Preprint from https://gipp.com/pub/

M. Ostendorff, T. Blume, T. Ruas, B. Gipp, and G. Rehm, "Specialized Document Embeddings for Aspect-based Similarity of Research Papers" in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2022.

Specialized Document Embeddings for Aspect-based Similarity of Research Papers

Malte Ostendorff
malte.ostendorff@dfki.de
DFKI GmbH
Berlin, Germany

Till Blume
till.blume@de.ey.com
Ernst & Young GmbH WPG – R&D
Berlin, Germany

Terry Ruas
ruas@uni-wuppertal.de
University of Wuppertal
Wuppertal, Germany

Bela Gipp
gipp@cs.uni-goettingen.de
University of Göttingen
Göttingen, Germany

Georg Rehm
georg.rehm@dfki.de
DFKI GmbH
Berlin, Germany

ABSTRACT

Document embeddings and similarity measures underpin content-based recommender systems, whereby a document is commonly represented as a single generic embedding. However, similarity computed on single vector representations provides only one perspective on document similarity that ignores which aspects make two documents alike. To address this limitation, aspect-based similarity measures have been developed using document segmentation or pairwise multi-class document classification. While segmentation harms the document coherence, the pairwise classification approach scales poorly to large scale corpora. In this paper, we treat aspect-based similarity as a classical vector similarity problem in aspect-specific embedding spaces. We represent a document not as a single generic embedding but as multiple specialized embeddings. Our approach avoids document segmentation and scales linearly w.r.t. the corpus size. In an empirical study, we use the Papers with Code corpus containing 157,606 research papers and consider the *task*, *method*, and *dataset* of the respective research papers as their aspects. We compare and analyze three generic document embeddings, six specialized document embeddings and a pairwise classification baseline in the context of research paper recommendations. As generic document embeddings, we consider FastText, SciBERT, and SPECTER. To compute the specialized document embeddings, we compare three alternative methods inspired by retrofitting, fine-tuning, and Siamese networks. In our experiments, Siamese SciBERT achieved the highest scores. Additional analyses indicate an implicit bias of the generic document embeddings towards the *dataset* aspect and against the *method* aspect of each research paper. Our approach of aspect-based document embeddings mitigates potential risks arising from implicit biases by making them explicit. This can, for example, be used for more diverse and explainable recommendations.

JCDL '22, June 20–24, 2022, Cologne, Germany

ACM ISBN 978-1-4503-9345-4/22/06…$15.00
https://doi.org/10.1145/3529372.3530912

CCS CONCEPTS

• **Information systems** → *Recommender systems*; *Similarity measures*; *Clustering and classification.*

KEYWORDS

Document embeddings, Document similarity, Content-based recommender systems, Papers With Code, Aspect-based Similarity

ACM Reference Format:
Malte Ostendorff, Till Blume, Terry Ruas, Bela Gipp, and Georg Rehm. 2022. Specialized Document Embeddings for Aspect-based Similarity of Research Papers. In *The ACM/IEEE Joint Conference on Digital Libraries in 2022 (JCDL '22), June 20–24, 2022, Cologne, Germany.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3529372.3530912

1 INTRODUCTION

In content-based recommender systems and other information retrieval applications, the retrieval of semantically similar documents is often performed based on document embeddings that can be derived from the text [14, 38], citations or links [24, 67], and combinations of text and citations [12, 51]. The similarity between documents is then calculated based on the similarity of their vector representations, e. g., with cosine similarity [16, 63]. Existing approaches represent a document with a single vector in the embedding space. This leads to a single notion of document similarity which neglects the many meanings represented within a document, e. g., different arguments or sub-topics. In the context of word embeddings, Camacho-Collados and Pilehvar [8] define "the inability to discriminate among different meanings of a word" as the meaning conflation deficiency. While the appearance of contextualized word embeddings has solved the meaning conflation for words [54, 69], document embeddings still suffer from this issue.

The coarse-grained similarity assessment (similar or not) neglects the many aspects in which two documents are related. Goodman [21] and Bär et al. [4] argue the concept of similarity is an ill-defined notion unless one can say what aspects are being considered to bind the compared items. In scientific recommender systems, the similarity is often concerned with multiple facets of the presented research, e. g., methods or findings [9]. Addressing these facets individually could help tailoring recommendations for specific information needs and increasing their diversity [18, 36]. Especially in the scientific domain, this could help bursting filter bubbles or facilitating new discoveries [48, 57].

Existing approaches derive aspect-based document similarity by splitting documents into aspect-specific segments and computing a segment-level similarity [9, 28, 34]. Since segmentation breaks the document coherence, our prior work [52] proposes to keep documents intact and to incorporate aspect information into similarity through a pairwise document classification task. In the prior work, we perform a pairwise multi-class classification task whereby aspects in two documents are represented with a single class label. Pairwise document classification has been successfully demonstrated for Wikipedia articles [53] and research papers [52]. However, with $O(n^2)$ comparisons for a corpus of n documents, the pairwise multi-class classification approach scales poorly to large scale corpora. A quadratic complexity requires extensive computation resources, in particular in combination with other computational expensive methods, e. g., large Transformer language models [69].

