. Summary-focused ranking

In most cases, users reading reviews are interested in getting an overview of the experiences of other users with the product. A ranking which gives a good overview summarizes the views expressed in all reviews. The goal for a review ranking system is therefore to approximate all reviews by the top-K in the ranking. Thus, the top-K reviews *summarize* the opinions about a product present in all reviews.

With the topic-rating models computed for each review we try to find a ranking of reviews that approximates the aggregated topic-rating models of all reviews for a product. This means we try to minimize the Kullback–Leibler divergence between the top-K ranked reviews and the aggregation of all reviews and the target distribution $\Phi = P(z, r|D)$:

$$KLD(P(z, r|D_K) \parallel P(z, r|D)) = \sum_{i=1}^{|Z|} \sum_{i=1}^{|R|} P(z_i, r_j|D_K) \log_2 \left(\frac{P(z_i, r_j|D_K)}{\frac{1}{|D|} \sum_{k=1}^{|D|} P(z_i, r_j|d_k)} \right)$$

with $P(z_i, r_j|D_K) = \frac{1}{|D_K|} \sum_{k=1}^{|D_K|} P(z_i, r_j|d_k)$ being the aggregated topic model of the first K documents.

5.2. Sentiment-focused ranking

Instead of approximating all reviews, the sentiment-focused ranking tries to summarize only one particular class of ratings, for example negative aspects as represented by the topic-rating model with rating one. It could also be interesting to see which features of a product are discussed mainly in a neutral review or which aspects are only discussed in positive reviews. Depending on the rating smoothing matrix aspects from reviews having a slightly different rating can influence the ranking.

The target distribution that we try to approximate with the review ranking in this case is a uniform distribution over all topics for one rating $\Phi = DU(|Z|)$. That means we get a diverse ranking covering all latent topics associated with a particular rating:

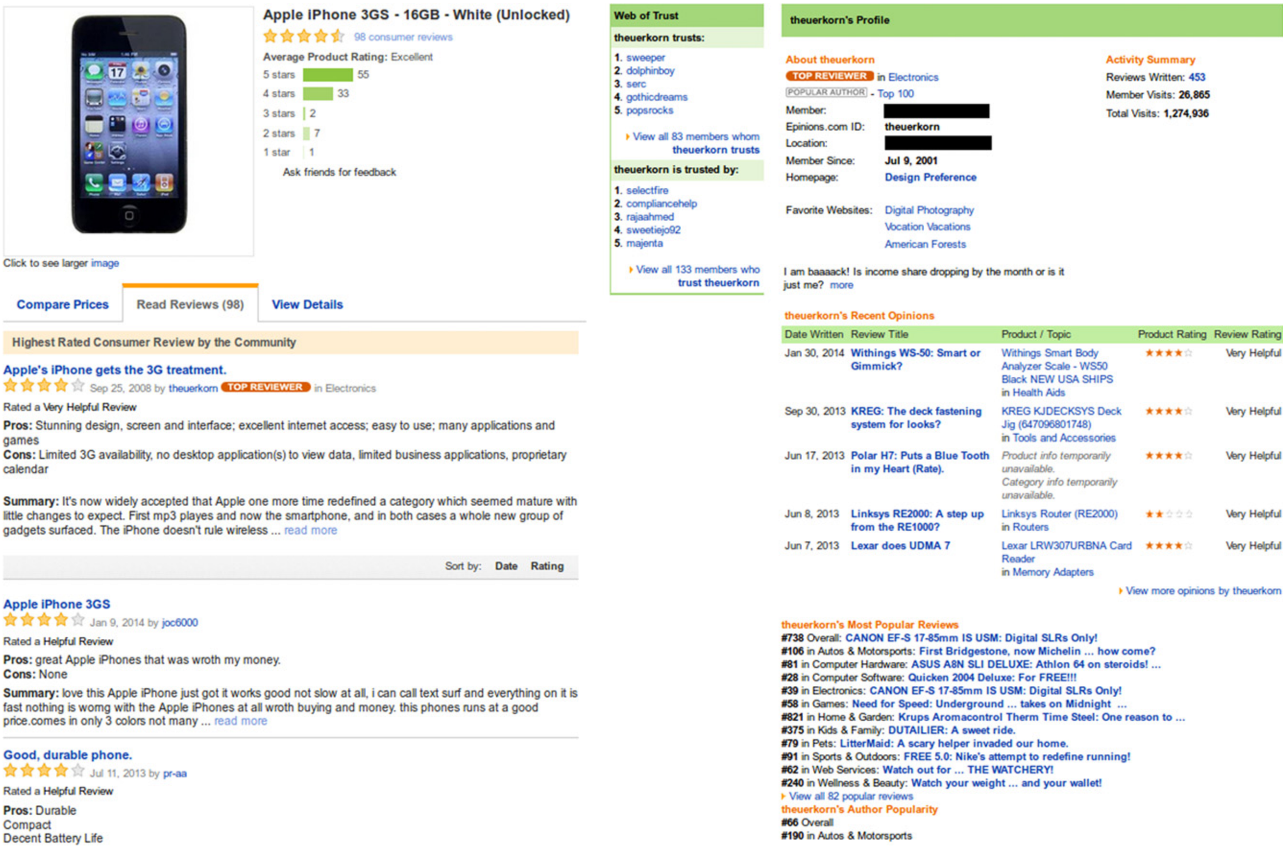
$$KLD(P(z, r = j|D_K) \parallel DU(|Z|)) = \sum_{i=1}^{|Z|} P(z_i, r = j|D_K) * \log_2 \left(\frac{P(z_i, r = j|D_K)}{DU(|Z|)} \right)$$

