Feature Sentiment Diversification of User Generated Reviews: The FREuD Approach

Nasir Naveed Thomas Gottron Steffen Staab
WeST – Institute for Web Science and Technologies
University of Koblenz-Landau
Universitätsstr. 1, 56070 Koblenz
{naveed, gottron, staab}@uni-koblenz.de

Abstract

Online discussions, user reviews and comments on the Social Web are valuable sources of information about products, services, or shared contents. The rapidly growing popularity and activity of Web communities raises novel questions of appropriate aggregation and diversification of such social contents. In many cases, users are interested in gaining an extensive overview over pros and cons of a particular track of contributions. We address the problem of social content diversification by combining latent semantic analysis with feature-centric sentiment analysis. Our FREuD approach provides a representative overview of sub-topics and aspects of discussions, characteristic user sentiments under different aspects, and reasons expressed by different opponents. In experiments with real world product reviews we compare FREuD to the typical implementation of ranking reviews by the usefulness rating provided by users as well as a naive sentiment diversification algorithms based on star ratings. To this end we had human users provide a fine-grained gold standard about the coverage of features and sentiments in reviews for several products in three categories. We observed that FREuD clearly outperforms the baseline algorithms in generating a sentiment-diversified set of user reviews for a given product.

1 Introduction

Web 2.0 provides an interactive way for online text publishing in different domains. Users can engage in online discussion on a wide range of topics and contribute their personal experiences and opinions. One such area is online product review portals such as reviews.cnet.com and epinion.com. These portals do not only publish editorial reviews of different products but also provide ways for the users to share their own experience of the use of the products. In these reviews users tend to cover different aspects or features of the products. Usually a review covers some features of the product along with associated sentiments or opinions about these features. Other than being positive or negative about features, users also discuss which features are more important than others and about which features they are more excited. Thus, these reviews provide rich information about different aspects of a product and can play an important role in the decision making process of new customers when buying a product.

Consider, for instance, a scenario where a customer intends to buy a smartphone. Nowadays, his first step would likely be to use a product review website and browse through user reviews for different smartphones to get an overview of the users' opinions about the usefulness of different features of each of the phones. A common observation is that successful products can have hundreds of reviews, where some reviews are more useful than others. Another observation is that the reviews are written in free text form and the extent of coverage is different towards different features with different degrees of authority or authenticity. To arrive at some decision of whether or not to buy the phone the customer has to contemplate a large amount of text to find the most valuable reviews or opinions; which requires a lot of reading and time. Thus, the challenge in this scenario is to come up with an optimal set of high quality reviews that cover as many relevant features as possible and provide diversified view points of opinions of different users about the product's features.

Currently, there are some ways to indicate which reviews are worth reading. One common way is to leverage user votes for reviews which indicate the usefulness of a review as seen by other users. But user votes do not guarantee that highly rated reviews cover all possible aspects and associated sentiments and that all the pros and cons are addressed. In fact, the collective dynamics may even lead to disproportionately high ratings of some, rather arbitrary, reviews (cf. the analogy with preferred downloads in Salganik and Watts (2009)).

Automatic solutions could overcome these problems, but mining sentiments about product aspects or features from user reviews poses certain challenges. The first challenge is to estimate which features are addressed in a review. The second difficulty is to mine the users' sentiments. And the third challenge is how to come up with an optimal set of reviews, which has already been shown in literature to be an NP-hard problem (Tsaparas, Ntoulas, and Terzi 2011).

To tackle the above mentioned challenges, we consider the review selection problem as an information retrieval task with specific emphasis on result diversification. Based on this mind-set we developed the FREuD approach[1]. In FREuD we use a combination of text pre-processing and probabilistic topic models to obtain latent topics discussed in a collection of reviews related to a single product. We observed, that these latent topics frequently align very well with the features of the product discussed in the reviews. Thus, the latent topics provide us with a very good approximation of the product features. We then employ a dictionary based approached for estimating the user sentiments in each single review. Finally, we select a subset of reviews to optimize the diversity criteria of covered product features and sentiments. To this end, we use a greedy algorithm operating on the features and sentiments discussed in each review.

In this paper we make two main contributions:

- We describe the problem setting of feature-centric sentiment diversification as an information retrieval task. We present the details of our FREuD approach and discuss its technical details. We demonstrate that FREuD outperforms two baselines of a user based ranking as it is currently implemented in productive systems as well as a naive sentiment diversification strategy.
- We develop of a novel gold standard data set for the task of feature-centric sentiment diversification over product reviews. To this end, we have had human assessors annotate a gold standard on the features covered and sentiments expressed about these features in reviews for twenty products in three different categories[2].

The rest of the paper is organized as follows. We will review related work in the next Section. Afterwards, we proceed with a formalization of the task of feature-centric sentiment diversification and its interpretation from an information retrieval viewpoint. Our approach is discussed in Section 4. The evaluation methodology, the employed data set and the construction of a gold standard is presented in Section 5, the results and comparison of FREuD with baseline methods in Section 6. We conclude the paper with a summary of our contributions and an outlook at future work.