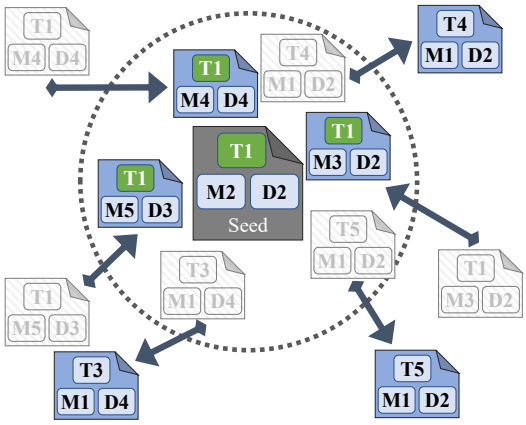

Figure 1: Papers are associated with tasks (T), methods (M), and datasets (D). With generic embeddings (gray), the k-nearest neighbors are papers similar in any aspect. Specializing the embeddings (blue) for the *task* aspect (arrows) lets papers with the same task (T1, green) be close to each other in the embedding space.

In this paper, we present a new approach for aspect-based document similarity. We propose to represent a document using multiple specialized embeddings – one embedding for each aspect. We construct an aspect-specific embedding space for each aspect. Thus, we are able to capture the similarity of documents regarding different aspects. We build upon the idea of specialization (sometimes referred to as retrofitting) of word embeddings [17, 19]. The specialization models leverage external lexical knowledge to specialize word embedding spaces for particular constraints, e. g., vectors of synonyms are close to each other. The use of multi-sense embeddings to better represent the different meaning of words is known to improve natural language understanding related tasks [40, 55, 61, 62, 73]. We apply the idea of specialization on documents and for each aspect-specific embedding space. Our goal is to leverage aspect information such that documents similar in a particular aspect are close to each other in the embedding space for that aspect (Figure 1). Thus, we refer to these embeddings as *specialized* for a specific aspect in contrast to *generic* embeddings that only reflect one unspecified aspect or view of a document.

Our approach keeps the documents intact as opposed to segmentation approaches [9, 28, 34] and it addresses the scalability issues of pairwise document classification [52]. The computational expensive encoding of aspect information is only performed once per document and aspect. Retrieving similar documents can be done through a nearest neighbor search in each aspect-specific embedding space. As a result, our approach has linear complexity, i. e., $O(n)$ w.r.t. to n documents in the corpus.

We evaluate our approach of specializing document embeddings on a content-based recommendation task using the Papers with Code[1] corpus. Research papers in Papers with Code are labeled with three aspects: the papers' *task*, the applied *method*, and the *dataset* used. We use these labels as aspects to specialize the embeddings of the research papers. As specialization methods, we rely on existing methods but apply them in a way diverging from their original purpose. Namely, we evaluate retrofitting [19] and jointly learned embeddings from Transformer fine-tuning [5, 12] and Siamese Transformers [60]. The specialized embeddings are compared against a pairwise multi-class document classification baseline and generic (non-specialized) embeddings from FastText word vectors [6], SciBERT [5], and SPECTER [12].

In summary, our contributions are: (1) We propose a new approach to aspect-based document similarity using specialized document embeddings. Opposed to pairwise document classification, we treat aspect-based similarity as a classical vector similarity problem in aspect-specific embedding spaces, which improves the scalability. (2) We empirically evaluate three specialization methods for three aspects on a newly constructed dataset based on Papers with Code for the use case of research paper recommendations. In our experiment, specialized embeddings improved the results in all three aspects, i. e., *task*, *method*, and *dataset*. (3) We find that recommendations solely based on generic embeddings had an implicit bias towards the *dataset* and against the *method* aspect. (4) We demonstrate the practical use of our approach in a prototypical recommender system[2]. (5) We make our code, dataset, and models publicly available[3].

2 RELATED WORK

In the field of information processing, *aspects* appear in various contexts and domains., e. g., sentiment analysis [15, 27, 43, 56], image recommender systems [10, 11], or reviewer matching [33, 45]. In these examples, the goal is to associate aspect information with single items (e. g., products, images) or between items and users (e. g., review matching). Unfortunately, very few works focus on aspect-based similarity of document pairs.

Segmentation. Chan et al. [9] investigate aspect-based recommendations as a segmentation task. They segment the abstracts of collaborative and social computing papers into four classes, depending on their research aspects: background, purpose, mechanism, and findings. Next, they represent a paper with four vectors, each derived from the corresponding segment's content. Computing the

[1] https://paperswithcode.com/
[2] Demo https://hf.co/spaces/malteos/aspect-based-paper-similarity
[3] Repository https://github.com/malteos/aspect-document-embeddings

cosine similarity between the segment vectors allows the retrieval of similar papers for a specific aspect. Huang et al. [28] apply the same segmentation approach but to biomedical research papers. Kobayashi et al. [34] classify sections into discourse facets and build document vectors for each facet. However, splitting documents into segments breaks the document coherence and can hurt the performance of NLP models as Gong et al. [20] showed. The individual segments can retain insufficient context to produce meaningful representations. Therefore, we consider segmentation as a sub-optimal approach for aspect-based similarity.