with $DU(|Z|)$ being a discrete uniform distribution over the topics and $P(z_i, r = j|D_K) = \frac{1}{|D_K|} \sum_{k=1}^{|D_K|} P(z_i, r = j|d_k)$.

5.3. Topic-focused ranking

Corresponding to the previous sentiment-focused ranking, we can focus the review ranking on a particular latent topic. This allows to get all opinions – positive, neutral, and negative – about a certain aspect. This might be useful for users who are interested in a particular feature of a product and the experience other users report in their reviews. Note that in this case there is no language model based ranking available.

This type of ranking can be achieved by minimizing the Kullback–Leibler divergence of the reviews in the ranking and a uniform target distribution over all ratings for one topic $\Phi = DU(|R|)$. We refrained from evaluating this strategy due to the lack



Apple iPhone 3GS - 16GB - White (Unlocked)
★★★★½ 98 consumer reviews
Average Product Rating: Excellent

5 stars	🟩	55
4 stars	🟩	33
3 stars	🟩	2
2 stars	🟩	7
1 star		1

Ask friends for feedback

Click to see larger image

Compare Prices | Read Reviews (98) | View Details

Highest Rated Consumer Review by the Community

Apple's iPhone gets the 3G treatment.
★★★★☆ Sep 25, 2008 by theuerkorn TOP REVIEWER in Electronics
Rated a Very Helpful Review
Pros: Stunning design, screen and interface; excellent internet access; easy to use; many applications and games
Cons: Limited 3G availability, no desktop application(s) to view data, limited business applications, proprietary calendar

Summary: It's now widely accepted that Apple one more time redefined a category which seemed mature with little changes to expect. First mp3 playes and now the smartphone, and in both cases a whole new group of gadgets surfaced. The iPhone doesn't rule wireless ... read more

Sort by: Date Rating

Apple iPhone 3GS
★★★★☆ Jan 9, 2014 by joc6000
Rated a Helpful Review
Pros: great Apple iPhones that was wroth my money.
Cons: None
Summary: love this Apple iPhone just got it works good not slow at all, i can call text surf and everything on it is fast nothing is womg with the Apple iPhones at all wroth buying and money. this phones runs at a good price.comes in only 3 colors not many ... read more

Good, durable phone.
★★★★☆ Jul 11, 2013 by pr-aa
Rated a Helpful Review
Pros: Durable
Compact
Decent Battery Life

Web of Trust
theuerkorn trusts:
1. sweeper
2. dolphinboy
3. serc
4. gothicdreams
5. popsrocks

View all 83 members whom theuerkorn trusts

theuerkorn is trusted by:
1. selectfire
2. compliancehelp
3. rajaahmed
4. sweetiejo92
5. majenta

View all 133 members who trust theuerkorn

theuerkorn's Profile

About theuerkorn
TOP REVIEWER in Electronics
POPULAR AUTHOR - Top 100

Member:	[redacted]
Epinions.com ID:	theuerkorn
Location:	[redacted]
Member Since:	Jul 9, 2001
Homepage:	Design Preference
Favorite Websites:	Digital Photography Vocation Vacations American Forests

Activity Summary
Reviews Written: 453
Member Visits: 26,865
Total Visits: 1,274,936