2 Related Work

In this contribution we specifically focus on mining product features, estimating sentiments from free text in an unsupervised way and using this information for product review diversification. Therefore, our work mainly relates to two areas of research: text mining and diversity ranking. So, in this section we concentrate on related work in these areas.

Feature extraction techniques mainly rely on the availability of structured or semi-structured documents. Guo et al. (2009), for instance, proposed an unsupervised product-feature extraction and categorization method from semi-structured reviews. Their method relies on extracting features mentioned explicitly in structurally indicated pros and cons sections in a review. Liu, Hu, and Cheng (2005) proposed a supervised method for detecting product features in semi-structured reviews. They used associative rules and manually labelled data for this purpose. Similarly, Shi and MingYu (2011) studied a theoretical framework based on product feature mining issues from customer reviews and proposed a DFM (Data, Function, Mining) model for mining product feature structures from such reviews. In another work Zhai et al. (2011) proposed a semi-supervised method for clustering product features for opinion mining. They used a semi-supervised approach for grouping synonym features.

In contrast we use an approach based on general domain text pre-processing and Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), which neither relies on the structure of the document nor requires any manually labelled training data for mining latent topics. Therefore, it is applicable to unstructured text and is generalizable to any text collection.

Opinion mining or sentiment classification is a widely studied field. Some of the previous approaches focused on sentiment-based classification of individual words, phrases, sentences or documents as whole and assume sentiment classification as a binary task (positive or negative) (Pang, Lee, and Vaithyanathan 2002; Pang and Lee 2005). Eirinaki, Pisal, and Singh (2011) presented an algorithm for analyzing the overall sentiment of reviews and also identified semantic orientation of the specific component of the review that leads to a specific sentiment. Qiu et al. (2009) come up with a self-supervised model for sentiment classification. They used a dictionary based approached for sentiment classification. Lin et al. (2012) present a weakly supervised sentiment classification approach which is directly incorporated into a topic analysis based on LDA. In another work, Ganesan, Zhai, and Viegas (2012) proposed an unsupervised approach for generating ultra-concise aspect related summaries of opinions.

The above mentioned approaches do classify a document as a whole or parts of it as positive or negative and identify the polarity of a text snippet in relation to some aspect. These approaches can be used to classify a document as positive or negative on the basis of positive or negative phrase counts but they do not provide the strength of the sentiment in some numeric form. In another work (Pang and Lee 2005) it is shown that sentiment classification can be generalized into a rating scale. Intuition is one can get better diversification of reviews using a rating scale rather than using binary classification. For sentiment analysis we apply a dictionary based approach (Bradley and Lang 1999) which has already been applied successfully in other social web scenarios (Nielsen 2011; Naveed et al. 2011). This approach not only helps in sentiment classification of a document but also provides sentiment scores which can be used to reflect the overall strength of the sentiment.

Result diversification has recently received a lot of attention in the Information Retrieval community. A good overview of the general task of search result diversification and its evaluation in particular is presented by Carterette (2011). The selection of product reviews as a diversification task is addressed by Tsaparas, Ntoulas, and

[1] An early draft of the idea behind the approach has already been presented in the form of an extended abstract, cf. Naveed et al. (2012).

[2] The gold standard data set developed and used in this paper is available for download and research use at: http://west.uni-koblenz.de/Research/DataSets/FREuD

Terzi (2011). The focus, however, is only on a good coverage of product features. In recent work (Krestel and Dokoohaki 2011) similar to our approach, LDA was used for detecting latent topics in the reviews and star ratings of the reviews as an indicator of sentiment polarity to be used in review diversification. This method relies on star ratings for determining overall sentiment of the review.

However, in our approach we do not need to rely on star ratings for determining sentiment polarity. Furthermore, none of the approaches has addressed the task of diversification of both: feature and sentiment coverage in selected documents.

3 Formal Task Definition

Before going in the details of our approach, we formalize the feature-centric sentiment diversification task we are addressing. The aim of this FSCOVERAGE(k) task is to generate a selection of k product reviews that cover as many product relevant features as possible and an as good as possible diversified range of sentiments over the features.

Let us consider a product $\mathcal{P}$. The set of reviews related and relevant to this product $\mathcal{P}$ forms a corpus $\mathcal{C}$. This corpus $\mathcal{C}$ constitutes the set of documents FSCOVERAGE(k) operates on. Furthermore, we consider a product $\mathcal{P}$ (e.g. a mobile phone) to be associated with a finite set $\mathcal{F} := \{f_1, f_2, f_3, \ldots, f_n\}$ of product relevant features (e.g. screen size, battery life time, usability, etc.). Finally, we define a set of sentiment dimensions $\mathcal{S} := \{s_1, s_2, s_3, \ldots, s_m\}$ (e.g. positive or negative valence, calm or excited arousal, etc.). In the FSCOVERAGE(k) task, the product features $\mathcal{F}$ and the sentiment dimensions $\mathcal{S}$ define the space we want to cover as extensive as possible with a fixed number of reviews.

