Sentiment Diversification with Different Biases

Elif Aktolga and James Allan
Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts Amherst
Amherst, Massachusetts
{elif, allan}@cs.umass.edu

ABSTRACT

Prior search result diversification work focuses on achieving topical variety in a ranked list, typically equally across all aspects. In this paper, we diversify with sentiments according to an explicit bias. We want to allow users to switch the result perspective to better grasp the polarity of opinionated content, such as during a literature review. For this, we first infer the *prior sentiment bias inherent in a controversial topic* – the 'Topic Sentiment'. Then, we utilize this information in 3 different ways to diversify results according to various sentiment biases: (1) Equal diversification to achieve a balanced and unbiased representation of all sentiments on the topic; (2) Diversification *towards* the Topic Sentiment, in which the actual sentiment bias in the topic is mirrored to emphasize the general perception of the topic; (3) Diversification *against* the Topic Sentiment, in which documents about the 'minority' or outlying sentiment(s) are boosted and those with the popular sentiment are demoted.

Since sentiment classification is an essential tool for this task, we experiment by gradually degrading the accuracy of a perfect classifier down to 40%, and show which diversification approaches prove most stable in this setting. The results reveal that the proportionality-based methods and our SCSF model, considering sentiment strength and frequency in the diversified list, yield the highest gains. Further, in case the Topic Sentiment cannot be reliably estimated, we show how performance is affected by equal diversification when actually an emphasis either towards or against the Topic Sentiment is desired: in the former case, an average of 6.48% is lost across all evaluation measures, whereas in the latter case this is 16.23%, confirming that bias-specific sentiment diversification is crucial.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms: Experimentation, Algorithms, Measurement

Keywords: Diversity, Opinions, Sentiment, Proportionality

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

1. INTRODUCTION

In previous work diversification has mainly been applied for better topical variety in search results [9, 25, 26, 27]. Equal preference is typically given to all aspects. How can opinionated content exhibiting sentiments be diversified? Initial approaches have been presented [10, 17, 18]; however these only consider equal diversification. In this paper, we view the problem from a high-level perspective to allow for sentiment diversification according to *different biases*, which will be vital for situations like a literature review on a controversial topic.

Consider the topic 'global warming.' In a typical use case, a user engages in a comprehensive literature review with the aim of understanding the positions on this topic. This involves – besides searching and finding relevant opinionated documents [16] – understanding and mentally categorizing opinionated content. This can be done by organizing the discussed arguments by topical content; or, they can also be grouped by sentiment, such as positive, negative, neutral, and mixed [18]. We focus on facilitating the latter approach for the user. For some topics that can clearly be generalized into 'pro' versus 'con' arguments, this sentiment categorization is more natural, whereas it can be less obvious for topics like global warming that are associated with various arguments. Focusing on the *sentiment dimension* of these arguments, we can see that negative sentiments for global warming typically express criticism and concern about it and its effects on the environment. Those with positive sentiment often claim that worries about global climate change are unjustified ("there is no such issue"), playing down the concerns in a 'calming' (i.e., positive) way. Mixed or neutral statements either express no sentiments or contain an equal amount of positive and negative arguments. Those could be "I don't care", or "It's a serious problem but we're handling it" kind of stances towards global warming.

Getting back to our use case: while a balanced and unbiased presentation of the results helps the user understand various viewpoints on a topic, discerning the topic's polarity is harder if minority opinions are 'buried' in the results [18]. Therefore, the user should be able to *switch the result perspective* as needed. This way, she can either obtain a balanced or a biased view on majority or minority opinions, make her own comparisons across the representations, and perform this task in a more informed manner. Note that this is different from showing all positive *or* all negative *or* all neutral/mixed documents at a time: with such a representation the user would still need to draw her own conclusions about which sentiments form majority or minority

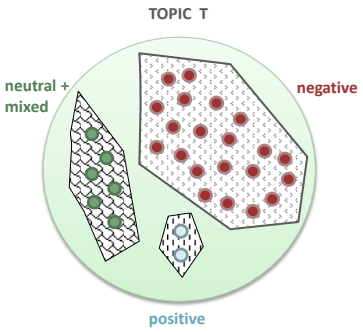

Figure 1: Topic Sentiment: Dots represent relevant documents for this topic, which are grouped according to their sentiments. The obtained sentiment distribution is used for sentiment diversification.

opinions. Our aim is to analyze this information for the user and to match the inferred trend as closely as possible in the results. For this, we need to have a good grasp of the topic a priori: we consider a large pool of relevant documents about the topic, which are grouped by sentiments as visualized in Figure 1. Then, we can infer the *topic's sentiment distribution or inherent bias* – Topic Sentiment – from this analysis. We categorize the aspects 'mixed' and 'neutral' together to represent the 'balanced' aspect, whereas 'positive' and 'negative' refer to arguments that are clearly biased towards one side only. If Figure 1 represented the Topic Sentiment for global warming, this could be interpreted as the issue being perceived with great concern since negative sentiments constitute the majority, and while there are some 'balanced' positions on it, the positive sentiments form a clear minority. By utilizing this information during diversification, three biases are emphasized in search results: (1) Equal diversification by preferring all sentiments equally. This allows for a balanced representation of all sentiments on the topic; (2) Diversification towards the Topic Sentiment, in which the resulting reranked list mirrors the actual sentiment bias in a topic. This approach highlights the general perception of a topic; (3) Diversification against the Topic Sentiment, in which documents about the minority sentiment(s) are boosted whereas those with the majority sentiment are demoted. Such a list highlights unusual and outlying opinions on the topic.

In this paper we propose different diversification models for sentiment diversity with these 3 biases, and perform experiments using the TREC Blog Track data [23]. Since sentiment classification is an essential tool for this task, we experiment by gradually reducing the accuracy of a perfect classifier down to 40%, and show which diversification approaches prove most stable in this setting. Further, in case the Topic Sentiment cannot be reliably estimated, we show how performance is affected by equal diversification when actually an emphasis either towards or against the Topic Sentiment is desired. The results reveal that particularly when highlighting minority sentiments, diversifying with the corresponding bias yields significant improvements.

2. RELATED WORK

There is a large amount of work in the area of topical diversity: the main aim is to eliminate topical redundancy in results while maximizing the number of documents containing novel information [1, 3, 6, 9, 25, 26, 27]. One of the earliest works is the maximal marginal relevance (MMR) approach [3], which employs content-based similarity measures to balance the tradeoff between novelty and relevance. More recently, researchers have shown explicit diversification approaches to be superior over implicit diversification techniques: well-known algorithms are xQuAD [25, 26, 27], IA-select [1], and more recently PM-1 and PM-2 [9]. Other approaches to topical diversity are language modeling based [31], probabilistic [4] and correlation-based [30]. Very recent research addresses personalized diversification [29], blog feed diversity [19, 28], and combined implicit and explicit aspect diversification [15].