I am baaaack! Is income share dropping by the month or is it just me? more

theuerkorn's Recent Opinions

Date Written	Review Title	Product / Topic	Product Rating	Review Rating
Jan 30, 2014	Withings WS-50: Smart or Gimmick?	Withings Smart Body Analyzer Scale - WS50 Black NEW USA SHIPS in Health Aids	★★★★☆	Very Helpful
Sep 30, 2013	KREG: The deck fastening system for looks?	KREG KJDECKSYS Deck Jig (647096801748) in Tools and Accessories	★★★★☆	Very Helpful
Jun 17, 2013	Polar H7: Puts a Blue Tooth in my Heart (Rate).	*Product info temporarily unavailable.* *Category info temporarily unavailable.*	★★★★☆	Very Helpful
Jun 8, 2013	Linksys RE2000: A step up from the RE1000?	Linksys Router (RE2000) in Routers	★★☆☆☆	Very Helpful
Jun 7, 2013	Lexar does UDMA 7	Lexar LRW307URBNA Card Reader in Memory Adapters	★★★★☆	Very Helpful

View more opinions by theuerkorn

theuerkorn's Most Popular Reviews
#738 Overall: CANON EF-S 17-85mm IS USM: Digital SLRs Only!
#106 in Autos & Motorsports: First Bridgestone, now Michelin ... how come?
#81 in Computer Hardware: ASUS A8N SLI DELUXE: Athlon 64 on steroids! ...
#28 in Computer Software: Quicken 2004 Deluxe: For FREE!!!
#39 in Electronics: CANON EF-S 17-85mm IS USM: Digital SLRs Only!
#58 in Games: Need for Speed: Underground ... takes on Midnight ...
#821 in Home & Garden: Krups Aromacontrol Therm Time Steel: One reason to ...
#375 in Kids & Family: DUTAILIER: A sweet ride.
#79 in Pets: LitterMaid: A scary helper invaded our home.
#91 in Sports & Outdoors: FREE 5.0: Nike's attempt to redefine running!
#62 in Web Services: Watch out for ... THE WATCHERY!
#240 in Wellness & Beauty: Watch your weight ... and your wallet!
View all 82 popular reviews
theuerkorn's Author Popularity
#66 Overall
#190 in Autos & Motorsports

Fig. 2. Product review ranking for Apple iPhone 3GS (left) and a user profile (right) from the Epinions Web site.

of large-scale user data to test this type of ranking and the difficult mapping of all latent topics to well-defined product features.

$$KLD(P(z=i,r|D_K)\parallel P(Y)) = \sum_{j=1}^{|R|} P(z=i,r_j|D_K) * \log_2\left(\frac{P(z=i,r_j|D_K)}{P(Y)}\right)$$

with $DU(|R|)$ being a discrete uniform distribution over the ratings and $P(z=i,r_j|D_K)=\frac{1}{|D_K|}\sum_{k=1}^{|D_K|}P(z=i,r_j|d_k)$.

6. Experiments

Diversification of rankings has been mostly studied in the field of Web search, where rankings should contain novel and diverse information. Sub-topic retrieval, e.g., aims at finding an optimal ranking to cover as many sub-topics as possible (see, for example, TREC 2009 Web Track, Diversity Task in Clarke, Craswell, and Soboroff (2009)). We adopt this setting by exchanging sub-topics with aspects and sentiment. The perfect ranking would cover all different aspects and all different opinions about the aspects. The first review in the ranking should cover many aspects of the product to serve as a good overview. This evaluation approach requires annotated results, namely each review needs to be annotated with the aspects (sub-topics) discussed in it and whether the polarity of each aspect is positive, neutral, or negative. In the following we describe our dataset along with the used annotation methods.

6.1. Dataset

Product review sites, such as Amazon or Epinions, provide millions of reviews and a single product can easily have a several thousand user-generated reviews. Despite this huge amount of data, current systems do not provide an automatic ranking for these reviews. Some offer to sort the result lists by date a review was written or based on the assigned product ratings; some use community-feedback letting users decide on the "helpfulness" of individual reviews. Epinions,[5] is a general review Web site which hosts trusted opinions from customers, focusing on a diverse range of products as well as services, such as airlines or shipping companies. Epinions has also become a central place for building a collective reputation for these products and services, making it very attractive for the task of opinion filtering and analysis. Fig. 2 (left) shows a ranking for a product from the Epinions Web site. The reviews can be ranked by rating or by date. The most helpful review based on community-feedback is shown at the top of the ranking. Fig. 2 (right) depicts an epinions user profile page. The top part includes the *about* and *activity* sections, which contain basic information about each user such as id, location and length of memberships, while the activity section contains the number of written reviews and number of profile visits. On the left, the *Web of Trust* displays information of whom the user trusts *truster*, as well as the people who trust the current user *trustee*. Finally the bottom section links to the reviews written by the user.

Gathering the corpus. We collected the review data by crawling the publicly available reviews and profile pages from the Epinions Web site. The crawler was run in a semi-supervised manner over a three months period between June and August 2009. At each run between 20 and 50 threads were created and were maintained until they finished retrieving the complete social (trust) graph of each user. Due to latency in response from the server, threads were kept alive to assure completeness of the crawl. Like for most

[5] Epinions. www.epinions.com.

