Hierarchical Re-estimation of Topic Models for Measuring Topical Diversity

Hosein Azarbonyad, Mostafa Dehghani, Tom Kenter, Maarten Marx, Jaap Kamps, and Maarten de Rijke

University of Amsterdam, Amsterdam, The Netherlands
{h.azarbonyad,dehghani,tom.kenter,maartenmarx,kamps,derijke}@uva.nl

Abstract. A high degree of topical diversity is often considered to be an important characteristic of interesting text documents. A recent proposal for measuring topical diversity identifies three elements for assessing diversity: words, topics, and documents as collections of words. Topic models play a central role in this approach. Using standard topic models for measuring diversity of documents is suboptimal due to *generality* and *impurity*. General topics only include common information from a background corpus and are assigned to most of the documents in the collection. Impure topics contain words that are not related to the topic; impurity lowers the interpretability of topic models and impure topics are likely to get assigned to documents erroneously. We propose a hierarchical re-estimation approach for topic models to combat generality and impurity; oure re-estimation approach operates at three levels: words, topics, and documents. Our re-estimation approach for measuring documents' topical diversity outperforms the state of the art on PubMed dataset which is commonly used for diversity experiments.

1 Introduction

Quantitative notions of topical diversity in text documents are useful in several contexts, e.g., to assess the interdisciplinarity of a research proposal [3] or to determine the interestingness of a document [2]. An influential formalization of diversity has been introduced in biology [16]. It decomposes diversity in terms of *elements* that belong to *categories* within a *population* [19] and formalizes the diversity of a population d as the expected distance between two randomly selected elements of the population:

$$div(d) = \sum_{i=1}^{T} \sum_{j=1}^{T} p_i p_j \delta(i,j), \tag{1}$$

where p_i and p_j are the proportions of categories i and j in the population and $\delta(i,j)$ is the distance between i and j. Bache et al. [3] have adapted this notion of diversity to quantify the topical diversity of a text document. Words are considered elements, topics are categories, and a document is a population. When using topic modeling for measuring topical diversity of text document d, Bache et al. [3] model elements based on the probability of a word w given d, $P(w|d)$, categories based on the probability of w given topic t, $P(w|t)$, and populations based on the probability of t given d, $P(t|d)$.

In probabilistic topic modeling, at estimation time, these distributions are usually assumed to be sparse. First, the content of a document is assumed to be generated by

a small subset of words from the vocabulary (i.e., $P(w|d)$ is sparse). Second, each topic is assumed to contain only some topic-specific related words (i.e., $P(w|t)$ is sparse). Finally, each document is assumed to deal with a few topics only (i.e., $(P(t|d)$ is sparse). When approximated using currently available methods, $P(w|t)$ and $P(t|d)$ are often dense rather than sparse [12, 18, 20]. Dense distributions cause two problems for the quality of topic models when used for measuring topical diversity: *generality* and *impurity*. General topics mostly contain general words and are typically assigned to most documents in a corpus. Impure topics contain words that are not related to the topic. Generality and impurity of topics both result in low quality $P(t|d)$ distributions.

We propose a hierarchical re-estimation process for making the distributions $P(w|d)$, $P(w|t)$ and $P(t|d)$ more sparse. We re-estimate the parameters of these distributions so that general, collection-wide items are removed and only salient items are kept. For the re-estimation we use the concept of *parsimony* [8] to extract only essential parameters of each distribution.

Our main contributions are: (1) We propose a hierarchical re-estimation process for topic models to address two main problems in estimating topical diversity of text documents, using a biologically inspired definition of diversity. (2) We study the efficacy of each level of re-estimation, and improve the accuracy of estimating topical diversity, outperforming the current state-of-the-art [3] on a publicly available dataset commonly used for evaluating document diversity [1].

2 Related work

Our hierarchical re-estimation method for measuring topical diversity relates to measuring text diversity, improving the quality of topic models, model parsimonization, and evaluating topic models.

Text diversity and interestingness. Recent studies measure topical diversity of document [2, 3, 7] by means of Latent Dirichlet Allocation (LDA) [4]. The main diversity measure in this work is Rao's measure [16] (Equation 1), in which the diversity of a text document is proportional to the number of dissimilar topics it covers. While we also use Rao's measure, we hypothesize that pure LDA is not good enough for modeling text diversity and propose a re-estimation process for adapting topic models for measuring topical diversity.