Pairwise Multi-Class Document Classification. In prior work, we propose to extend document similarity with aspect information using a pairwise multi-class document classification [52, 53]. The prior work evaluates the multi-class document classification approach on Wikipedia articles [53] and research papers [52]. For Wikipedia, the articles are treated as documents and Wikidata properties as labels for aspects describing their similarity [53]. For research papers, we derive aspect labels from citations and the titles of the sections in which the citations are located [52]. Due to the inconsistent use of section titles, the titles prevent a clear distinction among aspects. Unfortunately, no manually curated gold standard is available to date. In both studies, variations of BERT models [5, 14] using a sequence pair classification setting yielded the best results [52, 53]. Despite its good classification performance, pairwise classification with large language models, like BERT, is not suitable for large-scale similarity search applications. Pairwise classification requires passing all possible document pairs through the language model. Thus, this approach has a quadratic complexity, as discussed also by Reimers and Gurevych [60].

Document Embeddings. Various methods exist to encode semantic information of documents into numerical vector representations, commonly know as embeddings. Examples range from Bag-of-Words [25] over TF-IDF [30] to Paragraph Vectors [38]. Also, document embeddings from averaged word embeddings have been shown to be effective [2, 61]. Recently, pretrained language models based on the Transformer architecture [14, 69] have become more popular to generate embeddings based on the document text. But also other semantic information, e. g., citations [24, 51, 67], can be utilized for document embeddings.

Retrofitting. Faruqui et al. [17] show that word embedding learned in unsupervised fashion can be enriched with additional semantic information using retrofitting. Retrofitting is performed in a post-processing step with external knowledge in the form of linguistic resources, such as synonyms and antonyms. Retrofitting minimizes the distance between synonyms vectors and maximizes it between antonyms [19, 47]. Thereby, the multi-senses of words are integrated into their vector representations.

Joint Learning. Similarly, external knowledge can be directly integrated into a representation learning process. Reimers and Gurevych [60] show representations from BERT [14] can be improved with a Siamese architecture [7] when fine-tuned on semantic textual similarity datasets. Other approaches augment pre-trained models (e.g., BART [39], RoBERTa [42]) combining separate trained intermediate tasks and external knowledge sources to solve an additional final task, such as word sense disambiguation [73], paraphrase detection [71, 72], fake news detection [70], and media bias detection [35, 66]. Also, Cohan et al. [12] use citations as a pretraining objective for a scientific BERT language model.

Summary. Even though the mentioned methods provide substantial contributions in document embeddings, they produce generic embeddings that represent a single view of a document's content. This single view prevents to measure the similarity of document embeddings related to aspects. However, our approach aims for aspect-specialized embedding, i. e., for each document and for each of their aspects. Thereby, we address issues from existing approaches for aspect-based document similarity.

3 METHODOLOGY

In the following, we present our approach for aspect-based document similarity and the evaluated embedding methods.

3.1 Approach

Our document embedding specialization approach, illustrated in Figure 1, consists of two major components: (1) aspect information for a defined set of aspects $A = \bigcup_{j=1}^{n} a_j$, and (2) a specialization method that derives for any document d_i in the corpus D a set of n specialized embeddings $\vec{d}_i^{(a_j)}$ for each specific aspect a_j with $1 \leq j \leq n$. The aspect information is given in the form of triples $(d_a, d_b, y^{(a_j)})$ where the label $y^{(a_j)} = \{0, 1\}$ holds the binary information whether d_a and d_b are similar or dissimilar in aspect a_j. The training objective of the specialization method is to maximize the similarity of the embeddings of those document pairs (d_a, d_b) with $y^{(a_j)} = 1$, i. e., that are similar in aspect a_j.

We distinguish between *specialized* embeddings and *generic* embeddings. Generic embeddings can be considered aspect-free, i. e., $\vec{d}_i^{(a_1)} = \vec{d}_i^{(a_2)} = \vec{d}_i^{(a_n)}$. *Specialized* or *generic* similar documents are retrieved through a k-nearest neighbor search using the cosine similarity of the document embeddings. We evaluate our approach in the context of content-based recommender systems. Therefore, we refer to the results of the nearest neighbor search as *specialized* or *generic* recommendations.

With this approach, we treat aspect-based similarity as a classical vector similarity problem in aspect-specific embedding spaces. As a result, similar documents can be more efficiently retrieved as in the pairwise classification approach [52, 53]. Pairwise classification requires the classification of all document pairs, i. e., a corpus with $|D|$ documents is equivalent to $\frac{|D|*(|D|-1)}{2}$ classifications. Thus, the pairwise classification approach has a quadratic complexity, i. e., $O(|D|^2)$ w.r.t. the number of documents $|D|$. This quadratic complexity makes the computation infeasible even for a medium-sized corpus, in particular when Transformers are used for each classification. Our approach computes for each document $d \in D$ and each aspect $a \in A$ one specialized document embedding $\vec{d}^{(a)}$. Consequently, only $|D| * |A|$ Transformer forward-passes are sufficient for inference. Thus, our approach scales linearly w.r.t. the number of documents $|D|$. Retrieving the k most similar documents can be done efficiently in the vector space using cosine similarity [44]. For

larger corpora, approximate nearest neighbor search [3] could be also used.