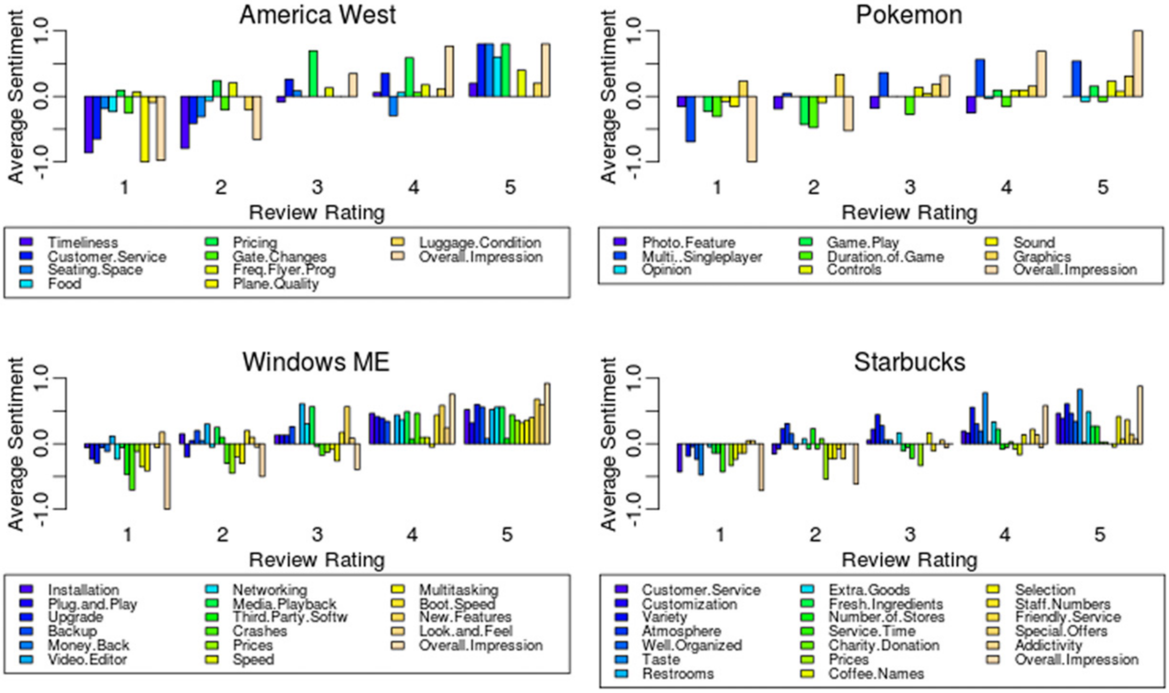


Fig. 3. Average number of positive (1.0) and negative (−1.0) mentions of aspects grouped by given rating.

Table 3
Distribution of star ratings for four test products.

Rating	Number of reviews			
	"Pokemon"	"America West"	"Starbucks"	"Windows ME"
1.0	13	43	21	17
2.0	23	29	13	20
3.0	22	23	18	23
4.0	32	17	36	41
5.0	13	5	41	25
$\sum$	103	117	129	126

web crawling tasks we used seed lists (Lauw, Lim, & Wang, 2007). We devised two seed lists: a general *community crawl list*, and a focused *negative crawl list*. The first seed list was generated based on the top reviewers list from Epinion.[6] The list assures that we have a large number of reviews along with complete profiles of users. The second seed list contains only negative reviews. This list is necessary to avoid a heavily biased dataset with only positive reviews. The complete corpus contains more than 300 products, with over 200,000 reviews.

Corpus annotation method. For our evaluation we randomly picked four out of the 300 products to manually annotated them: "America West Airline", "Pokemon Snap for Nintendo 64", "Starbucks", and "Microsoft Windows ME". This resulted in a total of 475 reviews. The manual annotation comprised the features of the products together with the polarity associated with them. Table 3 shows the distribution of ratings for these products in our corpus. To allow efficient annotation of the reviews, each product was analyzed to identify its features. To do this a fraction of the reviews based on length and expertise of reviewer were selected. After the most common features were identified, an annotation sheet was compiled for each product. It provides check boxes for the polarity of the opinions expressed in each review for a certain aspect. An example is depicted in Table 4 corresponding to the following text:

Table 4
Example of a "Starbucks" review (top) and corresponding filled out annotation sheet (bottom).