Improving the quality of topic models. The two most important issues with topic models are the *generality problem* and the *impurity problem* [5, 12, 18, 20]. Many approaches have been proposed to address the generality problem [20–22]. The main difference with our work is that previous work does not yield sparse topic representations or topic word distributions. Soleimani and Miller [18] propose parsimonious topic models (PTM) to address the generality and impurity problems. PTM achieves state-of-the-art results compared to existing topic models. Unlike [18], we do not modify the training procedure of LDA but propose a method to refine the topic models.

Model parsimonization. In language model parsimonization, the language model of a document is considered to be a mixture of a general background model and a document-specific language model [6, 8, 25]. The goal is to extract the document-specific part and remove the general words. We employ parsimonization for re-estimating topic models. The main assumption in [8] is that the language model of a document is a mixture of its

specific language model and a general language model:

$$P(w\,|\,d) = \lambda P(w\,|\,\tilde{\theta}_d) + (1-\lambda)P(w\,|\,\theta_C), \tag{2}$$

where w is a term, d a document, $\tilde{\theta}_d$ the document specific language model of d, θ_C the language model of the collection C, and λ is a mixing parameter. The main goal is to estimate $P(w\,|\,\tilde{\theta}_d)$ for each document. This is done in an iterative manner using EM algorithm. The initial parameters of the language model are the parameters of standard language model, estimated using maximum likelihood: $P(w\,|\,\tilde{\theta}_d) = \frac{tf_{w,d}}{\sum_{w'} tf_{w',d}}$, where $tf_{w,d}$ is the frequency of w in d. The following steps are computed iteratively:
E-step:

$$e_w = tf_{w,d} \cdot \frac{\lambda P(w\,|\,\tilde{\theta}_d)}{\lambda P(w\,|\,\tilde{\theta}_d) + (1-\lambda)P(w\,|\,\theta_C))}, \tag{3}$$

M-step:

$$P(w\,|\,\tilde{\theta}_d) = \frac{e_w}{\sum_{w'} e_{w'}}, \tag{4}$$

where $\tilde{\theta}_d$ is the parsimonized language model of document d, C is the background collection, $P(w\,|\,\theta_C)$ is estimated using maximum likelihood estimation, and λ is a parameter that controls the level of parsimonization. A low value of λ will result in a more parsimonized model while $\lambda = 1$ yields a model without parsimonization. The EM process stops after a fixed number of iterations or after convergence.

Evaluating topic models. We evaluate the effectiveness of our re-estimated models by measuring the topical diversity of text documents. In addition, in Section 6, we analyze the effectiveness of our re-estimation approach in terms of purity in document clustering and document classification tasks. For classification, following [9, 15, 18], we model topics as document features with values $P(t\,|\,d)$. For clustering, each topic is considered a cluster and each document is assigned to its most probable topic [15, 23, 24].

3 Measuring topical diversity of documents

To measure topical diversity of text documents, we propose HiTR (hierarchical topic model re-estimation). HiTR can be applied to any topic modeling approach that models documents as distributions over topics and topics as distributions over words.

The input to HiTR is a corpus of text documents. The output is a probability distribution over topics for each document in the corpus. HiTR has three levels of re-estimation: (1) **document re-estimation (DR)** re-estimates the language model per document $P(w\,|\,d)$; (2) **topic re-estimation (TR)** re-estimates the language model per topic $P(w\,|\,t)$; and (3) **topic assignment re-estimation (TAR)** re-estimates the distribution over topics per document $P(t\,|\,d)$. Based on applying or not applying re-estimation at different levels, there are seven possible re-estimation approaches; see Fig. 1. HiTR refers to the model that uses all three re-estimation techniques, i.e., TM+DR+TR+TAR. Next, we describe each of the re-estimation steps in more detail.

Document re-estimation (DR) re-estimates $P(w\,|\,d)$. Here, we remove unnecessary information from documents before training topic models. This is comparable to pre-processing steps, such as removing stopwords and high- and low-frequency words, that are typically carried out prior to applying topic models [4, 10, 14, 15].

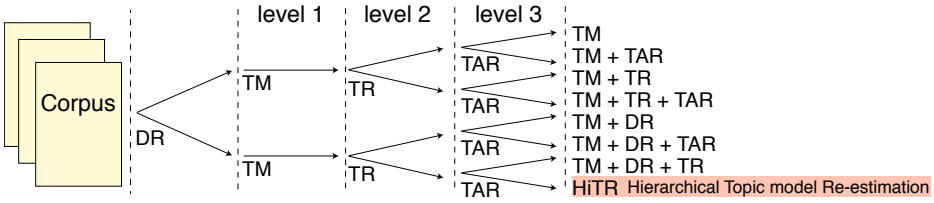

Fig. 1: Different topic re-estimation approaches. TM is a topic modeling approach like, e.g., LDA. DR is document re-estimation, TR is topic re-estimation, and TAR is topic assignment re-estimation.