We can assume the reviews in $\mathcal{C}$ to address the features and utter sentiments about them in various degrees. It is possible that a review expresses both, positive and negative, sentiments about a certain feature, e.g. in a statement like *"the screen of the phone is wonderfully large but its readability in sunlight is really bad"*. In review d we capture the strength of a positive sentiment $s \in S$ regarding feature $f \in \mathcal{F}$ with the value $v^+(f,s,d) \in [0,1]$, and the negative sentiment with $v^-(f,s,d) \in [0,1]$. Higher values of $v^+(f,s,d)$ and $v^-(f,s,d)$ correspond to stronger positive and negative sentiments, respectively. A value of 0, instead, means that no positive or negative sentiment is uttered about feature f.

For a given set $\mathcal{C}' \subset \mathcal{C}$ of reviews, we can now define a feature-sentiment-diversity score *Div*$(\mathcal{C}')$ by:

$$\mathit{Div}(\mathcal{C}') = \sum_{f \in \mathcal{F}} \sum_{s \in S} \left(\max_{d \in \mathcal{C}'} v^+(f,s,d) + \max_{d \in \mathcal{C}'} v^-(f,s,d) \right)$$

The FSCOVERAGE(k) task is to maximize the score $D(\mathcal{C}')$ under the constraint $|\mathcal{C}'| \leq k$. In analogy to the proof by Tsaparas, Ntoulas, and Terzi (2011) it can be shown that FSCOVERAGE(k) is NP-hard.

4 The FREuD Approach

We have already stated above the three main challenges for obtaining a feature-centric sentiment diversified selection of reviews. Using the formalization in the previous Section we can state these challenges more precisely:

Table 1: LDA topics for the category Camera

Topic No.	Top Topic Terms
1	lens, mode, iso, fps, value
2	camera, image, quality, picture, dslr
3	camera, video, quality, lens, slot
4	control, sensor, pixel, zoom, issue
5	photo, control, flash, option, set
6	zoom, panason, grip, focus, inch
7	camera, feature, body, kit, frame
8	nikon, focus, issue, lcd, meter
9	camera, water, shock, fog, claim
10	perform, review, viewfinder, comparison

- Identify the set of features $\mathcal{F}$ discussed in a set of reviews $\mathcal{C}$.
- Estimate the positive and negative sentiment values $v^+(f,s,d)$ and $v^-(f,s,d)$ for a given feature f in a specific document $d \in \mathcal{C}$.
- Provide a good approximative solution for FSCOVERAGE(k) based on these estimates.

We have developed the FREuD approach which combines machine learning techniques for product feature mining, a dictionary based approach for estimating the sentiments of a review and a greedy approach to provide a solution for FSCOVERAGE(k). We will now present the details for each of these steps in the following subsections.

4.1 Feature Extraction

To obtain a list of features discussed in a set of reviews, we use Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). LDA is extensively used for mining latent topics in text corpora. Each topic corresponds to a probability distribution over words. The motivation is, that each topic favours the use of specific words. Furthermore, in LDA a text is modelled as a mixture of latent topics Z. This means each text document can address various topics to a different degree. For a given text corpus, machine learning techniques for LDA provide the word distributions of the latent topics as well as the probabilistic topic composition for each document. These topic compositions come in the form of the probabilities $P(z|d)$ that a document d covers topic $z \in Z$.

Using LDA on product reviews we observed that the latent topics frequently align very well with product features discussed in the reviews. To refine the input for LDA, we use the Stanford Part-Of-Speech Tagger (Toutanova et al. 2003) to extract nouns from the reviews. Most of these nouns reflect the different aspects of products. In LDA, for each product category we set the number of topics to be 10. This number approximates the number of features we use for each category in the evaluation data set. Table 1 shows top terms of topics discovered by LDA which approximate to features discussed in reviews about cameras.

Thus, the latent topics Z provide us with a very good approximation of the set of discussed product features $\mathcal{F}$. Accordingly we can consider each topic $z \in Z$ to correspond to a feature $f \in \mathcal{F}$. Furthermore, the topic composition of each review gives us an estimate to which degree a review discusses a specific feature. In conclusion, we use the probability $P(f|d)$ as value for modelling the extent to which review d addresses feature f.

4.2 Sentiment Estimation

To estimate the sentiments in a review we employ the Affective Norms for English Words (ANEW) sentiment dictionary (Bradley and Lang 1999). ANEW assigns normative emotional ratings to a large number of English words under three categories of sentiments: valence, arousal and dominance. Valence scores reflect the polarity of the review ranging from negative to positive, arousal reflects the excitement level ranging from highly excited to calm and dominance reflects how certain the user is in expressing the sentiments. Thereby, these three categories provide us with a set $\mathcal{S}$ of sentiments to work on. The emotional rating values for words in ANEW cover a range between 1 and 10. We normalize these values to the interval $[-1, 1]$ and distinguish between the positive and negative values for the purpose of obtaining $v^+(s, w)$ and $v^-(s, w)$ for each individual word w and sentiment s. The global positive sentiment value $v^+(s, d)$ of an entire review d is then given by an aggregation of the positive sentiment values of the single words:

$$v^+(s, d) := \sum_{w \in d} v^+(s, w)$$

The value for $v^-(s, d)$ is defined equivalently.