3.2 Embedding Methods

We evaluate the document embeddings from three base models and three specialization methods. Besides the aspect information (Section 4.2), each method utilizes the title and abstract to generate the embeddings. We distinguish between generic and specialization methods, where the latter is divided into two categories: retrofitted and jointly learned embeddings. Source codes, trained models and instruction to reproduce our work are publicly available[3].

3.2.1 Generic Embeddings. We use *generic* document embeddings that do not leverage any aspect information. As base models, we rely on averaged FastText word vectors as document embeddings [6], document embeddings from SciBERT [5][4], and SPECTER [12]. SPECTER and SciBERT are BERT-inspired models [14] pretrained on scientific literature. In contrast to SciBERT, SPECTER uses citation prediction as an additional pretraining objective. SciBERT and SPECTER are used as published by their authors without any fine-tuning on our corpus and in their `BASE`-version.

3.2.2 Retrofitted Embeddings. Retrofitting refers to the postprocessing of existing embeddings such that they fit predefined constraints [17]. Constrains, e. g., synonyms or antonyms, define which vectors should be close or apart. For our experiments, we retrofit all generic embeddings with Explicit Retrofitting (ER) as proposed by Glavaš and Vulić [19]. In contrast to other retrofitting methods [17], ER generalizes to unseen vectors for which no predefined constraints exist. An ER model can be learned on a subset for which constraints exist (training set) and, then, be applied on all remaining embeddings (test set). The training constraints are the positive samples in the same fashion as the synonyms are used for the retrofitting of words.

3.2.3 Jointly Learned Embeddings. We refer to this category as jointly learned embeddings since aspect information is integrated into the representation learning process. Aspect-based embeddings are directly generated from textual input (title and abstract of a paper). We fine-tune SPECTER and SciBERT in a sequence-pair setup on positive and negative samples from our training set. The input is a pair of two papers separated with a `[SEP]`-token. The sequence pair is subject to a binary classification (similar in aspect or not). To derive embeddings for the test set, we use only a single paper as input to SPECTER and SciBERT. Aside from SPECTER and SciBERT, we also test a Siamese network based on SciBERT (see Sentence-BERT [60]). Siamese-SciBERT uses a Siamese architecture [7], in which the paper pair is separately fed as an input, their representations are concatenated, and then classified.[5]

[4]For SciBERT, we apply mean-pooling, i. e., a document vector is the mean of the hidden-states of the last layer of the SciBERT model. Documents embeddings from the `[CLS]`-token yielded significantly lower results, e. g., 0.001 MAP for the *task* aspect).

[5]For Siamese-SciBERT, we experimented with different loss functions and found the Multiple Negative Ranking Loss [26], with only positive samples from the train set, yielded the best results for our data.

4 EXPERIMENTS

For our experiments, we use the three generic embeddings Avg. FastText, SciBERT, SPECTER (see Section 3.2.1). As specialization methods, we retrofit the three generic embeddings, and also jointly learn specialized embeddings with Transformer fine-tuning and Siamese Transformers (see Section 3.2.2 and 3.2.3). Furthermore, we use the pairwise classification approach as a baseline.

4.1 Corpus

Our approach requires information about aspects that make a document pair similar. To the best of our knowledge, no appropriate dataset for the problem of aspect-based similarity is publicly available as they lack either quantity or quality. Chan et al. [9] provide a dataset that is too small in size for a machine learning approach. In our prior work [52], we rely on citations and section titles as a training signal. However, section titles are inconsistently used and, therefore, prevent a clear distinction among aspects.

Papers with Code hosts a hand-curated collection of research papers in the machine learning domain [31]. In addition to metadata on authors or bibliography, each research paper is labeled with the *task* a paper is focusing on, the papers' *method*, and the *dataset* used. We use these labels as aspects, $A = \{task, method, dataset\}$, as they address different information needs that are beneficial for research paper recommender systems [9]. For example, *Beltagy et al. [5]* and *Cohan et al. [12]* are labeled with *BERT [14]* as their *method*. Thus, we consider the pair of *Beltagy et al. [5]* and *Cohan et al. [12]* as similar regarding the *method* aspect. Other aspect labels are for example:

- **Tasks:** Low-Rank Matrix Completion, Q-Learning, Quantization, Speaker Recognition, Object Detection
- **Methods:** Residual Connection, Tanh Activation, Multi-Head Attention, LSTM, Transformer
- **Datasets:** Atari 2600 Atlantis, Cityscapes, SOP, MS MARCO, Labeled Faces in the Wild

4.2 Ground truth

Table 1: Ground truth for each aspect

Aspect	Papers	Labels	Avg. papers per label
Task	154,350	1,421	17.9
Method	108,687	788	12.4
Dataset	37,604	1,743	5.6