Proper pre-processing of documents, however, takes lots of effort and involves tuning many parameters. *Document re-estimation*, however, removes impure elements (general words) from documents automatically. If general words are absent from documents, we expect that the trained topic models will not contain general topics. After document re-estimation, we can train any standard topic model on the re-estimated documents.

Document re-estimation uses the parsimonization method described in §2. The re-estimated model $P(w\,|\,\tilde{\theta}_d)$ in (4) is used as the language model of document d, and after removing unnecessary words from d, the frequencies of the remaining words (words with $P(w\,|\,\tilde{\theta}_d) > 0$) are re-estimated for d using the following equation:

$$tf(w,d) = \Big\lfloor P(w\,|\,\tilde{\theta}_d) \cdot |d| \Big\rfloor,$$

where $|d|$ is the document length in words. Topic modeling is then applied on the re-estimated document-word frequency matrix.

Topic re-estimation (TR) re-estimates $P(w\,|\,t)$ by removing general words. The re-estimated distributions are used to assign topics to documents. The goal of this step is to increase the purity of topics by removing general words that have not yet been removed by DR. The two main advantages of the increased purity of topics are(1) it improves human interpretation of topics, and (2) it leads to more document-specific topic assignments, which is essential for measuring topical diversity of documents.

Our main assumption is that each topic's language model is a mixture of its topic-specific language model and the language model of the background collection. TR extracts a topic-specific language model for each topic and removes the part that can be explained by the background model. We initialize $\tilde{\theta}_t$ and θ_T as follows:

$$P(w\,|\,\tilde{\theta}_t) = P(w\,|\,\theta_t^{\mathcal{TM}}) \qquad P(w\,|\,\theta_T) = \frac{\sum_{t\in T} P(w\,|\,\theta_t^{\mathcal{TM}})}{\sum_{w'\in V}\sum_{t'\in T} P(w'\,|\,\theta_{t'}^{\mathcal{TM}})}$$

where t is a topic, $\tilde{\theta}_t$ is topic-specific language model of t, and θ_T is the background language model of T (the collection of all topics), $P(w\,|\,\theta_t^{\mathcal{TM}})$ is the probability of w belonging to topic t estimated by a topic model $\mathcal{TM}$. Having these estimations, the steps of TR are similar to the steps of parsimonization, except that in the E-step we estimate $tf_{w,t}$, the frequency of w in t, by $P(w\,|\,\theta_t^{\mathcal{TM}})$.

Topic assignment re-estimation (TAR) re-estimates $P(t\,|\,d)$. In topic modeling, most topics are usually assigned with a non-zero probability to most of documents. For

documents which are in reality about a few topics, this topic assignment is incorrect and overestimates its diversity. TAR addresses the general topics problem and achieves more document specific topic assignments. To re-estimate topic assignments, a topic model is first trained on the document collection. This model is used to assign topics to documents based on the proportion of words they have in common. We then model the distribution over topics per document as a mixture of its document-specific topic distribution and the topic distribution of the entire collection.

We initialize $P(t|\tilde{\theta}_d)$ and $P(t|\theta_C)$ as follows:

$$P(t|\tilde{\theta}_d) = P(t|\theta_d^{\mathcal{TM}}) \qquad P(t|\theta_C) = \frac{\sum_{d \in C} P(t|\theta_d^{\mathcal{TM}})}{\sum_{t' \in T} \sum_{d' \in C} P(t'|\theta_{d'}^{\mathcal{TM}})}.$$

Here, t is a topic, d a document, $P(t|\tilde{\theta}_d)$ the document-specific topic distribution, and $P(t|\theta_C)$ the distribution of topics in the entire collection C, and $P(t|\theta_d^{\mathcal{TM}})$ the probability of assigning topic t to document d estimated by a topic model $\mathcal{TM}$. The remaining steps of TAR follow the ones of parsimonization, the difference being that in the E-step, we estimate $f_{t,d}$ using $P(t|\theta_d^{\mathcal{TM}})$.

4 Experimental setup

Our main research question is: (RQ1) How effective is HiTR in measuring topical diversity of documents? How does it compare to the state-of-the-art in addressing the general and impure topics problem?