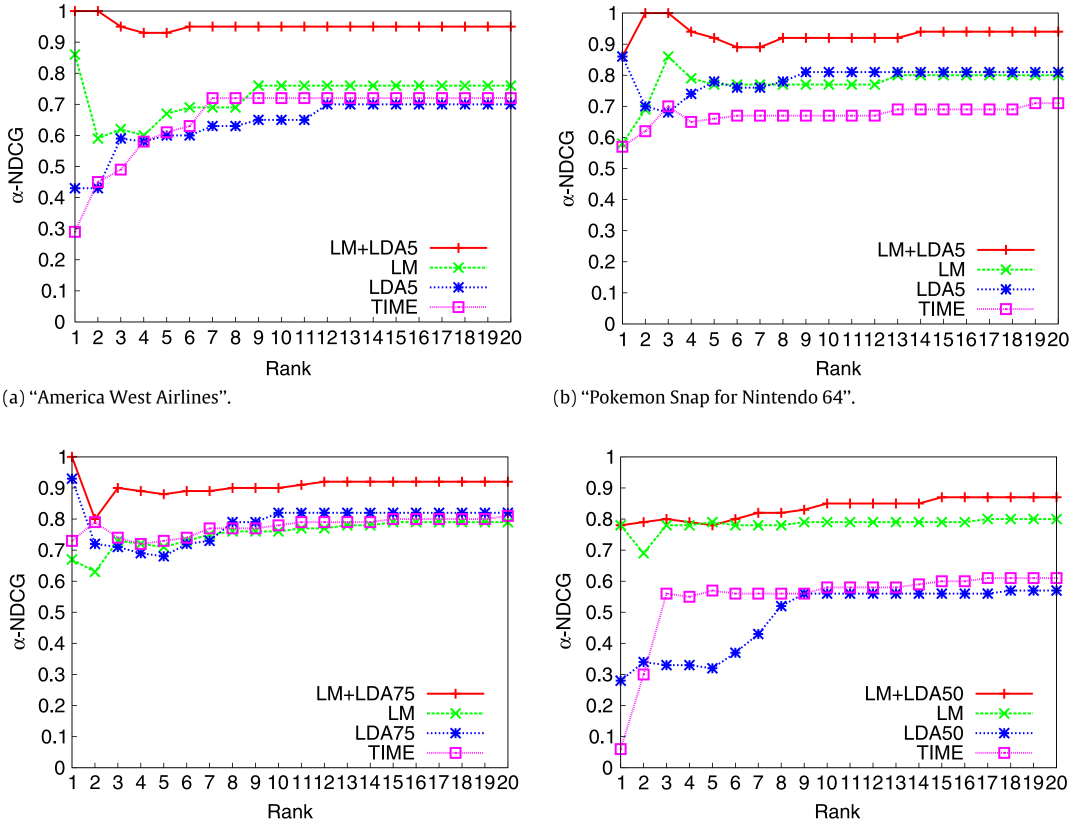
…Starbucks has excellent customer service, wonderful coffee and selection and the customer knows they are getting quality when they go to Starbucks. I recommend Starbucks to any coffee lover! It is absolutely great! So much so, I thought it would be a great career opportunity, and it is! Starbucks has between 3 and 5 people working in the front to cater to your coffee needs. Because it gets so busy sometimes that is needed. They recognize different tastes in coffee and try to accommodate, and their service is friendly and prompt. The decor of Starbucks is very trendy. It has soft lighting, nice comfy chairs for long stays and a bar that normally looks out the street window. All condiments are easy to find, and constantly full!…

Feature	Aspect	
	Pos	Neg
"Customer service"	X	
"Customization possibility"		
"Large selection"	X	
"Clean, relaxing atmosphere"	X	
"Being well-organized"		
"Taste of coffee"	X	
"Fresh ingredients"		
"Having extra merchandise"		
"Customer service time"		
"Various locations"		
"Charity donation"		
"Prices"		
"Enough staff"	X	
"Coupons/Special offers"		
"Overall opinion"	X	

6.2. Sentiment aspect validation

Fig. 3 depicts the distribution of positive and negative mentions of product aspects for the different rating classes. As to be expected, reviews with only one star report mainly negative experiences whereas five-star ratings correlate with mainly positive mentions of aspects in the review. Some features tend to be more positive across all rating classes, e.g. if "graphics" are mentioned within a Pokemon review it is mostly in a positive context, even if the associated rating was only "1". "Seating/Space" in American West

[6] http://www.epinions.com/member/community_lists.html/show_~6.



(a) “America West Airlines”.

(b) “Pokemon Snap for Nintendo 64”.

(c) “Microsoft Windows ME”.

(d) “Starbucks”.

Fig. 4. Summary strategy: comparing recency (TIME) with LDA, LM and LM + LDA.