4.3 Feature-Sentiment Estimation

To estimate the positive and negative sentiment s for a given feature f under each sentiment category in a review we combine the positive and negative global sentiment of a review d with the probability of the review to address feature f according to outcome of the LDA analysis:

$$\begin{aligned} v^+(f, s, d) &:= v^+(s, d) \cdot P(f|d) \\ v^-(f, s, d) &:= v^-(s, d) \cdot P(f|d) \end{aligned}$$

4.4 Review Subset Selection

As already mentioned FSCOVERAGE(k) is NP-hard. Therefore, to find a good solution for FSCOVERAGE(k) we use a greedy algorithm. The greedy algorithm starts from an empty result set $\mathcal{R}$ of selected reviews and iteratively adds a review that adds most value to the result set in the sense, that it extends the range of covered features and expressed sentiments most.

The degree to which a combination of sentiment s and feature f is already covered in the result set $\mathcal{R}$ is given by $\max_{d' \in \mathcal{R}} v^+(f, s, d')$ and $\max_{d' \in \mathcal{R}} v^-(f, s, d')$ for the positive and negative sentiment values respectively. The gain of adding document d to the result set corresponds to the subsequent increase of these two maxima. By summing up the increase over all combinations of sentiments and features we obtain a contribution score *contrib*(d) for document d:

$$\begin{aligned} &\mathit{contrib}(d) := \\ &\sum_{s \in \mathcal{S}} \sum_{f \in F} \left[\max\left(0, \left(v^+(f, s, d) - \max_{d' \in \mathcal{R}} v^+(f, s, d')\right)\right) \right. \\ &\quad \left. + \max\left(0, \left(v^-(f, s, d) - \max_{d' \in \mathcal{R}} v^-(f, s, d')\right)\right)\right] \end{aligned}$$

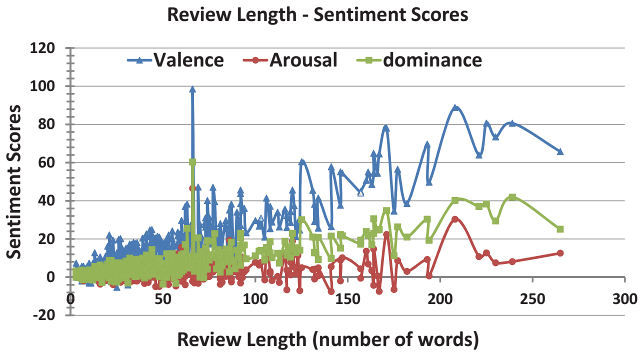

Figure 1: Relationship of review length and sentiment scores.

After having computed this score for all documents which have not been added to the result set so far, we select the one review d with the highest *contrib*(d) score for addition to the result set $\mathcal{R}$. In the next iteration we recalculate the *contrib*(d) scores of the remaining reviews to determine which review to add next. This iteration is repeated until the set $\mathcal{R}$ contains k reviews.

4.5 FREuD Variations

Looking at real world data, we observed a near linear relationship between review length and sentiment scores of the review. However, we also observed that there is a lot of variance in the strength of the sentiment expressed in the reviews. This variance cannot be explained by the document length alone. Figure 1 shows this relationship for valence, arousal and dominance.

This relationship between review length and sentiment scores has an effect on the *contrib*(d) function, which would favour longer documents. Therefore, to check the effect this bias has on the performance of our approach, we implemented FREuD in three different variations dealing with length normalization in different ways:

- **FREuD-noLN.** This variation does not make use of any length normalization technique for sentiment scores.
- **FREuD-stdLN.** In this implementation we use a standard length normalization for sentiment scores. We divide the global sentiment score of a review by the total number of words in the review.

Table 2: CNET product review data set used in our experiments.

Category	# Products	# Reviews
Cell Phone	7	1501
Printer	7	688
Camera	6	256

- **FREuD-sentiLN.** Length normalization is performed based on the number of sentiment words in the review, i.e. the number of words which have actually been annotated with a sentiment score according to the ANEW dictionary.

5 Evaluation Setup

In this section, we elaborate on the evaluation methodology for our FREuD approach. Our evaluation approach includes an objective evaluation using established metrics for measuring the performance of a search result diversification systems. We describe the compilation of a data set and gold standard suitable for evaluating approaches on feature-centric sentiment diversified selection of reviews. In this context we also introduce the two baseline systems which we use for comparison.

5.1 Data Set

As evaluation corpus we use end user product reviews collected from the CNET product review website[3]. CNET covers several product categories of consumer electronics. This website allows end users to write reviews about products and provide an overall rating of the products using a five-star rating system. The users also have the option to vote for the usefulness of existing reviews using a thumbs up and thumbs down voting system. By default, CNET uses these votes to rank the reviews in the user interface from the most helpful to the least helpful review.

We used the API of CNET[4] for obtaining and downloading product information and reviews about popular products under three categories: printers, cell phones and digital cameras. In each category we chose up to seven products with at least 40 reviews. We then crawled all the user reviews from these products along with their metadata, e.g. the star rating and the number of thumbs up and thumbs down votes for each review. In total we obtained 2,445 product reviews on 20 products for our evaluation data set. Except for the category of cameras each product had more than 50 reviews. Table 2 gives an overview of the corpus.