To address RQ1 we run our models on a binary classification task. We generate a synthetic dataset of documents with high and low topical diversity (the process is detailed below), and the task for every model is to predict whether a document belongs to the high or low diversity class. We employ HiTR to re-estimate topic models and use the re-estimated models for measuring topical diversity of documents. To gain deeper insights into how HiTR performs, we conduct a separate analysis of the last two levels of re-estimation, TR and TAR:[1] (RQ2.1) Does TR increase the purity of topics? If so, how does using the more pure topics influence the performance in topical diversity task? (RQ2.2) How does TAR affect the sparsity of document-topic assignments? And what is the effect of re-estimated document-topic assignments on the topical diversity task? To answer RQ2.1, we first evaluate the performance of TR on the topical diversity task and compare its performance to DR and TAR. To answer RQ2.2, we first evaluate TAR together with LDA in a topical diversity task and analyze its effect on the performance of LDA to study how successful TAR is in removing general topics from documents.

Dataset, pre-processing, evaluation metrics, and parameters: Following [3], we generate 500 documents with a high value of diversity and 500 documents with a low value of diversity. We select over 300,000 documents articles published between 2012 to 2015 from PubMed [1]. For generating documents with a high value of diversity, we first select 20 journals and create 10 pairs of journals. Each pair contains two journals that are relatively unrelated to each other (we use the pairs of journals selected in [3]). For each pair of journals A and B we select 50 articles to create 50 probability distributions over

[1] As the DR level of re-estimation directly employs the parsimonious language modeling techniques in [8], we omit it from our in-depth analysis.

topics: we randomly select one article from A and one from B and generate a document by averaging the selected article's bag of topic counts. Thus, for each pair of journals we generate 50 documents with a high diversity value. Also, for each of the chosen 20 journals, we repeat the procedure but instead of choosing articles from different journals, we select them from the same journal to generate 25 non-diverse documents.

For pre-processing documents, we remove stopwords included in the standard stop word list from Python's NLTK package. In addition, we remove the 100 most frequent words in the collection and words with fewer than 5 occurrences.

Measuring topical diversity: After re-estimating word distributions in documents, topics, and document topic distributions using HiTR, we use the final distributions over topics per document for measuring topical diversity. Diversity of texts is computed using Rao's coefficient [3] using Equation 1. We use the normalized angular distance δ for measuring the distance between topics, since it is a proper distance function [2].

To measure the performance of topic models on the topical diversity task, we use ROC curves and report the AUC values [3]. We also measure the *coherence* of the extracted topics; this measure indicates the purity of $P(w|t)$ distributions, where a high value of coherence implies high purity within topics. We estimate coherence using *normalized pointwise mutual information* between the top N words within a topic [10, 15]. As the reference corpus for computing word occurrences, we use the English Wikipedia.[2]

The topic modeling approach used in our experiments with HiTR is LDA. Following [3, 17, 18] we set the number of topics to 100. We set the two hyperparameters to $\alpha = 1/T$ and $\beta = 0.01$, where T is the number of topics, following [15]. In the re-estimation process, at each step of the EM algorithm, we set the threshold for removing unnecessary components from the model to 0.0001 and remove terms with an estimated probability less than this threshold from the language models, as in [8].

We perform 10-fold cross validation, using 8 folds as training data, 1 fold to tune the parameters (λ for DR, TR, and TAR), and 1 fold for testing. Our baseline for the topical diversity task is the method proposed in [3], which uses LDA. We also compare our results to PTM [18], which we use instead of LDA for measuring topical diversity. PTM is the best available topic modeling approach, and the current state of the art.

For statistical significance testing, we compare our methods to PTM using paired two-tailed t-tests with Bonferroni correction. To account for multiple testing, we consider an improvement significant if: $p \leq \alpha/m$, where m is the number of conducted comparisons and α is the desired significance. We set $\alpha = 0.05$. In §5, $^{\blacktriangle}$ and $^{\blacktriangledown}$ indicate that the corresponding method performs significantly better and worse than PTM, respectively.

5 Results

In this section, we report on the performance of HiTR on the topical diversity task. Additionally we analyze the effectiveness of the individual re-estimation approaches.

5.1 Topical diversity results Fig. 2 plots the performance of our topic models across different levels of re-estimation, and the models we compare to, on the PubMed dataset. We plot ROC curves and compute AUC values. To plot the ROC curves we use the diversity scores calculated for the generated pseudo-documents with diversity labels. HiTR improves the performance of LDA by 17% and PTM by 5% in terms of AUC. From Fig. 2 two observations can be made.

[2] We use a dump of June 2, 2015, containing 15.6 million articles.