Airlines reviews on the other hand is mostly mentioned in a negative way across rating classes. These numbers serve as an empirical validation of our hypothesis that high user assigned ratings correlate with positive mentions of aspects and low ratings correlate with negative mentions. The user assigned rating thus gives valuable information about the review text.

6.3. Results

We report results comparing the two proposed models for review data representation – topic models and language models – and a combination of both. We also compare these results with a baseline approach ranking the newest review highest. Recency-based ranking is a common way among commercial review ranking sites to rank product reviews and therefore a realistic baseline to compare against.

We evaluate the summary-focused ranking and the sentiment-focused ranking on the annotated test data. The topic-focused ranking cannot be automatically evaluated using the test data since there is no inherent mapping between latent topics and product aspects. Therefore we only have anecdotal evidence that the algorithm based on aspects works as well.

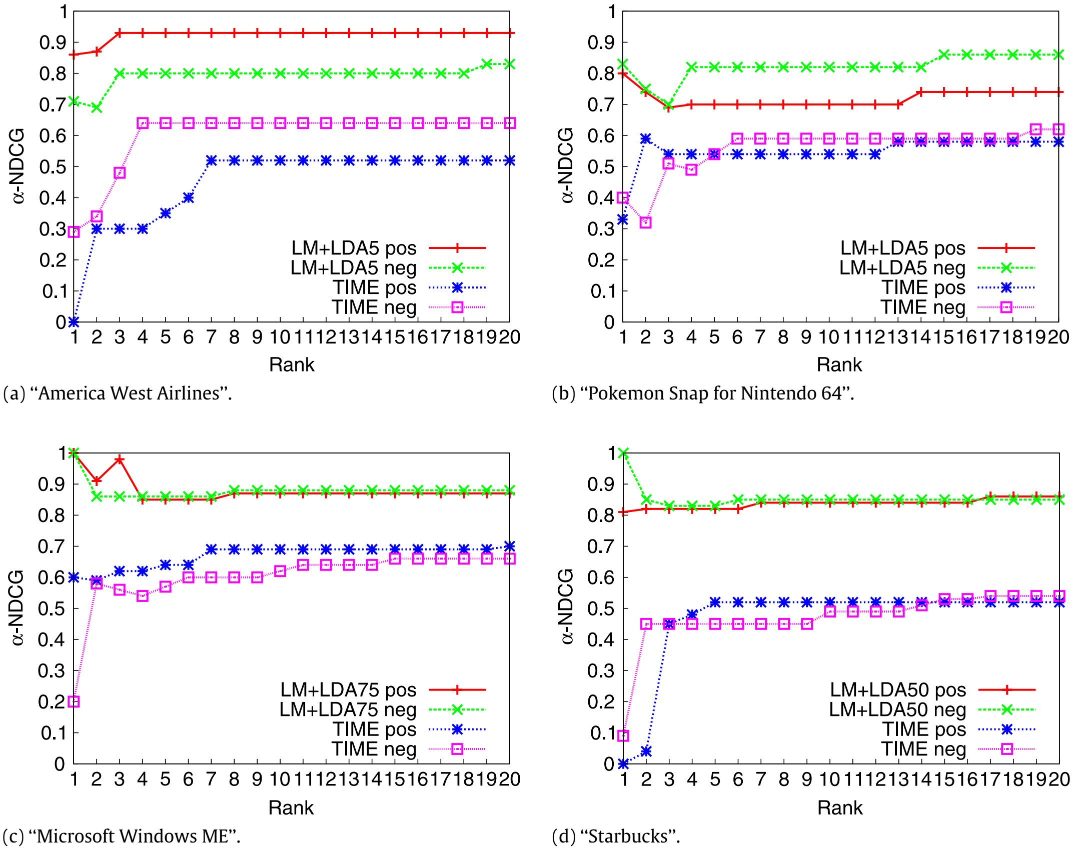
Summary-focused ranking. To evaluate the summary-based ranking we computed α-NDCG (Clarke et al., 2008) with *alpha* $= 0.99$. We assess the novelty and diversity of the rankings using the manually annotated reviews. α-NDCG only accounts for positive and negative features and does not take different degrees of polarity into account in contrast to our optimization approach.

The results for the top-20 reviews ranked based on recency (time), latent topics (LDA), language models (LM), and a combination (LM + LDA) are shown in Fig. 4. For all four products, combining LDA and LM outperforms the other approaches followed by using a language model representation. On average we outperform simple recency rankings by over 30% based on α-NDCG for the top 10 ranks. For a discussion on the optimal number of latent topics for our approach and the influence of the smoothing parameter for the user assigned crisp star ratings, see Krestel and Dokoohaki (2011).