5.2 Baseline Systems

In order to judge the quality of our FREuD approach we need realistic baseline systems to compare to. As mentioned above, CNET by default ranks the reviews for each product on the basis of the helpfulness of the review. The helpfulness is computed on the basis of the thumbs up and thumbs down votes for the reviews. Such an approach is also implemented in many other product review portals. We reconstruct this ranking from the metadata of the reviews and use this approach as our first baseline system: **CNET-default**. Furthermore, as CNET displays five reviews per page and as in a Web context only few user go beyond the first page (Jansen, Spink, and Saracevic 2000), we used a size of $k = 5$ for the set of reviews for all approaches. This baseline allows for comparing to a realistic scenario implemented in productive, real world systems.

A second baseline implements a naive sentiment diversification strategy. As mentioned above, each product review in CNET provides also a star rating on a scale between 1 and 5 stars. If a user assigns 5 stars to the product, this implies he is highly positive about the product while 1 star means he is highly unsatisfied with it. For our **CNET-diversified** baseline, we picked one review from each of the five star rating categories. As there are typically multiple reviews with the same star rating, we always chose the one with the highest usefulness score according to the thumbs up votes.

5.3 Developing a Gold Standard for the Data Set

We are interested in diversity based on both: sentiments and features. As the data set does not directly exhibit objective and machine readable information about the covered features or expressed sentiments we needed to obtain a gold standard in a different way. To this end we first collected a list of product features for each of the categories and then employed crowdsourcing in order to obtain human feedback on whether or not a feature is discussed in a review and what are the sentiments about this feature.

The preliminary requirement for our evaluation setup was to obtain a list of typical product features for each product category. For cell phones, we obtained a list of features from *gsmarena*[5], which uses a predefined structured list of features for cell phone comparisons. For digital cameras we used an equivalent list of features employed on *dpreview*[6] for reviewing cameras. Finally, we used three printer websites to collect the most commonly discussed features for the printer category.

In the next step we had all five approaches (CNET-default, CNET-diversified, FREuD-noLN, FREuD-stdLN and FREuD-sentiLN) compute a set of top-5 reviews for each product. We pooled the reviews obtained in this way and had them evaluated by human assessors in a crowdsourcing fashion. The assessors were then asked to mark which of the features from the given lists were covered in a review and which sentiments were expressed about these features. The characteristics of the data set used for this evaluation is shown in Table 3. The reviews for each product in each category were mixed and anonymized for systems names. The assessors were not aware of the originating system of the review.

The human assessors were iteratively asked to pick a product category and select a product in that category for which they wanted to read a review and provide details on

[3] http://reviews.cnet.com/

[4] http://developer.cnet.com/

[5] http://www.gsmarena.com/

[6] http://www.dpreview.com

Table 3: Data set used for objective evaluation

Category	# Products	# Reviews	# Features
Cell Phone	7	175	13
Printer	7	175	12
Camera	6	150	11

which features are discussed in the review and which sentiments are expressed about these features. The sentiment choices available to assessors for selection were ‘positive’, ‘negative’, ‘neutral’ and ‘both’. The option ‘both’ meant that a reviewer is positive and negative about a given feature. Assessors could evaluate any number of reviews. An identification of the assessors avoided that the same assessor could work on the same review twice. Assessors’ prior use and knowledge of the products was also recorded. To collect the assessors feedback on the review, we used the process as described by Algorithm 1.

Algorithm 1: Process used to obtain assessors feedback while developing gold standard.

```
assessor selects a product;
for each unassessed remaining review do
    randomly pick one review at a time and present it to
    the assessor side by side with the product specific
    preselected features;
    assessor reads review and ;
    for each feature (from the list): assessor checks do
        if feature is discussed in the review at all then
            for all found utterances discussing this
            feature do
                assessor ticks the appropriate option
                (positive, neutral, negative, both) to
                annotate the sentiment polarity of the
                feature;
                mark the location of the utterance in the
                review;
            end
        end
    end
end
```

There were 179 unique assessors who voluntarily participated in the evaluation[7]. Each of the review in the evaluation was presented to three different assessors. A feature is deemed as covered in the review if two out of three assessors agreed that the given feature was discussed. For the sentiment polarity of the feature we employed a similar majority decision. Table 4 shows some details of the participating assessors.

Table 4: Statistics about the assessors who participated in the evaluation

Gender	Percentage
Males	68.72%
Females	29.05%
Undisclosed	2.23%

Product Knowledge	Percentage
No	51.79%
Little	29.91%
Yes	18.30%

Table 5: Category-wise inter-rater agreement over coverage of the feature in the review.

Category	Fleiss Kappa (k)
Printer	0.45
Cell Phone	0.44
Digital Camera	0.41

5.4 Inter-rater Agreement

To gain confidence in our gold standard we checked the inter-rater agreement among the evaluators over the feature and sentiment coverage they identified. For this purpose we used Fleiss Kappa (Fleiss and others 1971). Fleiss Kappa is a widely used inter-rater reliability measure employed to check for nominal scale agreement between a fixed number of raters. The product category wise inter-rater agreement is show in Table 5, which according to the Fleiss benchmark (Fleiss 1981) is an *intermediate to good agreement* for all the categories.