First, HiTR benefits from the three re-estimation approaches it encapsulates by successfully improving the quality of estimated diversity scores. Second, the performance of LDA+TAR, which tries to address the generality problem, is higher than the performance of LDA+TR, which addresses impurity. General topics have a stronger negative effect on measuring topical diversity than impure topics. Also, LDA+DR outperforms LDA+TR. So, removing impurity from $P(t\,|\,d)$ distributions is the most effective approach in the topical diversity task, and removing impurity from $P(w\,|\,d)$ distributions is more effective than removing impurity from $P(w\,|\,t)$ distributions. Table 1 illustrates the difference between LDA and HiTR with the topics assigned by the two methods for a non-diverse document that is combined from two documents from the same journal, entitled "Molecular Neuroscience: Challenges Ahead" and "Reward Networks in the Brain as Captured by Connectivity Measures," using the procedure described in §4. As only a very basic stopword list being applied, words like *also* and *one* still appear. We expect to have a low diversity value for the combined document. However, using Rao's diversity measure, the topical diversity of this document based on the LDA topics is 0.97. This is due to the fact that there are three document-specific topics—topics 1, 2 and 4—and four general topics. Topics 1 and 2 are very similar and their δ is 0.13. The other, more general topics have high δ values; the average δ value between pairs of topics is as high as 0.38. For the same document, HiTR only assigns three document-specific topics and they are more pure and coherent. The average δ value between pairs of topics assigned by HiTR is 0.19. The diversity value of this document using HiTR is 0.16, which indicates that this document is non-diverse. Hence, HiTR is more effective than other approaches in measuring topical diversity of documents; it successfully removes generality from $P(t\,|\,d)$.

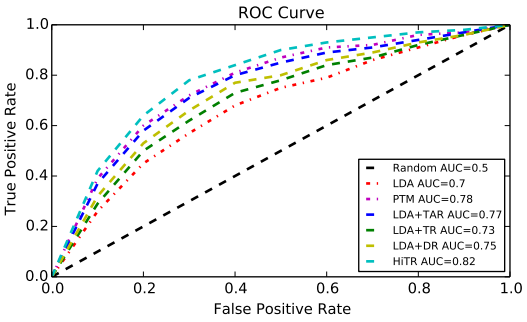

Fig. 2: Performance of topic models in topical diversity task on the PubMed dataset. The improvement of HiTR over PTM is statistically significant ($p < 0.05$) in terms of AUC.

5.2 Topic re-estimation results To answer **RQ2.1**, we focus on topic re-estimation (TR). Since TR tries to remove impurity from topics, we expect it to increase the coherence of the topics by removing unnecessary words from topics. We measure the

Table 1: Topic assignments for a non-diverse document using LDA and HiTR. Only topics with $P(t\,|\,d) > 0.05$ are shown.

	LDA		HiTR	
Topic	$P(t\,\|\,d)$	Top 5 words	$P(t\,\|\,d)$	Top 5 words
1	0.21	brain, anterior, neurons, cortex, neuronal	0.68	brain, neuronal, neurons, neurological, nerve
2	0.14	channel, neuron, membrane, receptor, current	0.23	channel, synaptic, neuron, receptor, membrane
3	0.10	use, information, also, new, one	0.09	network, nodes, cluster, community, interaction
4	0.08	network, nodes, cluster, functional, node		
5	0.08	using, method, used, image, algorithm		
6	0.08	time, study, days, period, baseline		
7	0.07	data, values, number, average, used		

Table 2: Topic model coherence in terms of average normalized mutual information between top 10 words in the topics on the PubMed dataset.

LDA	PTM	LDA+TR	LDA+DR+TR
8.17	9.89	9.46	10.29▲

purity of topics based on the coherence of words in $P(w\,|\,t)$ distributions. Table 2 shows the coherence of topics according to different topic modeling approaches, in terms of average mutual information. TR significantly increases the coherence of topics by removing the impure parts from topics. The coherence of PTM is higher than of TR. However, when we first apply DR, train LDA, and finally apply TR, the coherence of the extracted topics is significantly higher than the coherence of topics extracted by PTM. We conclude that TR is effective in removing impurity from topics. Moreover, DR also contributes in making topics more pure.

5.3 Topic assignment re-estimation results To answer **RQ2.2**, we focus on TAR (topic assignment re-estimation). We are interested in seeing how HiTR deals with general topics. We sum the probability of assigning a topic to a document, over all documents: for each topic t, we compute $\sum_{d\in C} P(t\,|\,d)$, where C is the document collection. Fig. 3 shows the distribution of probability mass before and after applying TAR; topics are sorted based on the topic assignment probability of LDA. LDA assigns a vast proportion of the probability mass to a relatively small number of topics, mostly general topics that are assigned to most documents. We expect that many topics are represented in some documents, while relatively few topics will be relevant to all documents. After applying TAR, the distribution is less skewed and the probability mass is more evenly distributed.