Sentiment-focused ranking. The results for sentiment-focused ranking focusing on either positive or negative aspects are shown in Fig. 5. To compute the α-NDCG values we only considered the positive, respectively negative, manually annotated aspects to be relevant. As can be seen in the figure, summarizing the negative opinions with the top-K reviews in the ranking for “America West” is easier than the positive opinions. For “America West” the negative and positive aspects are quite well covered after the first three reviews using LDA+LM whereas the coverage of negative opinions for TIME reaches a local maximum after four reviews and for positive opinions it takes even more (eight).

6.4. Discussion

The starting point for our review rankings are the bag-of-words representations of the review texts. Therefore we do not need to extract individual aspect–sentiment pairs, as for example (Hu & Liu, 2004). Alternatives based on aspect-based opinion mining are possible and could be integrated into our framework by greedily ranking reviews based on the aspects and associated sentiments they contain. A more common way to employ information about aspect–sentiment pairs is to summarize the pairs directly and present them to the user instead of showing a ranking of original review texts. Nevertheless, using the extracted features and sentiments directly to rank reviews would work as well and allow summary-focused, as well as sentiment-focused rankings. The



(a) "America West Airlines".

(b) "Pokemon Snap for Nintendo 64".

(c) "Microsoft Windows ME".

(d) "Starbucks".

Fig. 5. Sentiment strategy: comparing recency (TIME) with LM + LDA focusing only on positive or negative aspects respectively.

topics would then correspond to the extracted features and the topic-focused ranking would be feature-focused, which could lead to slightly different results due to the agglomerating character of topics subsuming related features. Having said that, we believe that combining our bag-of-words approach with state-of-the-art aspect-based opinion mining could further improve the overall results.

7. Conclusions and future work

We presented in this paper an approach to rank reviews for products based on latent topics and language models in combination with user-assigned ratings. The main goal was to summarize the opinions expressed in all reviews for a product in the top-K results of a ranking. In contrast to recommending single reviews we aimed at recommending an optimal diverse set of reviews using methods from information retrieval. We showed that diversified rankings of reviews allow users to grasp the overall opinions about a product faster and more reliably, thus unburden the user from having to read many reviews to get an overview. We investigated reviews for products from four very different product categories, which exhibit different characteristics in terms of features and user experience. Manual annotation of the reviews allowed an automatic evaluation of the proposed approaches and a comparison of different ranking strategies and different algorithms.

For future work we will investigate the possibility of personalizing the review rankings by taking personal preferences of users into account. For example, a user might be more interested in the battery life of a product than the screen size. Another interesting direction is analyzing and categorizing product reviews on a large scale to identify different types of reviews. As trust is a factor that directly affects the user's confidence, we will also investigate the possibility to incorporate trust between users and authors of reviews.

Large scale processing in the context of big data is becoming more and more important for intelligent systems (Hussain, Cambria, Schuller, & Howard, 2014). A major challenge for future work will be to tackle real-time information processing and analysis for large volumes of streaming text data in particular for social Web content such as product reviews.

References

Aciar, S., Zhang, D., Simoff, S., & Debenham, J. (2007). Informed recommender: basing recommendations on consumer product reviews. *IEEE Intelligent Systems*, *22*, 39–47.

Beineke, P., Hastie, T., Manning, C., & Vaithyanathan, S. (2003). An exploration of sentiment summarization. In *Proc. of spring symposium on exploring attitude and affect in text: theories and applications* (pp. 1–4). New York, USA: AAAI.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bross, J., & Ehrig, H. (2010). Generating a context-aware sentiment lexicon for aspect-based product review mining. In *Proc. of the conf. on web intelligence and intelligent agent technology (WI-IAT)* (pp. 435–439). Washington, DC, USA: IEEE.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based *n*-gram models of natural language. *Computational Linguistics*, *18*, 467–479.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*, 15–21.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the conf. on research and development in information retrieval (SIGIR)* (pp. 335–336). New York, USA: ACM.

Chaovalit, P., & Zhou, L. (2005). Movie review mining: a comparison between supervised and unsupervised classification approaches. In *Proc. of the Hawaii intl. conf. on system sciences (HICSS)* (pp. 1–9). Washington DC, USA: IEEE.

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2009). Overview of the trec 2009 web track. In *(TREC): Vol. Special Publication 500-278*. *Proc. of the text retrieval conf.*. Gaithersburg, MD: NIST.

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., et al. (2008). Novelty and diversity in information retrieval evaluation. In *Proc. of the conf. on research and development in information retrieval (SIGIR)* (pp. 659–666). New York, NY, USA: ACM.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proc. of the conf. on world wide web (WWW)* (pp. 519–528). New York, NY, USA: ACM.