The used Papers with Code corpus contains in total 157,606 unique papers. For each aspect, we construct separated ground truths containing positive and negative samples. Positive samples are unique unordered paper pairs with the same label, i. e., $y = 1$. For each label, the number of pairs is $\binom{L}{2}$ where L is the number of papers per label. Negative samples are randomly sampled paper pairs without the same label, i. e., $y = 0$. The number of negative samples is 50% of the number of positive samples. Some labels are too frequent in the corpus, e. g., the *method* label *Softmax* is assigned to 5,324 papers. To ensure the specificity of aspect information, we discard all labels which are assigned to more than 100 papers. The

removal of too frequent labels increases the task's difficulty and ensures an appropriate dataset size. The dataset would become too large otherwise, e. g., *Softmax* alone would account for 1.2M paper pairs. We conduct our experiments as 4-fold cross-validation and split the data into 75% training and 25% test papers. The resulting ground truth consists of on average of 1,227,058 *task*, 284,193 *method*, and 58,984 *dataset* paper pairs.

4.3 Baseline

To compare our approach with prior work, we use the pairwise multi-class classification approach as a baseline [52]. We train a pairwise classification model based on SPECTER. We selected SPECTER over SciBERT as its generic version outperformed SciBERT. With a document pair as input, the model predicts the probability distribution over the aspect labels. The pairwise approach is not directly applicable on our dataset as its quadratic complexity would require the classification of 1.3 billion document pairs. To reduce the number of candidate pairs, we first retrieve the $n = 300$ nearest neighbors d_n for any seed document d_s based on the generic SPECTER embeddings. The pairs of seed and neighbor documents (d_s, d_n) are selected as candidates for the classifier. This candidate filtering reduces the number of classifications to 11.3 million document pairs.

4.4 Evaluation Methodology

Each of the n aspects is evaluated separately (n train, n test sets). All documents from the test set are used as seeds. For a given aspect a_j and the vector $\vec{d}_s^{(a_j)}$ of seed d_s, we retrieve k candidate documents, with a k nearest neighbor search [13]. The similarity of documents is computed as the cosine similarity of their vectors [63]. The only exception is the pairwise baseline approach, for which the predicted class probabilities are used instead of cosine similarity. A candidate document d_c is relevant for the seed d_s if they are associated with the same label for aspect a_j, i. e., $(d_s, d_c, y^{(a_j)} = 1)$ is part of the ground truth. We compute precision, recall, mean average precision, and mean reciprocal rank based on this relevance definition [44].

5 EXPERIMENTAL EVALUATION

In the following, we present our experimental results. We start with the evaluation of the pairwise approach baseline and continue with the comparison of all aspect-similarity methods, analyze the differences between generic and specialized embeddings, and finally verify our findings with qualitative examples.

5.1 Pairwise Baseline Evaluation

In order to retrieve similar documents with the pairwise approach, we first need to train a classification model that can be separately evaluated on the test set. Table 2 shows the classification performance of Pairwise SPECTER in terms of precision, recall, and F1-score. With a micro F1-score of 0.74, the performance is comparable the previous experiments [52]. A discrepancy can be seen between the aspects. For *task* the F1-scores are the highest with 0.84, followed by *method* with 0.50. The worst performance yields the *dataset* aspect with an F1-score of only 0.16.

Table 2: Classification report for Pairwise SPECTER.

Aspect ↓ / Metric →	Precision	Recall	F1-Score
Task	0.88	0.81	0.84
Method	0.56	0.46	0.50
Dataset	0.11	0.33	0.16
Micro Avg.	0.79	0.74	0.76
Macro Avg.	0.52	0.35	0.50

To make the pairwise approach applicable to our dataset, we introduced an artificial constraint since the prediction for all document pairs is not possible due to the quadratic complexity and limited resources. We retrieve the $n = 300$ nearest neighbors based on generic SPECTER to filter for candidate pairs for that we predict the aspect labels. As this constraint potentially harms the performance, we plot Pairwise SPECTER's performance as MAP@k=10 depending on the size of nearest neighbor filter in Figure 2. The performance generally increases as n increases. However, the larger n the smaller the increase is. Thus, we expect the performance not to increase significantly for large n. The high MAP for the *dataset* aspect and small n is due to the good performance of generic SPECTER for this aspect.

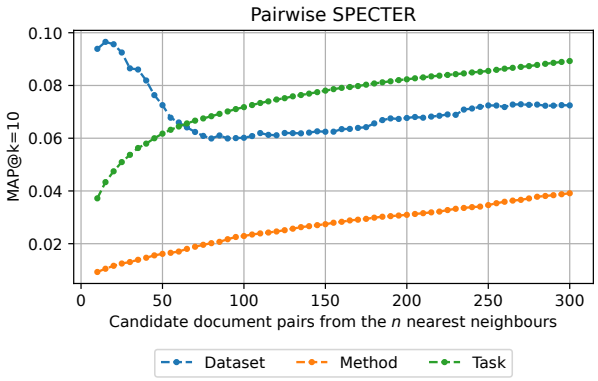

Figure 2: Performance of Pairwise SPECTER in terms MAP@k=10 depending on the candidate filtering for different n nearest neighbors.