5.5 Diversity Evaluation Metrics

Measuring the performance of algorithms which combine relevance and diversity together requires metrics which can incorporate relevance and diversity in a ranked retrieval evaluation setup. One established metric is $\alpha-$nDCG (Clarke et al. 2008) which builds on standard nDCG. The assumption underlying $\alpha-$nDCG is that each query has multiple known intents or facets and these intents are of equal importance. The $\alpha-$nDCG metric regards the documents in a result set to cover these query intents to different degrees. A highly relevant document is one which covers many intents. Additionally, $\alpha-$nDCG promotes an increase in diversity by reducing redundancy.

The main difference between standard nDCG and $\alpha-$nDCG lies in the definition of the gain values. For $\alpha-$nDCG, the gain $G\left[k\right]$ of the document at rank k is a vector over all query intents. Furthermore, this gain is discounted for intents which have already been covered by higher ranked documents. Thus, for $\alpha-$nDCG, the gain $G\left[k\right]$ at rank k is defined as:

[7] A large share of the evaluation was completed by research fellows from various research group who were typically interested in developing such a gold standard data set to be used later in other experiments.

$$G\left[k\right] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}}$$

where $J(d_k, i)$ is a binary value describing if the document at rank k is relevant to the query intent i according to the gold standard and $r_{i,k-1}$ denotes how many higher ranked documents have already addressed the intent i. The parameter α is used to balance redundancy and novelty. Its value ranges between 0 and 1, where lower values of α increase redundancy and decrease novelty and higher values of α favour novelty at the cost of reduced redundancy.

Based on these gain values the discounted cumulative gain $DCG\left[k\right]$ at rank k is given by:

$$DCG\left[k\right] = \sum_{j=1}^{k} G\left[j\right] / (log_2(1+j))$$

Defining DCG' as the ideal gain obtained by sorting the documents in the ideal order according to the gold standard allows for a normalization equivalently to the one of standard nDCG. This leads to the definition of $\alpha - nDCG\left[k\right]$ as:

$$\alpha - nDCG\left[k\right] = \frac{DCG\left[k\right]}{DCG'\left[k\right]}$$

In our setting we can assume each product to serve as query and product features as different known intents of the query with each intent (feature) having equal likelihood or importance. For a given product we consider each feature-sentiment pair as one intent and if a review covers more such intents, it should be ranked higher than the others. In our settings we used the standard value of α used also in the TREC diversity task, i.e. $\alpha = 0.5$.

For computing the diversification performance of the different approaches we used the TREC evaluation framework provided for the diversity task of the Web Track[8]. We generated appropriate input files (*qrels*, *topics* and *results*) for the TREC tool from our gold standard data set and the result files from the rankings provided by all the competing approaches. Given our setting and the motivation described above we cut off the result list for all approaches after five results. Accordingly we compare the performance based on $\alpha-$nDCG@5.

6 Experimental Results

In this section we present and discuss the experimental results for the evaluation of the variations of FREuD and the baseline approaches in selecting the sentiment diversified top-5 reviews. As mentioned in the Section 5.5, we measure the performance based on the α-nDCG@5 metric. Table 6 compares the α-nDCG@5 scores for all approaches and for each individual product in the three categories. We have highlighted the best performing approach for each product and for each category. FREuD-noLN and FREuD-sentiLN achieve high scores in general and dominate the baseline approaches for most products. FREuD-stdLN still provides very good results in some cases, but the values are less stable and exhibit a larger variance.

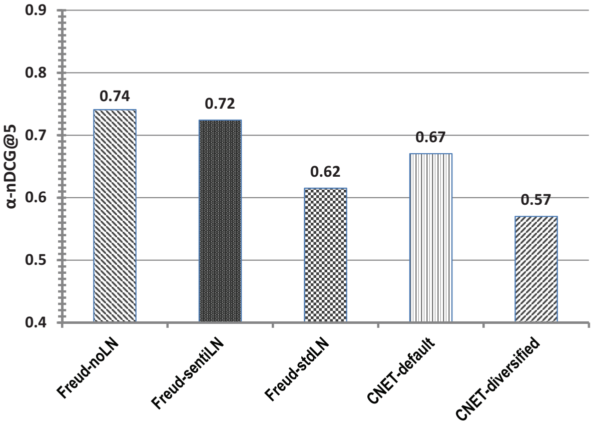

Figure 2: Feature-sentiment diversification performance of all approaches aggregated for each category of products.

Table 7: Relative percentage improvement in α-nDCG scores achieved by FREuD variations against the baseline systems.

FREuD Variations	CNET-default	CNET-diversified
Freud-noLN	10.52%	29.99%
Freud-sentiLN	8.02%	27.05%
Freud-stdLN	-8.24%	7.92%

This behaviour is also reflected when considering the average performance of the approaches. Figure 2 shows the overall average α-nDCG values. Here we observe that FREuD-noLN dominates all other systems including the two baseline systems as well as the other two variations of FREuD. However, the values of 0.74 and 0.72 for FREuD-noLN and FREuD-sentiLN, respectively, are very close to each other. For FREuD-stdLN, instead, we see that the average performance is actually below the CNET-default baseline. The naive sentiment-diversification of CNET-diversified performs worst. The poor performance of FREuD-stdLN can be explained by the fact that standard length normalization favors the short length reviews to be ranked higher as their length normalized sentiment scores are higher than the sentiment scores of longer reviews. As shorter reviews typically cover a lower number of features, therefore, the collective feature coverage of the reviews recommend by FREuD-stdLN is less than the other two FREuD approaches.