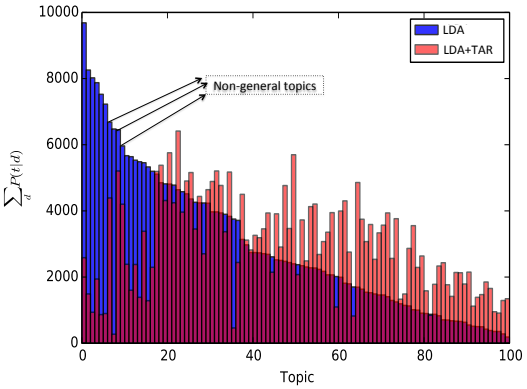

Fig. 3: The total probability of assigning topics to the documents in the PubMed dataset estimated using LDA and LDA+TAR. (The two areas are equal to the number of documents ($N \approx 300K$)).

There are topics that have a high $\sum_d P(t\,|\,d)$ value in LDA's topic assignments and a high $\sum_d P(t\,|\,d)$ value after applying TAR too; we marked them as "non-general topics" in Fig. 3. Table 3, column 2 shows the top five words for these topics. TAR is able to find these three non-general topics and their assignment probabilities to documents in the $P(t\,|\,d)$ distributions is not changed as much as the actual general topics. Thus, TAR removes general topics from documents and increases the probability of document-specific topics for each document. To further investigate whether TAR really removes general topics, Table 3, column 3 shows the top five words for the first 10 topics in Fig. 3, excluding the "non-general topics." These seven topics have the highest decrease

Table 3: Top five words for the topics detected by TAR as general topics and non-general topics.

Topic	Non-general topics	General topics
1	health, services, public, countries, data	use, information, also, new, one
2	surgery, surgical, postoperative, patient, performed	ci, study, analysis, data, variables
3	cells, cell, treatment, experiments, used	time, study, days, period, baseline
4		group, control, significantly, compared, groups
5		study, group, subject, groups, significant
6		may, also, effects, however, would
7		data, values, number, average, used

in $\sum_d P(t\,|\,d)$ values due to TAR. Clearly, they contain general words and are not informative. Fig. 3 shows that after applying TAR, the $\sum_d P(t\,|\,d)$ values have decreased dramatically for these topics, without creating new general topics.

5.4 Parameter analysis Next, we analyze the effect of the λ parameter on the performance of DR, TR, and TAR. Fig. 4 displays the performance at different levels of re-estimation. With $\lambda = 1$, no re-estimation occurs, and all methods equal LDA. We see that DR peaks with moderate values of λ ($0.4 \leq \lambda \leq 0.45$). This reflects that documents contain a moderate amount of general information and that DR is able to successfully deal with it. For $\lambda \geq 0.8$, the performance of DR and LDA is the same and for these values of λ DR does not increase the quality of LDA. Also, the best performance of TR is achieved with high values of λ ($0.65 \leq \lambda \leq 0.75$). From this observation we conclude that topics typically need only a small amount of re-estimation. With this slight re-estimation, TR is able to improve the quality of LDA. However, for $\lambda \geq 0.75$ the accuracy of TR degrades. Lastly, TAR achieves its best performance with low values of λ ($0.02 \leq \lambda \leq 0.05$). Hence, most of the noise is in the $P(t\,|\,d)$ distributions and aggressive re-estimation allows TAR to remove most of it.

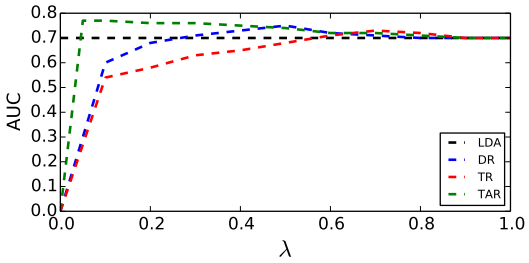

Fig. 4: The effect of the λ parameter on the performance of topics models in the topical diversity task on the PubMed dataset.

6 Analysis

In this section, we want to gain additional insights into HiTR and its effects on topic computation. The purity of topic assignments based on $P(t\,|\,d)$ distributions has the highest effect on the quality of estimated diversity scores. Thus, we investigate how pure estimated topic assignments are using HiTR. To this end, we compare document clustering and classification results, based on the topics assigned by HiTR, LDA and PTM. For clustering, following [15], we consider each topic as a cluster. Each document d is assigned to the topic that has the highest probability value in $P(t\,|\,d)$. For classification, we use all topics assigned to the document and consider $P(t\,|\,d)$ as features for a supervised classification algorithm; we use SVM. We view high accuracy in clustering

or classification as an indicator of high purity of topic distributions. Our focus is not on achieving a top clustering or classification performance: these tasks are a means to assess the purity of topic distributions using different topic models.