Deligne, S., & Bimbot, F. (1995). Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams. In *Proc. of the conf. on acoustics, speech, and signal processing* (pp. 169–172). Los Alamitos, CA, USA: IEEE.

Fellbaum, C. (1998). *WordNet—an electronic lexical database*. Cambridge, MA, USA: MIT Press.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: mining customer opinions from free text. In *Proc. of the conf. on advances in intelligent data analysis (IDA)* (pp. 121–132). Heidelberg, Germany: Springer-Verlag.

Ghose, A., & Ipeirotis, P. G. (2007). Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proc. of the conf. on electronic commerce (ICEC)* (pp. 303–310). New York, NY, USA: ACM.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proc. of the conf. on research and development in information retrieval (SIGIR)* (pp. 50–57). New York, NY, USA: ACM.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proc. of the conf. on knowledge discovery and data mining (SIGKDD)* (pp. 168–177). New York, USA: ACM.

Hussain, A., Cambria, E., Schuller, B., & Howard, N. (2014). Guest editorial introduction: affective neural networks and cognitive learning systems for big data analysis. *Neural Networks*, *58*, 1–3.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proc. of the conf. of the cognitive science society* (pp. 103–106). Erlbaum.

Kim, H., Ganesan, K., Sondhi, P., & Zhai, C. (2011). *Comprehensive review of opinion summarization. Technical report*. Illinois Environment for Access to Learning and Scholarship.

Kim, H., & Zhai, C. (2009). Generating comparative summaries of contradictory opinions in text. In *Proc. of the conf. on information and knowledge management (CIKM)* (pp. 385–394). New York, USA: ACM.

Krestel, R., & Dokoohaki, N. (2011). Diversifying product review rankings: getting the full picture. In *Proc. of the conf. on web intelligence and intelligent agent technology (WI-IAT)* (pp. 138–145). Washington, DC, USA: IEEE.

Lauw, H., Lim, E., & Wang, K. (2007). Summarizing review scores of "unequal" reviewers. In *Proc. of the conf. on data mining (SDM)* (pp. 539–544). Philadelphia, PA, USA: SIAM.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proc. of the conf. on information and knowledge management (CIKM)* (pp. 375–384). New York, NY, USA: ACM.

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of natural language processing* (pp. 1–38) Boca Raton, FL, USA: CRC Press, Taylor and Francis Group, (Chapter 28).

McCallum, A.K. (2002). MALLET: a machine learning for language toolkit. URL: http://mallet.cs.umass.edu.

Moghaddam, S., & Ester, M. (2011). ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. In *Proc. of the conf. on research and development in information (SIGIR)* (pp. 665–674). New York, USA: ACM.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*, 1–135.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proc. of the conf. on research and development in information retrieval (SIGIR)* (pp. 275–281). New York, NY, USA: ACM.

Popescu, A.-M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proc. of the conf. on human language technology and empirical methods in natural language processing (HLT-EMNLP)* (pp. 339–346). Morristown, NJ, USA: ACL.

Rajagopal, D., Olsher, D., Cambria, E., & Kwok, K. (2013). Commonsense-based topic modeling. In *Proc. of the workshop on issues of sentiment discovery and opinion mining (WISDOM)*. New York, NY, USA: ACM.

Stoyanov, V., & Cardie, C. (2008). Topic identification for fine-grained opinion analysis. In *Proc. of the conf. on computational linguistics (COLING)* (pp. 817–824). Morristown, NJ, USA: ACL.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of the conf. of the North American chapter of the ACL on human language technology (NAACL)* (pp. 173–180). Morristown, NJ, USA: ACL.

Wang, D., Zhu, S., & Li, T. (2013). SumView: a web-based engine for summarizing product reviews and customer opinions. *Expert Systems with Applications*, *40*, 27–33.

Xu, X., Meng, T., & Cheng, X. (2011). Aspect-based extractive summarization of online reviews. In *Proc. of the symposium on applied computing (SAC)* (pp. 968–975). New York, USA: ACM.

Yu, J., Zha, Z., Wang, M., & Chua, T. (2011). Aspect ranking: identifying important product aspects from online consumer reviews. In *Proc. of the meeting of the association for computational linguistics ACL* (pp. 1496–1505). Morristown, NJ, USA: ACL.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. of the conf. on research and development in information retrieval SIGIR* (pp. 334–342). New York, NY, USA: ACM.

Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. In *Proc. of the conf. on information and knowledge management (CIKM)* (pp. 43–50). New York, USA: ACM.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proc. of the conf. on world wide web (WWW)* (pp. 22–32). New York, USA: ACM.