Among these approaches, it is common to equally or uniformly diversify across all query aspects or subtopics due to the lack of data [9, 25, 26, 27]. Although the TREC Web Track diversity task provides topical query aspects [8], distributions over these aspects are not included. Agrawal et al. [1] use their own classifiers and judgments for obtaining query intent aspect distributions. Further, the NTCIR-9 Intent task provides non-uniform aspect probabilities [24]. One of our contributions in this paper is to present alternatives to the equal distribution approach (Section 3.4): a query's topic's sentiment distribution can be employed in various ways to yield an emphasis for a certain bias. Topical diversity could also benefit from these ideas.

The proportionality-based approaches PM-1 and PM-2 [9] distinguish themselves from prior research by explicitly matching the aspect distribution in the diversified list to the overall popularity of these aspects, thus yielding a proportionally diversified list. We adapt this approach to sentiment diversity, and propose a minor variation for dealing with retrieval limitations (Section 3.3).

Extensive work on opinion detection and retrieval has dealt with techniques to boost opinionated documents in retrieval [14, 28, 32, 33]. Prior work focusing on *opinion diversity* is very recent: Demartini and Siersdorfer describe a study about opinions in search results as given by popular search engines for controversial queries [11]. In later work, Demartini then tackles opinion diversification [10]: his approach is based on the xQuAD framework [25]. Retrieved search results are classified into the sentiment categories positive, negative, or objective. Diversification ensures maximum variety among these aspects with uniform preference. We implement this approach as the SCS model and combine it with the 3 biases (Section 3.2.1). The SCSF model is a further extension, presented in Section 3.2.2.

Kacimi and Gamper propose a different opinion diversification framework for controversial queries [17, 18]: three criteria are considered for diversification: topical relevance, semantic diversification, and sentiment diversification. Their model favors documents most different in sentiment direction and in the arguments they discuss. The sentiments are again one of positive, negative, and neutral. In the model the components are linearly combined; however, in order to find the documents maximizing the distances for all criteria the authors consider all subsets of documents. Our work differs from this work in several points: (1) We perform *sentiment diversification* only and not opinion diversification. Opinions refer to topical content, whereas sentiments are a non-topical aspect that we focus on in this paper. (2) This choice allows us to study sentiment diversification perfor-

mance with different biases, which has not been researched in prior work.

In this context, unlike topical diversity we make a simplifying assumption that each query belongs to one topic and therefore represents one topical aspect. We avoid dealing with ambiguity by using long and specific queries in our experiments, as explained in Section 4. That is, the topical dimension is kept static so we can focus on the varied sentiment dimension. We leave it to future work to explore the interplay of topical and sentiment aspects together for diversification.

3. SENTIMENT DIVERSIFICATION

3.1 Introduction

In order to diversify a retrieved list with respect to the distribution of sentiments in a query's topic, we need to introduce a few concepts first. Let Q be a query, and let T be the query's topic $T(Q)$, abbreviated as T for simplicity. As visualized in Figure 1, we define T to include all the relevant documents that can be retrieved for Q, i.e., $T = rel(Q)$. Further, let each document D in T have a sentiment, i.e., each document is positive, negative, neutral or mixed. These can be generalized to countable sentiment criteria $\sigma \in sent(T)$. We will use this sentiment information from T in our models to diversify search results according to the distribution of sentiments in the topic.

Sentiment criteria of the form positive, negative, and neutral/mixed can take different shapes when converted into a sentiment score. In the literature [10, 17, 23] we identified a document to either have a single discrete sentiment from $\{-1, 0, 1\}$, or the sentiment is broken down into three scores *positivity*, *negativity*, and *neutrality* such that they sum to 1.0 for a single document. We refer to these latter ones as finer grained "fractional scores" in the rest of the paper. Our models are designed for these kinds of scores, but discrete scores can also be handled by simple conversion as we will show later.

Below we consider two different diversification frameworks and present several modifications to them.

3.2 Retrieval-Interpolated Diversification

Algorithm 1 Retrieval Interpolated Diversification Framework.

```
1  S = ∅
2  while|S| < τ and |R| > 0
3        do
4           D* = arg max_{D∈R}  λRetC(Q) + (1 − λ)SentC(T)
5           R = R \ {D*}
6           S = S ∪ {D*}
7  return S
```

Algorithm 1 shows the Retrieval-Interpolated Diversification Framework, which is similar to xQuAD, first introduced by Santos et al. [25] for topical diversity. In this diversification framework, documents retrieved in R are iteratively added to the new ranked list S. The τ documents are chosen according to the maximization objective function in line 4:

$$D^* = argmax_{D \in R}\ \lambda \cdot \text{RetC}(Q) + (1 - \lambda) \cdot \text{SentC}(T) \quad (1)$$

where $\text{RetC}(Q)$ is the *retrieval contribution*, which is always estimated with $P(D|Q)$ – how likely D is to be relevant to Q by content, and $\text{SentC}(T)$ is the *sentiment contribution*, which we will define in two different ways below. The scores from these two components are interpolated for diversity estimation.

3.2.1 Sentiment Contribution by Strength (SCS)

In this version of the model we estimate the sentiment contribution in the maximization objective function (Equation 1) as follows:

$$\text{SentC}(T) = P(D, \bar{S}|T) \quad (2)$$

Here $P(D, \bar{S}|T)$ measures how much D can contribute to the sentiment diversity of S. Structurally, this resembles xQuAD [25] with the difference that the estimation is conditioned on the query's topic T.

In order to make the model more flexible towards sentiment scores, we define each document to have a fractional score for each sentiment criterion $\sigma \in sent(T)$. For example, a document may be classified as positive with 75% confidence. Then, this can be converted into a trinary score $P(D|\sigma = \text{positive}) = 0.75$, $P(D|\sigma = \text{neutral}) = 0.25$, and $P(D|\sigma = \text{negative}) = 0$. Fractional classification scores directly obtained from a classifier (such as logistic regression) fit in nicely into this framework. If documents are manually judged, they are often associated with only one 'dominant' sentiment score from $\{-1, 0, 1\}$ such as -1, which can be converted into a 100% negative score. Given this information, we can further decompose $P(D, \bar{S}|T)$ as follows:

$$P(D, \bar{S}|T) = \sum_{\sigma \in sent(T)} P(D, \bar{S}|\sigma) \cdot P(\sigma|T) \quad (3)$$

$$\stackrel{rank}{=} \sum_{\sigma \in sent(T)} P(D|\sigma) \cdot P(\bar{S}|\sigma) \cdot P(\sigma|T) \quad (4)$$