5.2 Aspect-based Similarity Evaluation

Table 3 presents the overall results based on the most $k = 10$ similar documents from each method. Results for other k values are depicted in Figure 3. In the following, unless stated otherwise, we refer to the MAP results since it takes the rank of multiple relevant candidates into account.

Siamese-SciBERT is for all metrics and aspects the best method by a large margin. Among the generic embeddings, SPECTER is on average better than Avg. FastText. For *task* and *dataset*, SPECTER outperforms Avg. FastText, while for *method* the opposite is the case. SciBERT yields the lowest scores in the generic category. As Reimers and Gurevych [60] showed, BERT-based embeddings perform poorly without task-specific fine-tuning. Even the computational less complex Avg. FastText outperforms SciBERT. Despite

Table 3: Overall results for the most $k = 10$ similar documents for nine embedding methods and the Pairwise SPECTER baseline. Precision (P), recall (R), mean reciprocal rank (MRR), mean average precision (MAP) are reported as average over a 4-cross-validation. The highest score among aspects in each metric is underlined for the individual method, and bold shows the highest score among methods for a single metric. Fine-tuned Siamese-SciBERT yields the best results.

Aspects →		*Task*				*Method*				*Dataset*			
Methods ↓		**P**	**R**	**MRR**	**MAP**	**P**	**R**	**MRR**	**MAP**	**P**	**R**	**MRR**	**MAP**
Pairwise SPECTER baseline [52]		0.298	0.110	0.545	0.089	0.152	0.048	0.400	0.039	0.124	0.119	0.316	0.072
Generic	Avg. FastText	0.208	0.071	0.419	0.046	0.096	0.029	0.233	0.016	0.170	0.260	0.439	0.152
	SPECTER	0.231	0.080	0.448	0.053	0.077	0.023	0.205	0.012	0.175	0.277	0.446	0.164
	SciBERT	0.083	0.027	0.241	0.015	0.044	0.012	0.142	0.006	0.079	0.112	0.251	0.059
Specialized	Retrofitted Avg. FastText	0.233	0.081	0.445	0.054	0.133	0.040	0.294	0.024	0.202	0.290	0.481	0.174
	Retrofitted SPECTER	0.201	0.071	0.414	0.046	0.067	0.020	0.186	0.010	0.130	0.227	0.364	0.129
	Retrofitted SciBERT	0.106	0.035	0.284	0.019	0.067	0.018	0.189	0.009	0.103	0.140	0.304	0.073
	Fine-tuned SPECTER	0.279	0.095	0.497	0.067	0.063	0.017	0.171	0.010	0.092	0.134	0.279	0.070
	Fine-tuned SciBERT	0.091	0.031	0.258	0.020	0.052	0.013	0.156	0.007	0.070	0.088	0.224	0.045
	Fine-tuned Siamese-SciBERT	**0.569**	**0.242**	**0.708**	**0.224**	**0.407**	**0.168**	**0.588**	**0.137**	**0.270**	**0.374**	**0.533**	**0.235**

requiring the largest computational effort, the Pairwise SPECTER baseline yields only the second-best scores for *task* and *method* while for *datasets* the scores are even the fourth-lowest.

The retrofitting approach [19] has a mixed effect on the performance. For Avg. FastText and SciBERT, the retrofitting increases all scores (on average +26% MAP for Avg. FastText, +34% MAP for SciBERT), while for SPECTER the retrofitting decreases the performance compared to its generic version (on average -16% MAP). The fine-tuning of SPECTER and SciBERT has a different effect depending on the aspects. Compared to its generic counterpart, fine-tuned SPECTER's MAP score is 25% higher for the *task* aspect but 57% lower for the *dataset* aspect. For SciBERT, the fine-tuning also decreases its MAP score by 23% for the *dataset* aspect. Moreover, we do not only see performance differences between the methods but also between the aspects. All methods yield the highest precision for *task*, whereas recall and MAP are the highest for *dataset*. A high MRR can be found for *task* and *dataset*, while the *method* aspect shows the lowest scores throughout all metrics. The poor *method* results can be partially attributed to the unbalanced distribution of the aspects (Section 4.2). Most samples are available for *task*, explaining its good performance compared to *method*. However, *dataset* has the least number of samples but still outperforms *method*. As we specialize the embeddings, we also notice a decrease in performance difference between the aspects. While SPECTER has a high MAP difference from *dataset* to *method* (92%) and from *dataset* to *task* (68%), the same difference is lower for Siamese-SciBERT (42% and 5% respectively). The better the specialization effect the lower is the performance gap the between aspects.

To analyze the aspect-specific performance, Figure 3 depicts the performance ranking as MAP and precision for different k values for Avg. FastText, SPECTER, Retrofitted SPECTER, and Siamese-SciBERT. The performance among the aspect remains stable independent of k for all methods, except Siamese-SciBERT. With Siamese-SciBERT, the *task* aspect yields a higher MAP than *dataset* for $k > 15$. In terms of precision, Siamese-SciBERT is another exception since the precision of *method* is higher than in *dataset*. For all other methds, *method* has the lowest precision.