Table 7 illustrates the relative improvement in α-nDCG@5 scores achieved by FREuD variations over the two baselines. Also in this case we see a noticeable gain in performance by FREuD-noLN and FREuD-sentiLN.

To analyse differences in the different product categories, we computed the performance of all approaches at product category level. Figure 3 shows the category wise aver-

[8] http://plg.uwaterloo.ca/~trecweb/2009.html

Table 6: Performance comparison of all the approaches over the complete set of products in all categories using α-nDCG@5

Category	Product-ID	FREuD-noLN	FREuD-sentiLN	FREuD-stdLN	CNET-default	CNET-diversified
Camera	1	0.7993	**0.8163**	0.3878	0.7173	0.3064
	2	**0.8355**	0.8122	0.6339	0.6827	0.7157
	3	**0.7596**	0.6576	0.6576	0.6332	0.4740
	4	0.7383	0.8674	**0.8735**	0.8562	0.7673
	5	**0.7974**	0.7281	0.7917	0.6698	0.7714
	6	0.8622	0.8539	0.7666	**0.9291**	0.7127
	7	0.8045	**0.8509**	0.7725	0.7308	0.8301
	Avg.	**0.7996**	0.7981	0.6977	0.7456	0.6539
Cell Phone	8	**0.7485**	0.6251	0.7006	0.7206	0.3677
	9	0.5719	0.4747	**0.7365**	0.4988	0.6746
	10	**0.7713**	0.4823	0.5746	0.7376	0.3146
	11	0.6373	**0.7090**	0.2895	0.6501	0.3098
	12	0.6406	**0.9048**	0.7827	0.8835	0.8112
	13	0.8224	0.5991	**0.9173**	0.4369	0.6124
	14	0.8443	**0.9071**	0.4240	0.8371	0.3966
	Avg.	**0.7195**	0.6717	0.6322	0.6807	0.4981
Printer	15	**0.6521**	0.5747	0.6338	0.4701	0.4879
	16	0.7153	**0.7591**	0.5850	0.5974	0.4487
	17	**0.7660**	0.7315	0.2154	0.6059	0.5524
	18	**0.7652**	0.4254	0.4333	0.4715	0.7017
	19	0.5019	**0.8333**	0.3285	0.4955	0.2987
	20	0.7866	**0.8730**	0.7996	0.7859	0.8476
	Avg.	0.6979	**0.6995**	0.4993	0.5711	0.5562

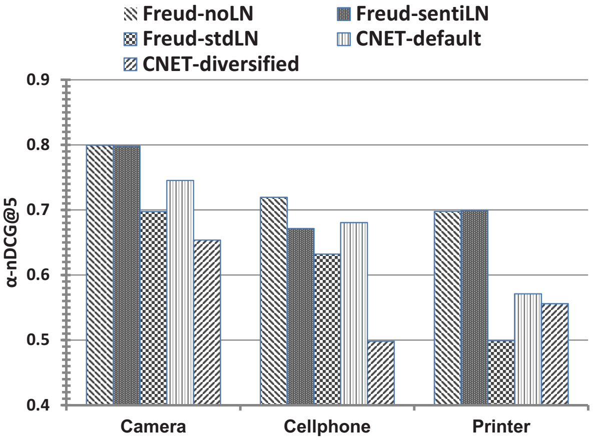

Figure 3: Feature-sentiment diversification performance of all system under individual product categories.

age α-nDCG@5 scores. Here we observe the same trend as before, i.e. also category-wise FREuD-noLN dominates all other systems when it comes to coverage and diversity performance. In the category-wise split-up we also see that the performance of FREuD-noLN and FREuD-sentiLN is at par for the categories camera and printer, while for cell phones FREuD-noLN has minor advantage.

Table 8: Results of statistical significance using a t-test at 5% significance level

FREuD Variations	CNET-default	CNET-diversified
Freud-noLN	*	*
Freud-sentiLN	*	*
Freud-stdLN	-	-

To check whether the difference in performance is significant, we conducted a paired t-test on the α-nDCG@5 scores. The results are show in Table 8. We see that FREuD-noLN and FREuD-sentiLN performed significantly better over the two baselines at 5% significance level. While the performance difference of FREuD-stdLN against the baselines is not significant.

Furthermore, we also tested the category-wise differences in performance for significance. These results are reported in Table 9 against CNET-default and in Table 10 against CNET-diversified at 5% significance level. Compared to CNET-default, a significant difference is observed only in the printer category for FREuD-noLN and FREuD-sentiLN. All other differences are not significant[9]. Compared to CNET-diversified, FREuD-noLN showed a significant improvement in all three categories and FREuD-sentiLN per-

[9] Compared to the global performance, this can be explained with the smaller sample size which makes it harder to demonstrate statistical significance.