where $P(\bar{S}|\sigma)$ denotes the likelihood of σ not being satisfied by the documents already chosen into S (see below for further derivation) and $P(\sigma|T)$ stands for the importance of sentiment σ to topic T. This is discussed in detail in Section 3.4. From Equation 3 to Equation 4 we make the same independence assumption as Santos et al. [25]: the diversity estimation of D with respect to the sentiments σ can be made independently of the documents already selected into S. We continue with Equation 4:

$$\begin{aligned} &\sum_{\sigma \in sent(T)} P(D|\sigma) \cdot P(\bar{S}|\sigma) \cdot P(\sigma|T) \\ &= \sum_{\sigma \in sent(T)} P(D|\sigma) \cdot P(\sigma|T) \cdot \prod_{D_j \in S} P(\overline{D_j}|\sigma) \\ &= \sum_{\sigma \in sent(T)} P(D|\sigma) \cdot P(\sigma|T) \cdot \prod_{D_j \in S} 1 - P(D_j|\sigma) \end{aligned} \quad (5)$$

Here we make another independence assumption for $P(\overline{D_j}|\sigma)$ as Santos et al. [25]: the likelihood of not sampling D_j's sentiment from T is independent of the sentiments of the other documents in S. Since each D_j was independently chosen into S, this is a reasonable assumption.

To summarize, Equation 5 estimates the diversity of a document D by considering how well D represents each sentiment criterion, which is weighted by how important that

Algorithm 2 Diversity by Proportionality (PM-2).

$S = \emptyset$
$\forall \sigma \;\; s_\sigma = 0$
while $|S| < \tau$ and $|R| > 0$
  do
    for $\sigma \in sent(T)$
      do
        $quotient[\sigma] = \frac{v_\sigma}{2s_\sigma + 1}$
    $\sigma^* = \arg\max_\sigma quotient[\sigma]$
    $D^* = \arg\max_{D \in R} \;\; \lambda \cdot quotient[\sigma^*] \cdot P(D|\sigma^*) + (1 - \lambda) \sum_{\sigma \neq \sigma^*} quotient[\sigma] \cdot P(D|\sigma)$
    $R = R \setminus \{D^*\}$
    $S = S \cup \{D^*\}$
    for $\sigma \in sent(T)$
      do
        $s_\sigma = s_\sigma + \frac{P(D^*|\sigma)}{\sum_{\gamma \in sent(T)} P(D^*|\gamma)}$
return S

sentiment criterion is to T. This whole part is demoted according to how many documents of the same sentiment S already contains.

3.2.2 Sentiment Contribution by Strength and Frequency (SCSF)

We consider an alternative formulation of the sentiment contribution component above in Equation 1 in which the punish/reward factor is estimated slightly differently:

$$\text{SentC}(T) = P(D|T) \cdot (1 - P(S|T)) \tag{6}$$

where $P(D|T)$ stands for how important D's sentiment is for T, and $1 - P(S|T)$ describes how well the sentiments from T are already represented in S. We further derive:

$$\begin{aligned} & P(D|T) \cdot (1 - P(S|T)) \\ = \;& P(D|T) - P(D|T) \cdot P(S|T) \\ = \;& \sum_{\sigma \in sent(T)} P(D|\sigma) \cdot P(\sigma|T) - P(D|\sigma) \cdot P(\sigma|T) \cdot P(S|\sigma) \end{aligned} \tag{7}$$

Here we apply the Bayes' Rule to $P(S|\sigma)$:

$$P(S|\sigma) = \frac{P(\sigma|S) \cdot P(S)}{P(\sigma)} \stackrel{rank}{=} P(\sigma|S) \tag{8}$$

which is rank-equivalent since $P(S)$ is a constant across all documents in an iteration, and $P(\sigma)$, the prior probability of a particular sentiment, is equal across all sentiments. Hence we obtain from Equation 7:

$$\begin{aligned} & \sum_{\sigma \in sent(T)} P(D|\sigma) \cdot P(\sigma|T) - P(D|\sigma) \cdot P(\sigma|T) \cdot P(\sigma|S) \\ = \;& \sum_{\sigma \in sent(T)} P(D|\sigma) \cdot P(\sigma|T) \cdot (1 - P(\sigma|S)) \\ = \;& \sum_{\sigma \in sent(T)} P(D|\sigma) \cdot P(\sigma|T) \cdot P(\bar{\sigma}|S) \end{aligned} \tag{9}$$

Now we can see that the first part of Equation 9 is identical to Equation 5. We can estimate the components $P(D|\sigma) \cdot P(\sigma|T)$ the same way as described in Sections 3.2 and 3.4. However, $P(\bar{\sigma}|S)$, the likelihood of S not having sentiment σ, is new. We define its complement as follows:

$$P(\sigma|S) = \frac{sent(\sigma, S)}{|S|} \tag{10}$$

which is the number of documents in S having *dominant sentiment* σ. Each document in S can be mapped into its dominant or most confident sentiment class $\sigma \in sent(T)$, typically positive, negative, or neutral/mixed. Given this, we count the number of times a particular sentiment σ occurs in S as $sent(\sigma, S)$. We set $P(\sigma|S) = 0$ if $S = \emptyset$ to avoid zero division in the first iteration.

To summarize, this formulation calculates the punish/ reward factor directly from the *frequency* of documents present in the whole set S with certain sentiments. Contrarily, in the SCS model the *strength of sentiments* of each document in S is considered individually, whereas the frequency of such documents is implicit in the multiplication over all documents in S. In the experiments we empirically verify the effectiveness of the two models in sentiment diversification to draw conclusions about their usefulness.

3.3 Diversity by Proportionality

As a second diversification framework we consider Algorithm 2, the best-performing approach in Dang and Croft's work [9]. This framework is based on the Sante-Laguë method for seat allocation and is adapted here to sentiment diversification. In each iteration documents are chosen based on the proportionality of the diversified list. We only describe modified components due to space limitations. Instead of applying the algorithm to topical aspects, here it is employed together with sentiments $\sigma \in sent(T)$. Further, $P(D|\sigma)$ is estimated by means of fractional sentiment scores as defined in Section 3.2 instead of estimating the relevance of the document with respect to a (sub)topical aspect. Note that under this modification, a document is purely evaluated on the basis of its sentiments and not according to topical relevance.

The variables v_σ and s_σ in the quotient are important. The former is the number of relevant documents the sentiment σ should have, whereas the latter represents the estimated number of documents actually present in the list for σ. v_σ at a particular rank i can easily be inferred from $P(\sigma|T)$ as follows:

$$v_\sigma = \lfloor i \cdot P(\sigma|T) + 0.5 \rfloor \tag{11}$$

According to PM-2, s_σ is updated with the fractional sentiment scores of the chosen document, since each sentiment takes up a 'portion' of the seats in S. This denotes how well the chosen document represents each sentiment. For

the relationship between document sentiments and the topic sentiment distribution, please refer to Section 3.4.