In summary, Siamese-SciBERT achieves, for all metrics and aspects, the highest scores. Thus, we consider Siamese-SciBERT the best method out of the analyzed methods to handle specialized embeddings even outperforming the Pairwise SPECTER baseline.

5.3 Specialization Evaluation

The performance discrepancy among the aspects could indicate a systematic difference between the documents retrieved through the similarity of generic embeddings and the specialized ones. Therefore, we conduct an additional experiment on their overlap. We use the trained models from Table 3 but infer vectors for all documents in the whole corpus. Then, retrieve $k = 50$ recommendations and count the overlap between each method's nearest neighbors on a seed-level. The large k value is selected to increase the chance of overlapping retrieved documents. Table 4 presents the intersection ratio between the generic retrieved documents from Avg. FastText and SPECTER, and the specialized ones from Siamese-SciBERT. For the remaining methods, we report the intersection in the supplemental materials[3]. The lower the overlap, the more distinct the recommendations are from each other.

On the one hand, most overlaps can be found between Avg. FastText and SPECTER. This suggests little difference within the generic retrieved documents. On the other hand, Siamese-SciBERT's *method*-specific recommendations overlap the least with the generic ones. The discrepancy among the aspects is significant. Compared to SPECTER, Siamese-SciBERT has an overlap of 12%, 5%, and 17% for *task*, *method*, and *dataset* respectively. Thus, indicating *dataset*-specific recommendations are overrepresented in generic recommendations, while *method*-specific ones are underrepresented.

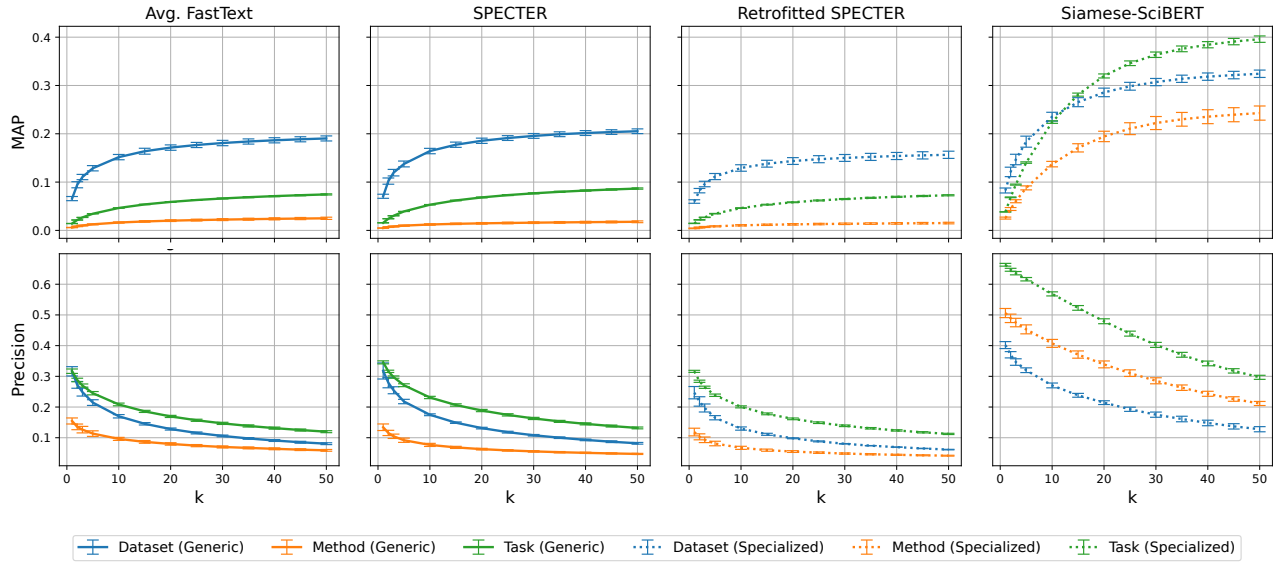

Figure 3: Precision and MAP@k for two generic (Avg. FastText and SPECTER) and two specialized embeddings (Retrofitted SPECTER and Siamese-SciBERT). For generic embeddings, each line presents the scores of the generic method evaluated on different aspect-datasets. For specialized embeddings, a line presents a separately trained model. Generic embeddings and retrofitted SPECTER yield similar results on different k and aspects, while for Siamese-SciBERT, the *task* aspect yields a higher MAP compared to *dataset* for $k > 15$.

Table 4: Intersection of $k = 50$ recommendations from A and B. Most overlap between generic methods (Avg. FastText and SPECTER). Only 5% of Siamese-SciBERT's *method* recommendations also also retrieved by generic methods.