Table 9: Category-wise statistical significance test of performance difference against *CNET-default* (at 5% significance level)

System	Camera	Cellphone	Printers
Freud-noLN	-	-	*
Freud-sentiLN	-	-	*
Freud-stdLN	-	-	-

Table 10: Category-wise statistical significance test of performance difference against *CNET-diversified* (at 5% significance level)

System	Camera	Cellphone	Printers
Freud-noLN	*	*	*
Freud-sentiLN	*	-	*
Freud-stdLN	-	*	-

formed significantly better in the categories camera and printer.

Summarizing our experiments, we can clearly see that FREuD-noLN performs best and significantly outperforms the baseline algorithms in selecting feature-centric sentiment diversified reviews. The improvement is consistent over the several product categories and significant at a global level.

7 Summary

In this paper we looked at the task of selecting a feature-centric sentiment diversified set of end user discussion contributions. The objective of this task is to rank a set of contributions such that the top-k entries cover a wide range of sub-topics or features addressed in a discussion as well as a diversified range of sentiments. We formalized the task and investigated it in the context of product reviews. With the FREuD approach we proposed a solution to this task. We constructed a real life data set composed of CNET product reviews and developed a gold standard data set for the purpose of evaluating feature-centric sentiment diversification approaches. We evaluated our proposed FREuD approach on this data set and compared it against two baselines systems. In this empirical evaluation we have been able to show that FREuD significantly outperforms both baseline systems.

In future work we will work on refinements of the estimation of sentiments expressed about a given feature. Our current approach operates with a document global, coarse grained sentiment value which is broken down to the feature level. Using a more fine-grained detection of sentiments in document segments might allow for a more detailed annotation of features with sentiments. Furthermore, we will consider alternative optimization strategies for selecting the top-k reviews. In this way we might overcome some of the situations in which our greedy algorithms provides sub-optimal solutions.

8 Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 257859, ROBUST.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3.

Bradley, M. M., and Lang, P. J. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

Carterette, B. 2011. An analysis of np-completeness in novelty and diversity ranking. *Information Retrieval* 14:89–106.

Clarke, C. L.; Kolla, M.; Cormack, G. V.; Vechtomova, O.; Ashkan, A.; Büttcher, S.; and MacKinnon, I. 2008. Novelty and diversity in information retrieval evaluation. In *Proc. Conference on Research and development in Information Retrieval*.

Eirinaki, M.; Pisal, S.; and Singh, J. 2011. Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*.

Fleiss, J., et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

Fleiss, J. 1981. *Statistical Methods for Rates and Proportions. Second Edition*. Wiley, John and Sons, Incorporated, New York, N.Y.

Ganesan, K.; Zhai, C.; and Viegas, E. 2012. Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, 869–878. New York, NY, USA: ACM.

Guo, H.; Zhu, H.; Guo, Z.; Zhang, X.; and Su, Z. 2009. Product feature categorization with multilevel latent semantic association. In *Proc. of CIKM*, 1087–1096.

Jansen, B. J.; Spink, A.; and Saracevic, T. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management* 36(2):207–227.

Krestel, R., and Dokoohaki, N. 2011. Diversifying product review rankings: Getting the full picture. In *Proc. International Conferences on Web Intelligence and Intelligent Agent Technology*, 138–145. Ieee.

Lin, C.; He, Y.; Everson, R.; and Rüger, S. M. 2012. Weakly supervised joint sentiment-topic detection from text. *IEEE Trans. Knowl. Data Eng.* 24(6):1134–1145.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, 342–351. New York, NY, USA: ACM.

Naveed, N.; Gottron, T.; Kunegis, J.; and Che Alhadi, A. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proc. of the 3rd International Conference on Web Science*.

Naveed, N.; Gottron, T.; Sizov, S.; and Staab, S. 2012. Freud: Feature-centric sentiment diversification of online discussions. In *WebSci '12: Proceedings of the 4th International Conference on Web Science*.

Nielsen, F. A. 2011. A new anew: evaluation of a word list for sentiment analysis in microblogs. In *Proc. of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, 93–98.

Pang, B., and Lee, L. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 115–124. Stroudsburg, PA, USA: Association for Computational Linguistics.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, 79–86.

Qiu, L.; Zhang, W.; Hu, C.; and Zhao, K. 2009. Selc: a self-supervised model for sentiment classification. In *Proc. Conference on Information and Knowledge Management*.

Salganik, M. J., and Watts, D. J. 2009. Web-based experiments for the study of collective social dynamics in cultural markets. *Topics in Cognitive Science* 1:439–468.

Shi, L., and MingYu, J. 2011. A dfm model of mining product features from customer reviews. In *Control, Automation and Systems Engineering (CASE), Conf. Proc.*

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, 173–180. Stroudsburg, PA, USA: Association for Computational Linguistics.

Tsaparas, P.; Ntoulas, A.; and Terzi, E. 2011. Selecting a comprehensive set of reviews. In *Proc. of the ACM international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM.

Zhai, Z.; Liu, B.; Xu, H.; and Jia, P. 2011. Clustering product features for opinion mining. In *Proc. International conference on Web search and data mining*. ACM.