Datasets and metrics. We use RCV1 [11], 20-NewsGroups,[3] and Ohsumed.[4] RCV1 contains 806,791 documents with category labels for 126 categories. For clustering and classification of documents, we use 55 categories in the second level of the hierarchy. 20-NewsGroups contains ∼20,000 documents (20 categories, around 1,000 documents per category). Ohsumed contains 50,216 documents grouped into 23 categories. For measuring the purity of clusters we use *purity* and *normalized mutual information* (NMI) [13]. We use 10-fold cross validation and the same pre-processing as in §4.

Purity results. The top part of Table 4 shows results on the document clustering task. As we can see, the topic distributions extracted using HiTR score higher than the ones extracted using LDA and PTM in terms of both purity and NMI. This shows the ability of HiTR to make $P(t\,|\,d)$ more pure. The two-level re-estimated topic models achieve higher purity values than their respective one-level counterparts except the combination of DR and TR, which indicates that re-estimation at each level contributes to the purity of $P(t\,|\,d)$. The combination of TR and DR is not effective in increasing purity over its one-level counterparts on most of the datasets, indicating that TR and DR address similar issues. But when each of them is combined with TAR, the purity of the topic distributions increases, implying that DR/TR and TAR address complementary issues.

The bottom part of Table 4 shows results on the document classification task. HiTR is more accurate in estimating $P(t\,|\,d)$; its accuracy is higher than that of other topic models. The higher values in classification task, compared to clustering task, indicate that the most probable topic does not necessarily contain all information about the content of a document. If a document is about more than one topic, the classifier utilizes all $P(t\,|\,d)$ information and performs better. Therefore, the higher accuracy of HiTR in this task is an indicator of its ability to assign document-specific topics to documents.

7 Conclusions

We have proposed HiTR, an approach for measuring topical diversity of text documents. It addresses two main issues with topic models, topic generality and topic impurity, which negatively affect measuring topical diversity scores in three ways. First, the existence of document-unspecific words within $P(w\,|\,d)$ (the distribution of words within documents) yields general topics and impure topics. Second, the existence of topic-unspecific words within $P(w\,|\,t)$ (the distribution of words within topics) yields impure topics. Third, the existence of document-unspecific topics within $P(t\,|\,d)$ (the distribution of topics within documents) yields general topics. We have proposed three approaches for removing unnecessary or even harmful information from probability distributions, which we combine in our method for Hierarchical Topic model Re-estimation (HiTR).

Estimated diversity scores for documents using HiTR are more accurate than those obtained using the current state-of-the-art topic modeling method PTM, or a general purpose topic model such as LDA. HiTR outperforms PTM because it adapts topic models for the topical diversity task. The quality of topic models for measuring topical

[3] Available at http://www.ai.mit.edu/people/~jrennie/20Newsgroups/

[4] Available at http://disi.unitn.it/moschitti/corpora.htm

Table 4: Re-estimated topic models for document clustering (top) and document classification (bottom). For significance tests, we consider p-value $< 0.05/7$; comparisons are against PTM.