3.3.1 Diversity by Proportionality with Minimum Available Votes (PM-2M)

Unlike the seat allocation problem in a voting system, in a retrieved list of documents there is an additional constraining factor. The top K documents retrieved from a search system constitute the source for diversification, so it is possible that a particular sentiment is underrepresented in this list. Unless the system requests more documents, the desired proportionality in the diversified list may not be optimally achieved with the current set of documents. In this situation, with respect to PM-2 the given votes v_σ overestimate l_σ, the *actual number of documents* with sentiment σ in the retrieved top K set. For a large enough rank K, this may result in a suboptimally diversified list where documents with an over-emphasized sentiment are exploited early in the ranks. Therefore, we propose a small change to the quotient defined in Algorithm 2:

$$quotient[\sigma] = \frac{\min(v_\sigma, l_\sigma)}{2s_\sigma + 1} \tag{12}$$

which ensures that the quotient does not over-emphasize the importance of a sentiment if data is missing in the retrieved list. This technique has a remote resemblance to disproportionate stratified sampling in that documents are chosen slightly differently than dictated by the topic sentiment distribution in favor of improved overall diversity. We refer to this modified diversification approach as PM-2M and compare its effectiveness to PM-2, SCS and SCSF in the experimental section 4.

3.4 Favoring Different Biases in Search Results

In the presentation of the diversification models above $P(\sigma|T)$ plays a central role in defining which sentiment bias is favored in search results. Intuitively, this component stands for the importance of sentiment σ to topic T. Below we present three different possible biases in search results that the estimation of $P(\sigma|T)$ impacts.

3.4.1 Equal Sentiment Diversification (BAL)

This is our baseline approach, which does not give preference to any sentiment, but weights them equally or uniformly. Therefore, this approach does *not* utilize information from the query's topic about its prior sentiment distribution. We set

$$P(\sigma|T) = \frac{1}{|sent(T)|} \tag{13}$$

which results in each sentiment criterion $\sigma \in sent(T)$ to be considered equally important. We refer to this bias method as 'Balance' (BAL) in Section 4.

We assume that with this balanced estimation the SCS model is equivalent to Demartini's [10] approach. Since this detail is not explicitly described in their work, it is most reasonable to assume an equal bias as in prior research.

3.4.2 Diversifying Towards the Topic Sentiment (CRD)

In this approach we choose to diversify the retrieved list *towards* the distribution of sentiments in the query's topic. Such results strongly represent the crowd's opinion(s). For this, we need information about the sentiments in T. Recall from Section 3.1 that T is defined as a topic space to include all the relevant documents that can be retrieved for the query Q, i.e., $T = rel(Q)$. Then, we can map each relevant document into its dominant or most confident sentiment class $\sigma \in sent(T)$. Given this, we count the number of times a particular sentiment σ occurs in T as $sent(\sigma, T)$. This allows us to interpret $P(\sigma|T)$ as the likelihood of sentiment σ being drawn from T:

$$P(\sigma|T) = \frac{sent(\sigma, T)}{|T|} \tag{14}$$

which represents the fraction of documents in T with dominant sentiment σ; for instance the fraction of positive documents in T. We name this bias as 'Crowd' (short: CRD).

3.4.3 Diversifying Against the Topic Sentiment (OTL)

What if a user is interested in viewing minority sentiments on the topic? For favoring outlying sentiments, we need to diversify the search results *against* the Topic Sentiment. For this, we introduce one minor modification to CRD above: Let the n sentiment estimations for $\sigma \in sent(T)$ be sorted in increasing order of $P(\sigma|T)$. Then, for each σ at rank i we swap its estimation $P(\sigma|T)$ with the one at rank $n - i$. This 'reverses' the values in the topic distribution without changing the properties of the distribution. Consequently, if originally in T positive documents are strongly favored and negative documents are least favored, this trend is reversed through the value swap in T so that outlying sentiments (negative documents) will be strongly preferred during diversification. We refer to this bias as 'Outlier' (OTL) in the experiments (Section 4).

Irrespective of the preferred bias, we apply Add-1 Smoothing [5] to $P(\sigma|T)$ estimates to account for zero probabilities. In order to correct such unrealistic estimations, an unobserved sentiment class is assigned a very small probability, and the estimations for the other sentiment classes are adjusted accordingly.

4. EXPERIMENTS

4.1 Setup

Retrieval Corpus As retrieval corpus we use the TREC Blog Track data from 2006 and 2008 [23] for all our experiments. For preparation, the DiffPost algorithm is applied to the corpus for better retrieval as shown in prior work [20, 22]. Further, we perform stop word removal and Porter stemming.

Queries and Retrieval Model We split the 150 TREC Blog Track 2008 queries into 3 non-overlapping randomly chosen sets of size 50 each in order not to bias training or testing towards a specific year: split 1 is used for training and tuning parameters; the results in this paper are reported on split 2, and split 3 is reserved for sentiment classifier training. For our diversification experiments, we use a strong retrieval baseline: the queries' stopped title and description texts are combined for use with the Sequential Dependence Model in Lemur/Indri [21], smoothed using Dirichlet ($\mu = 10,000$). All diversification models are applied to the top $K = 50$ retrieved documents as determined

during training. The retrieval scores are normalized to yield document likelihood scores.

Sentiment Classification The sentiment classifier is trained as a logistic regression model using Liblinear [13] with default settings. For this, we utilize the judged documents from the 50 split 3 TREC Blog Track queries. Training is done for three classes – positive, negative, and neutral to obtain probability estimates that are employed as fractional scores for sentiment estimation (Section 3.2.1). As features we extract Sentiwordnet 3.0 terms with their length-normalized term frequencies in the documents [12].

Topic Sentiment Estimation Given a query, the topic sentiment distribution can be estimated in various ways: (1) in the form of opinion relevance judgments for a pool of documents where all judged relevant documents are included in the distribution; (2) by retrieving the top M documents from a separate corpus or web search engine and tagging them with sentiment judgments. We experimented with both approaches but only present the results for (1) here due to space limitations: we use the relevance judgments from the TREC 2008 Blog Track [23], which are divided into the same sentiment aspects as required in the models.