	RCV1		20-Newsgroups		Ohsumed	
Method	Purity	NMI	Purity	NMI	Purity	NMI
LDA	0.55	0.40	0.52	0.36	0.50	0.30
PTM	0.61	0.43	0.57	0.38	0.55	0.33
LDA+DR	0.57$^{\blacktriangledown}$	0.41$^{\blacktriangledown}$	0.56	0.39	0.53$^{\blacktriangledown}$	0.32$^{\blacktriangledown}$
LDA+TR	0.57$^{\blacktriangledown}$	0.42$^{\blacktriangledown}$	0.56	0.38	0.53$^{\blacktriangledown}$	0.31$^{\blacktriangledown}$
LDA+TAR	0.60	0.43	0.57	0.39	0.54	0.33
LDA+DR+TR	0.58	0.42$^{\blacktriangledown}$	0.57	0.38	0.54	0.32
LDA+DR+TAR	0.60	0.43	0.58	0.40	0.55	0.35$^{\blacktriangle}$
LDA+TR+TAR	0.61	0.43	0.58	0.40$^{\blacktriangle}$	0.56$^{\blacktriangle}$	0.34$^{\blacktriangle}$
HiTR	**0.64**$^{\blacktriangle}$	**0.45**$^{\blacktriangle}$	**0.60**$^{\blacktriangle}$	**0.42**$^{\blacktriangle}$	**0.57**$^{\blacktriangle}$	**0.35**$^{\blacktriangle}$
	Acc.	Change	Acc.	Change	Acc.	Change
LDA	0.76	-8%	0.81	-7%	0.50	-11%
PTM	0.82	–	0.87	–	0.56	–
LDA+DR	0.79$^{\blacktriangledown}$		0.83$^{\blacktriangledown}$	-5%	0.52$^{\blacktriangledown}$	-7%
LDA+TR	0.78$^{\blacktriangledown}$	-5%	0.83$^{\blacktriangledown}$	-5%	0.53$^{\blacktriangledown}$	-5%
LDA+TAR	0.82	0%	0.85$^{\blacktriangledown}$	-2%	0.54	-4%
LDA+DR+TR	0.80$^{\blacktriangledown}$	-2%	0.84$^{\blacktriangledown}$	-3%	0.53$^{\blacktriangledown}$	-5%
LDA+DR+TAR	0.83	+1%	0.86	-1%	0.56	0%
LDA+TR+TAR	0.82$^{\blacktriangle}$	0%	0.87	0%	0.58$^{\blacktriangle}$	+4%
HiTR	**0.85**$^{\blacktriangle}$	+4%	**0.89**$^{\blacktriangle}$	+2%	**0.60**$^{\blacktriangle}$	+7%

diversity degrades mainly because of general topics in the $P(t|d)$ distributions. Our topic assignment re-estimation (TAR) approach successfully removes general topics, leading to higher performance on the topical diversity task.

We analyzed the purity of topic assignments on clustering and classification tasks, where $P(t|d)$ distributions were directly used as features. The results confirm that HiTR is effective in removing impurity from documents; it removes impure parts from the three probability distributions mentioned, using three re-estimation approaches.

Acknowledgments. This research was supported by Ahold Delhaize, Amsterdam Data Science, Blendle, the Bloomberg Research Grant program, the Dutch national program COMMIT, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 283465 (ENVRI) and 312827 (VOX-Pol), the Microsoft Research Ph.D. program, the Netherlands eScience Center under project number 027.012.105, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs 314.99.108, 600.006.014, HOR-11-10, CI-14-25, 652.002.001, 612.001.551, 652.001.003, 314-98-071, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

8 References

[1] National Center for Biotechnology Information, U.S. National Library of Medicine. Pubmed Central Open Access Initiative. 2010.

[2] H. Azarbonyad, F. Saan, M. Dehghani, M. Marx, and J. Kamps. Are topically diverse documents also interesting? In *CLEF*, 2015.

[3] K. Bache, D. Newman, and P. Smyth. Text-based measures of document diversity. In *KDD*, 2013.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3 (4–5):993–1022, 2003.

[5] J. Boyd-Gaber, D. Mimno, and D. Newman. Care and feeding of topic models. In *Mixed Membership Models & Their Applic.* CRC Press, 2014.

[6] M. Dehghani, H. Azarbonyad, J. Kamps, and M. Marx. On horizontal and vertical separation in hierarchical text classification. In *ICTIR*, 2016.

[7] M. Derzinski and K. Rohanimanesh. An information theoretic approach to quantifying text interestingness. In *NIPS MLNLP workshop*, 2014.

[8] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *SIGIR*, 2004.

[9] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2009.

[10] J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, 2014.

[11] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: a new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, 2004.

[12] T. Lin, W. Tian, Q. Mei, and H. Cheng. The dual-sparse topic model: Mining focused topics and focused terms in short text. In *WWW*, 2014.

[13] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[14] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR*, 2013.

[15] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson. Improving topic models with latent feature word representations. *Trans. Assoc. Comput. Linguistics*, 3:299–313, 2015.

[16] C. Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982.

[17] M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *WSDM*, 2015.

[18] H. Soleimani and D. Miller. Parsimonious topic models with salient word discovery. *IEEE Trans. on Knowl. and Data Eng.*, 27(3):824–837, 2015.

[19] A. Solow, S. Polasky, and J. Broadus. On the measurement of biological diversity. *Journal of Environmental Economics and Management*, 24(1):60–68, 1993.

[20] H. M. Wallach, D. M. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In *NIPS*, 2009.

[21] C. Wang and D. M. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *NIPS*, 2009.

[22] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010.

[23] P. Xie and E. P. Xing. Integrating document clustering and topic modeling. In *UAI*, 2013.

[24] X. Yan, J. Guo, Y. Lan, and X. Cheng. A biterm topic model for short texts. In *WWW*, 2013.

[25] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM*, 2001.