4.2 Evaluation Measures

The sentiment diversification approaches are evaluated using standard evaluation measures that were designed for topical diversity: Precision-IA [1], s-recall [31], α-NDCG [6], ERR-IA [2], and NRBP [7]. The former two measures are set-based, whereas the remaining ones are cascade measures as described by Ashkan and Clarke [2], punishing redundancy through parameters α (α-NDCG, ERR-IA, NRBP) and additionally β (NRBP), which represents user patience. In order to measure sentiment diversity with a chosen bias, we implement all the measures in their intent-aware (or for us, 'sentiment-aware') version [1, 2]. Hence, the weighted average over the sentiment-dependent scores of a measure is computed as given by measure-IA for a query Q and topic T:

$$\text{measure-IA}(Q,T) = \sum_{\sigma \in sent(T)} P(\sigma|T) \cdot \text{measure}(Q|\sigma) \quad (15)$$

where $P(\sigma|T)$ defines the weight for the sentiment-specific result yielded by measure$(Q|\sigma)$.

Intent-aware measures can be rank-specific such as Precision-IA@k or α-NDCG@k for example, or rank-independent as NRBP. We utilize another rank-specific measure defined by Dang and Croft [9], Cumulative Proportionality (CPR) at rank K:

$$CPR@K = \frac{1}{K} \sum_{i=1}^{K} PR@i \quad (16)$$

in which $PR@i$ is computed as the inverse normalized disproportionality at rank i (see [9] for details). Here, we define the disproportionality at rank i as follows:

$$DP@i = \sum_{\sigma \in sent(T)} c_\sigma (v_\sigma - s_\sigma)^2 + \frac{1}{2} n_{NR}^2 \quad (17)$$

where v_σ is the number of relevant documents the sentiment σ should have, s_σ is the number of relevant documents actually found for σ, n_{NR} is the number of documents that are non-relevant (to any sentiment), and $c_\sigma = 1$ if $v_\sigma \geq s_\sigma$, 0 otherwise. This measure allows us to assess how proportional the diversified list is with respect to the desired topic distribution. v_σ can be inferred from the true topic sentiment distribution $P(\sigma|T)$ in the same way as detailed in Equation 11. As noted by Dang and Croft [9], CPR penalizes the under-representation of aspects (here: sentiments) and the over-representation of non-relevant documents.

4.3 Results

In this section we discuss the results of the retrieval baseline SDM and all the proposed diversification models in Section 3, SCS, SCSF, PM-2 and PM-2M, with the three biases, Crowd (CRD), Balance (BAL) and Outlier (OTL). The interpolation parameter $\lambda \in \{0.0, ..., 1.0\}$ is tuned in 0.1 steps separately for each model and bias on our training split. The results are presented with fixed parameters K and λ on test split 2, and the evaluation is performed with the TREC 2008 Blog Track judgments at rank 20. α-NDCG, ERR-IA, and NRBP require parameters, which are set to $\alpha = 0.5$ and $\beta = 0.5$.

4.3.1 Straight-Bias Experiments

Our primary aim in the experiments is to evaluate sentiment *diversification* performance. Sentiment classification is an important part of the system since both the to-be-diversified documents need to be tagged with sentiments, as well as those for the topic sentiment distribution estimation. Since a 'full evaluation' of sentiment diversification techniques on a publicly available dataset has not been done yet in prior work, it is important to understand how sentiment classification quality affects diversification performance. Therefore, we start with a "perfect system" in which classification accuracy is 100% for judged documents. For unjudged documents the trained sentiment classifier described in Section 4.1 is applied. We then gradually reduce the overall classification performance in 10% steps until 40% as follows: given the top $K = 50$ retrieved documents for a query, before diversification we randomly sample the ranks at which the true classification label is switched to another label randomly to achieve the desired classification error for each query.

Figure 2 shows the results for the straight-bias experiment, in which the topic sentiment distribution employed in experiment and evaluation *underlies the same favored bias*. For instance, the left-most column in Figure 2 shows the results for diversifying towards Crowd in the experiments, and measuring performance for Crowd in the evaluation (short-CRD-CRD). The middle column shows the same for Balance (short: BAL-BAL), and the right-most column is for the Outlier bias (short: OTL-OTL).

At the top-most row in the Precision-IA@20 graphs we observe a big gap between the SDM baseline and SCS model versus the rest of the models. For Crowd, the SCSF model only dominates when classification accuracy is at least 60% while it achieves the best (however not statistically significant) numbers in the Outlier graph. PM-2M and PM-2 also perform well and dominate some of the lower accuracy ranges. Statistical significance with the paired two-sided t-test (p-value$<$ 0.05) is indicated in the graphs with circles:

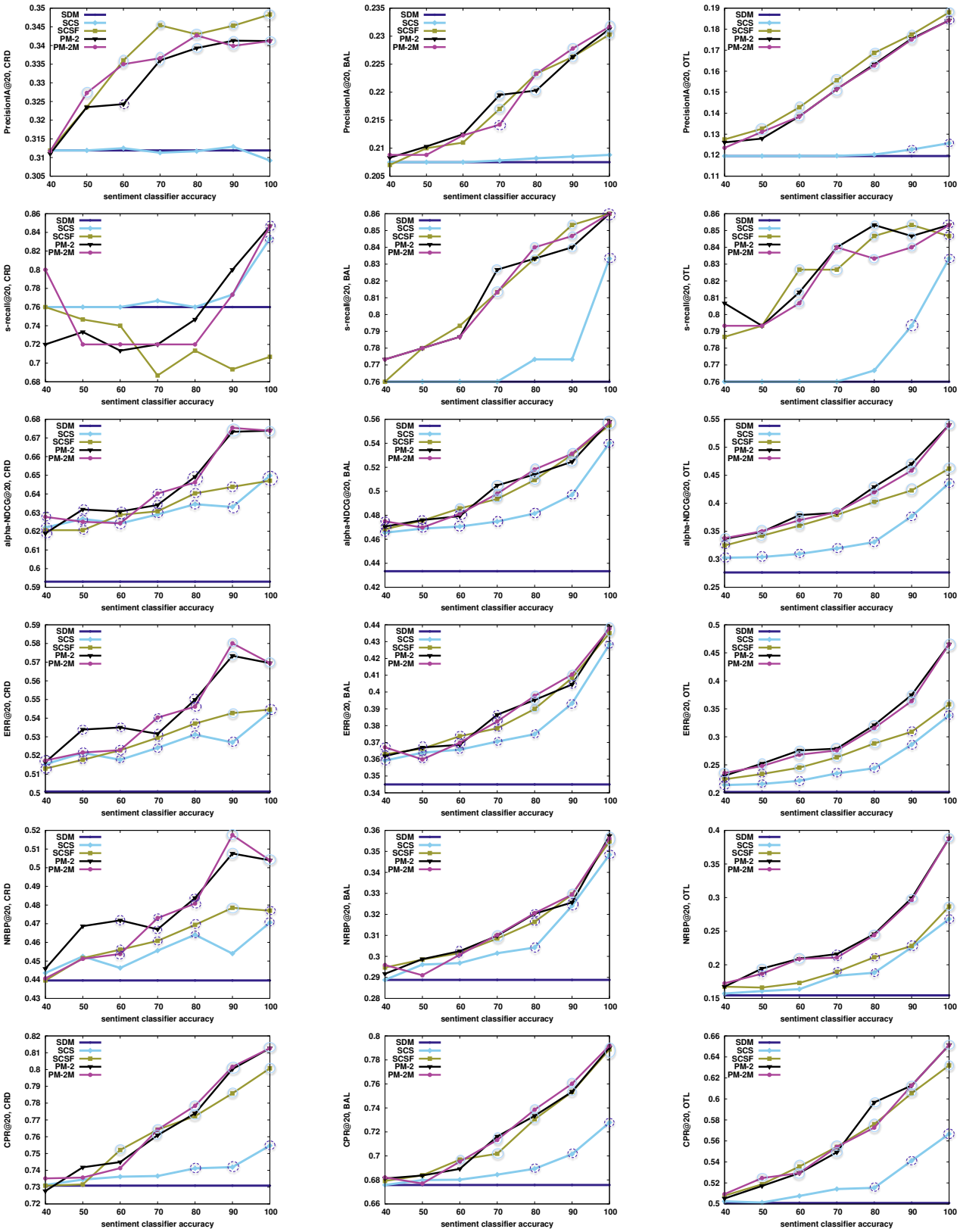

Figure 2: Straight-Bias Experiment over test split varying sentiment classifier accuracies on the x-axis and each one measure and bias on the y-axis. The leftmost column is for the Crowd bias (CRD), the middle one for Balance (BAL), and the rightmost one for Outlier (OTL). Dark purple dotted circled points indicate statistical significance over the SDM baseline with p-value < 0.05 using the paired two-sided t-test, whereas light blue circled points indicate the same over the SCS and SDM models.

Measure	**Precision-IA@20**		**α-NDCG@20**		**ERR-IA@20**		**NRBP**		**CPR@20**	
Exp-Eval	BAL-CRD	CRD-CRD	BAL-CRD	CRD-CRD	BAL-CRD	CRD-CRD	BAL-CRD	CRD-CRD	BAL-CRD	CRD-CRD
SDM baseline	0.312	0.312	0.593	0.593	0.501	0.501	0.440	0.440	0.731	0.731
SCS	0.308	0.309	0.642	**0.650**	0.532	**0.543**	0.453	**0.471**	0.750	**0.755**
SCSF	0.298	**0.348**	0.648	0.647	0.533	**0.545**	0.456	**0.477**	0.774	**0.801**
PM-2	0.302	**0.341**	0.642	**0.674**	0.526	**0.570**	0.446	**0.504**	0.772	**0.813**
PM-2M	0.298	**0.341**	0.639	**0.674**	0.521	**0.570**	0.440	**0.504**	0.767	**0.813**

Table 1: Cross-Bias Experiment over test split with perfect sentiment classifier to compare performance loss when diversifying equally (BAL-CRD) if actually diversification for the Crowd bias is desired (CRD-CRD). Bold entries in CRD-CRD columns are statistically significant over corresponding entries in BAL-CRD with p-value < 0.004 using the paired two-sided t-test.

the lighter blue circles refer to the result being significant over the SCS and SDM models, whereas the darker dotted circles indicate significance over the SDM model only. In the Precision-IA@20 graphs the results for SCSF and the proportionality-based methods are significant over SCS and SDM even for lower accuracies. We conclude that if precision is important, the SCSF diversification model should be used.

Among the s-recall@20 graphs the one for Crowd is the most arbitrary one. Performance drops well below the baseline for the SCSF and proportionality-based methods with medium quality classification: this indicates that the majority sentiment(s) are being emphasized too strongly, whereas minority sentiments appear much later in the ranked list for the first time, which is when the subtopic-recall measure is affected. This is expected, since we explicitly diversify in favor of majority sentiments. In the Balance and Outlier graphs for s-recall@20 there is no such trend, however precision is not as high for those biases as it is for Crowd. This is a typical precision versus recall tradeoff observation.

The next row shows results for α-NDCG@20, followed by ERR-IA@20 and NRBP: we note that the trends in these graphs look very similar, although the ranges of the values differ greatly. It is interesting to observe that the peak performance for the proportionality-based methods for the Crowd bias is not at 100% classification accuracy, but at 90%. What these three measures have in common is punishing redundancy based on the rank and sentiment criterion in addition to non-relevance. Since usually there are many documents with the majority sentiment in the retrieved list to start with, a strong emphasis on a single sentiment criterion results in more redundancy. With the 10% error in classification documents with other sentiments are slightly boosted, yielding better overall varied ranking. In the Balance and Outlier graphs this trend cannot be observed, since the Balance bias does not strongly emphasize a single sentiment criterion to begin with. Concerning the Outlier bias, there are fewer documents with minority sentiments in the retrieved list to cause the same 'clustered' ranking effect as for Crowd. Summarizing the trends across the α-NDCG@20, ERR-IA@20, and NRBP graphs we make the following conclusion: if ranking is important, the PM-2 and PM-2M methods should be chosen.

Finally, we look at the last row of graphs with the CPR@20 results: this measure evaluates how proportional the overall list is with respect to the chosen bias. PM-2 and PM-2M achieve the best results, which is closely followed by SCSF. PM-2 and SCSF are more appropriate for lower classification accuracies ($\leq 70\%$), whereas PM-2M performs slightly better with better classification quality.

Looking at the fixed values of the interpolation parameter λ during training for this experiment, the following insights can be drawn: for the SCS model, across all classifier accuracies and biases generally $\lambda \geq 0.6$ values are preferred. So this model performs best with a weaker emphasis on diversity, which pulls it closer to the SDM baseline as observed in the graphs of Figure 2. SCSF on the other hand has a good mixture of higher and lower λ values across classifier accuracies and biases, with many of them being < 0.5, particularly when the classifier is more accurate. So a heavier emphasis on the diversification part helps this model. The distinguishing feature between SCS and SCSF is the consideration of sentiment frequencies in addition to sentiment strength contributions. When the classifier is noisy however ($< 60\%$) and thus sentiment frequency counts are not accurate, SCSF also benefits from higher λ values. In the PM-2 and PM-2M models the role of λ is different: it balances the emphasis on the chosen aspect $\sigma*$ versus all the other aspects $\sigma \in sent(T), \sigma \neq \sigma*$. Here, consistently higher λ values are preferred for both models, i.e., a high emphasis on the chosen aspect and a minimal weight on the other ones seems most beneficial. The effectiveness of these two models solely relies on sentiment estimations: given our adaption of PM-2 from its original definition ([9]) to sentiment diversity, the retrieval scores are not used for building the diversified list.

4.3.2 Cross-Bias Experiments

Consider the following real-world setting: for certain topics, it may not be feasible to collect data for calculating Topic Sentiment estimations, or suitable corpora may currently not be available. This could happen if the topic is very new and the data is not substantial enough for drawing general conclusions. If judgments shall be obtained, the data tagging effort may also be a burden. In such a situation we can fall back to the Balance bias or equal diversification approach [9, 25, 26, 27]. Naturally, the next question to answer is how much performance is lost when diversifying with Balance instead of the desired bias such as Outlier. The cross-bias experiments in this section investigate this case, and enable us to draw conclusions about the value of collecting and using information about topic sentiment distributions for controversial topics.

We analyze two cases. The first, presented in Table 1 shows the results for equally diversifying for Balance, but performance is measured for the Crowd bias (BAL-CRD). This is contrasted with diversifying for the Crowd bias, and evaluating for the same (CRD-CRD). Bold entries in CRD-CRD indicate statistical significance over BAL-CRD with a p-value of < 0.004 (t-test, as before). The SDM baseline is included for comparison. We omit s-recall@20 due to space limitations. All CRD-CRD results for the proportionality-based methods are significant over BAL-CRD results, whereas for the SCSF and SCS models there are a few exceptions.

Measure	Precision-IA@20		α-NDCG@20		ERR-IA@20		NRBP		CPR@20	
Exp-Eval	BAL-OTL	OTL-OTL	BAL-OTL	OTL-OTL	BAL-OTL	OTL-OTL	BAL-OTL	OTL-OTL	BAL-OTL	OTL-OTL
SDM baseline	0.120	0.120	0.277	0.277	0.202	0.202	0.155	0.155	0.501	0.501
SCS	0.126	0.126	0.413	**0.436**	0.309	**0.338**	0.237	**0.268**	0.562	**0.567**
SCSF	0.164	**0.188**	0.433	**0.462**	0.320	**0.358**	0.243	**0.287**	0.624	0.632
PM-2	0.166	**0.184**	0.447	**0.540**	0.337	**0.465**	0.262	**0.388**	0.632	**0.651**
PM-2M	0.166	**0.184**	0.446	**0.540**	0.336	**0.465**	0.261	**0.388**	0.634	**0.651**

Table 2: Cross-Bias Experiment over test split with perfect sentiment classifier to compare performance loss when diversifying equally (BAL-OTL) if actually diversification for the Outlier bias is desired (OTL-OTL). Bold entries in OTL-OTL columns are statistically significant over corresponding entries in BAL-OTL with p-value < 0.05 using the paired two-sided t-test.

	SDM baseline		SCS	
Rank	Excerpt	Sent.	Excerpt	Sent.
1	The Religious Policeman: Mutt the Muttawa	-	The Religious Policeman: Mutt the Muttawa	-
2	Happy Feminist: PROTESTING GENDER...	o	Happy Feminist: PROTESTING GENDER...	o
3	Between tradition and demands for change	o	First women to win in Saudi elections	+
4	Saudi mobile carriers ban SMS voting...	-	Between tradition and demands for change	o
5	Saudi Arabia, Ever Our Friends And Allies	-	Saudi mobile carriers ban SMS voting...	-
6	Orientalism and Islamophobia	o	Saudi Arabia, Ever Our Friends And Allies	-
7	Laws discriminate against women...	-	Orientalism and Islamophobia	o
8	...who urged SA to improve women's rights...	o	Laws discriminate against women...	-
9	Being a Child in Saudi Arabia	o	...who urged SA to improve women's rights...	o
10	Depressing Post: ...woman filed a case against...	-	Being a Child in Saudi Arabia	o
	SCSF		PM-2	
Rank	Excerpt	Sent.	Excerpt	Sent.
1	The Religious Policeman: Mutt the Muttawa	-	The Religious Policeman: Mutt the Muttawa	-
2	Happy Feminist: PROTESTING GENDER...	o	Saudi Arabia, Ever Our Friends And Allies	-
3	Saudi Arabia, Ever Our Friends And Allies	-	First women to win in Saudi elections	+
4	Orientalism and Islamophobia	o	Happy Feminist: PROTESTING GENDER...	o
5	First women to win in Saudi elections	+	Orientalism and Islamophobia	o
6	Laws discriminate against women...	-	Laws discriminate against women...	-
7	Depressing Post: ...woman filed a case against...	-	Depressing Post: ...woman filed a case against...	-
8	Their shabby treatment of women...	-	Their shabby treatment of women...	-
9	Oprah is being smuggled into Saudi Arabia...	-	Thumbs up for the Saudi ladies.	+
10	Between tradition and demands for change	o	Between tradition and demands for change	o

Table 3: Crowd Bias: Top 10 results with 4 models for query number 1007, 'women in Saudi Arabia.' - denotes a negative document, o refers to mixed/neutral, and + to positive.

We observe a maximum loss of 16.92% for Precision-IA@20 with SCSF, and an average loss of 6.48% across all measures and diversification approaches.

The second case is presented in Table 2: we observe the results for equally diversifying for Balance, but performance is measured for the Outlier bias (BAL-OTL). This is contrasted with diversifying for the Outlier bias, and evaluating for the same (OTL-OTL). Similar to Table 1 the results are statistically significant for OTL-OTL over BAL-OTL, but the losses with equal diversification are more heavily pronounced here: there is a maximum loss of 48.79% for NRBP with PM-2M, and an average loss of 16.23% across all measures and diversification approaches. So for highlighting minority sentiments through diversification it is even more important to collect biased data about topic sentiment distributions than it is for emphasizing majority sentiments as observed in Table 1. This way diversification can be performed with the intended bias rather than with equal diversification, which yields significantly worse results.

We presented the cross-bias experiments with perfect sentiment classification to reveal the maximum performance loss. As classification accuracy degrades, the losses become smaller but remain noticeable.

4.3.3 *Analysis with Specific Queries*

To see the models in action, we look at the output for one query in Table 3, number 1007 from the TREC Blog Track: 'women in Saudi Arabia', asking for opinions about the treatment of women in Saudi Arabia. We show titles or characteristic excerpts from the documents together with their overall sentiment. The topic of this query has the following Topic Sentiment: 67% negative, 17% mixed/neutral, and 16% positive. Here we diversify for the Crowd Bias, so the aim is to mirror this distribution in the results. The top 10 retrieved results with the SDM baseline are presented at the top left: this result list does not include any positive documents, and an equal amount of negative and mixed/neutral documents, which is clearly unsatisfactory for a Crowd bias representation of the results. The SCS model includes one positive document at rank 3, since lower ranked documents through the SDM baseline can be pulled up by the diversification models. Although the documents are nicely shuffled around across ranks, the ratio of the sentiments is still not close to the Topic Sentiment. The SCSF model is able to correct this, explicitly considering the frequency of documents with their dominant sentiments: we have 6 negative documents, 3 mixed/neutral, and 1 positive. But 4 negative documents are clustered right after each other, which slightly affects measures such as α-NDCG@10. The PM-2 results (bottom right) use the overall proportionality of the sentiments in the list as a guidance for choosing further documents: here, a second positive document is pulled up from lower ranks, yielding the best CPR@10 score among the 4 models for this query at a cost of slightly lower Precision-IA@10 than SCSF. With 5 negative documents, 3 mixed/neutral ones, and 2 positive documents we are very close to the desired distribution of sentiments.

5. CONCLUSIONS & FUTURE WORK

In this paper we demonstrate how to diversify search results according to sentiments by considering a pre-defined bias. This allows us to emphasize either majority or minority sentiments during diversification, or to give an unbiased representation across all sentiment aspects. For this, we introduce several diversification models that use sentiments and topic sentiment distributions. Diversifying the output of a strong retrieval baseline, the results on the TREC Blog Track data reveal that the proportionality-based methods and the SCSF model perform best according to most measures, but an individual choice should be made based on the quality of the sentiment classifier at hand. Finally, we demonstrate the value of using biases and collecting topic sentiment distribution estimations by means of cross-bias experiments in which equal diversification is performed instead of the desired bias.

The ideas presented in this paper are not only valuable for sentiment diversity, but can also be applied to topical diversity with modifications. To what extent does it make sense to consider biases for topical diversity? For instance, with an Outlier bias-like approach underrepresented query aspects could be highlighted in search results. Further, we have proposed different extensions to existing diversification models such as xQuAD and PM-2 with the SCSF and PM-2M models, which may be effective for topical diversity as well.

There are many directions for future work: (1) Exploring other biases applicable for sentiment diversity; (2) We found that our trained 3-class sentiment classifier and ready-to-use classifiers on the web perform rather poorly at document-level sentiment classification. State-of-the-art sentiment classification works better on sentences or short text, but interpreting the overall sentiment of a document is more difficult, particularly on the web. Therefore, advances in this area would greatly benefit sentiment diversification so that it can be applied to the web beyond the TREC Blog Track; (3) In case this is difficult to realize, how can the diversification models be adapted to yield higher gains with noisy classification input; (3) Analyzing opinion or topical arguments and sentiments together with biases. One question to solve is what kind of biases could be defined to capture both, and whether more fine-grained topic-specific biases would be required.

6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part under subcontract #19-000208 from SRI International, prime contractor to DARPA contract #HR0011-12-C-0016, and in part by NSF grant #IIS-11217281. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proc. of WSDM*, 2009.

[2] A. Ashkan and C. L. Clarke. On the informativeness of cascade and intent-aware effectiveness measures. In *Proc. of WWW*, 2011.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of SIGIR*, 1998.

[4] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *Proc. of SIGIR*, 2006.

[5] S. F. Chen and J. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proc. of ACL*, 1996.

[6] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proc. of SIGIR*, 2008.

[7] C. L. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Proc. of ICTIR*, 2009.

[8] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In *Proc. of TREC-2009*, 2009.

[9] V. Dang and W. B. Croft. Diversity by proportionality: an election-based approach to search result diversification. In *Proc. of SIGIR*, 2012.

[10] G. Demartini. Ares: a retrieval engine based on sentiments sentiment-based search result annotation and diversification. In *Proc. of ECIR*, 2011.

[11] G. Demartini and S. Siersdorfer. Dear search engine: what's your opinion about...?: sentiment analysis for semantic enrichment of web search results. In *Proc. of SEMSEARCH*, 2010.

[12] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proc. of LREC*, pages 417–422, 2006.

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[14] B. He, C. Macdonald, and I. Ounis. Ranking opinionated blog posts using opinionfinder. In *Proc. of SIGIR*, 2008.

[15] J. He, V. Hollink, and A. de Vries. Combining implicit and explicit topic representations for result diversification. In *Proc. of SIGIR*, 2012.

[16] X. Huang and W. B. Croft. A unified relevance model for opinion retrieval. In *Proc. of CIKM*, 2009.

[17] M. Kacimi and J. Gamper. Diversifying search results of controversial queries. In *Proc. of CIKM*, 2011.

[18] M. Kacimi and J. Gamper. Mouna: mining opinions to unveil neglected arguments. In *Proc. of CIKM*, 2012.

[19] M. Keikha, F. Crestani, and W. B. Croft. Diversity in blog feed retrieval. In *Proc. of CIKM*, 2012.

[20] Y. Lee, S.-H. Na, J. Kim, S.-H. Nam, H.-Y. Jung, and J.-H. Lee. Kle at trec 2008 blog track: Blog post and feed retrieval. In E. M. Voorhees and L. P. Buckland, editors, *Proc. of TREC, Gaithersburg, Maryland, USA, November 18-21, 2008*, volume Special Publication 500-277. National Institute of Standards and Technology (NIST), 2008.

[21] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proc. of SIGIR*, 2005.

[22] S.-H. Nam, S.-H. Na, Y. Lee, and J.-H. Lee. Diffpost: Filtering non-relevant content based on content difference between two consecutive blog posts. In *Proc. of ECIR*, 2009.

[23] Ounis, M. de Rijke, C. Macdonald, G. Mishne, and Soboroff. Overview of the TREC-2006 Blog Track. In *Proc. of TREC*, 2006.

[24] T. Sakai and H. Joho. Overview of ntcir-9, 2011.

[25] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proc. of WWW*, 2010.

[26] R. L. Santos, C. Macdonald, and I. Ounis. Selectively diversifying web search results. In *Proc. of CIKM*, 2010.

[27] R. L. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proc. of SIGIR*, 2011.

[28] R. L. T. Santos, C. Macdonald, R. McCreadie, I. Ounis, and I. Soboroff. Information retrieval on the blogosphere. *Found. Trends Inf. Retr.*, 6(1), Jan. 2012.

[29] D. Vallet and P. Castells. Personalized diversification of search results. In *Proc. of SIGIR*, 2012.

[30] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proc. of SIGIR*, 2009.

[31] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proc. of SIGIR*, 2003.

[32] W. Zhang, L. Jia, C. Yu, and W. Meng. Improve the effectiveness of the opinion retrieval and opinion polarity classification. In *Proc. of CIKM*, 2008.

[33] W. Zhang, C. Yu, and W. Meng. Opinion retrieval from blogs. In *Proc. of CIKM*, 2007.