Recommendations A	Recommendations B	A∩B
Avg. FastText	SPECTER	0.29
	Siamese-SciBERT$^{(Task)}$	0.11
	Siamese-SciBERT$^{(Method)}$	0.05
	Siamese-SciBERT$^{(Dataset)}$	0.14
SPECTER	Siamese-SciBert$^{(Task)}$	0.12
	Siamese-SciBert$^{(Method)}$	0.05
	Siamese-SciBert$^{(Dataset)}$	0.17

6 QUALITATIVE VERIFICATION

Considering the quantitative findings, we also qualitatively analyze randomly sampled seed papers and their most similar documents in the context of research paper recommendations. Table 5 presents one of these samples with its top-$k = 3$ recommendations. Generic recommendations are taken from SPECTER and *task*-, *method*-, and *dataset*-specific ones from Siamese-SciBERT. For other examples, we provide a Web-based demo to browse the recommendations for all papers from the dataset[2].

Gupta [22] is the seed paper to which Papers with Code associates three *task* labels (*data augmentation*, *sentiment analysis*, *text generation*), two *method* labels (*convolution* and *generative models (GAN)*), and none *dataset* label. As the labels and the title suggests, *Gupta [22]* uses generative adversarial networks as a data augmentation method to generate textual training data for the sentiment classification task. The four different recommendation sets illustrate the many facets in that papers can be related.

The generic recommendations are all about GAN as an augmentation method. While the first and third recommendations *Karimi et al. [32]* and *Zhang et al. [77]* are both also about sentiment classification, the second *Zhu et al. [78]* investigates emotion classification. Even though sentiment and emotion can be considered as related, the former is based on text and the latter on image data.

All *task*-specific recommendations *Anaby-Tavor et al. [1]*, *Regina et al. [58]*, and *Wu et al. [74]* have data augmentation on text classification as a central theme. However, in contrast to the seed, GANs are not used for augmentation, and the classification task is not concerned with sentiment. The *method*-specific recommendations *Zahan et al. [76]*, *Husmann et al. [29]*, and *Shen et al. [65]* are at first sight unrelated to the seed since they focus on unrelated topics such as hashing or the classification of biomedical or financial data. Nonetheless, the seed and the *method*-specific recommendation all use t-distributed Stochastic Neighbor Embedding (t-SNE) for visualization. Despite of being different in central themes, the paper pairs have similar methodologies. The similarity between the seed and the *dataset*-specific recommendations is evident. *Gupta et al. [23]*, *Xiang et al. [75]*, and *Meisheri and Khadilkar [46]* are all about sentiment classification in low resource settings. Instead of data augmentation with GAN, they utilize external knowledge or transfer learning.

In summary, we consider all recommendations as generally relevant since they share one or more aspects with the seed. Due to the subjectiveness of relevance, a recommender system would

Table 5: Example recommendations from SPECTER (generic) and Siamese-SciBERT (aspect-specific) for the seed *"Data augmentation for low resource sentiment analysis using generative adversarial networks"* by Gupta [22]

	Generic	Task	Method	Dataset
1	Adversarial Training for Aspect-Based Senti. Analysis with BERT [32]	Not Enough Data? Deep Learning to the Rescue! [1]	DNA Methylation Data to Predict Suicidal and Non-Suicidal Deaths: A ML. Approach [76]	Semi-Supervised and Transfer Learning Approaches for Low Resource Senti. Class. [23]
2	Emotion Classification with Data Augmentation Using Generative Adversarial Networks [78]	Towards better detection of spear-phishing emails [58]	Company Class. using Machine Learning [29]	Affection Driven Neural Networks for Senti. Analysis [75]
3	Hierarchical Attention Generative Adversarial Networks for Cross-domain Senti. Class. [77]	Conditional BERT Contextual Augmentation [74]	Inductive Hashing on Manifolds [65]	Learning Representations for Senti. Class. using Multi-task framework [46]

need to relate the recommendations to its users' individual information needs. However, when new user data is unavailable, this is not feasible. This is a general problem of purely content-based recommendations. Our example illustrates how different aspects can approximate similar research papers in a granular and more detailed perspective. The specialization from Siamese-SciBERT also leads to diverse recommendations between aspect-specific recommendations and generic ones. SPECTER's generic recommendations have a relatively narrow focus on data augmentation with GAN for classification. The *method*-specific recommendations even reveal the implicit shared use of the t-SNE visualization.

7 DISCUSSION

Our quantitative and qualitative results reveal the effect of specialized document embeddings. The performance gains between the best generic and the best specialized embeddings, i. e., generic SPECTER and Siamese-SciBERT, are substantial. We anticipated this outcome as the generic embeddings are not optimized for this task compared to the specialized ones. Still, our findings do not mean generic embeddings lead to unrelated recommendations, but only that they are not similar concerning *task*, *aspects*, or *dataset*. Siamese-SciBERT also outperforms the Pairwise SPECTER baseline. The pairwise SPECTER with a unbounded n may yield better results than our baseline implementation. However, due to the quadratic complexity, we have to perform 1.3 billion comparisons, which would take approximately 46 days on the hardware used in our experiments (GeForce RTX 2080 Ti with 11GB memory). Thus, the potential performance gains do not justify the increase in computational effort.

Specialization Performance. In terms of specialization, the Siamese Transformer (Siamese-SciBERT) outperforms retrofitting and non-Siamese Transformer fine-tuning. This outcome can be explained by several reasons. The retrofitting method from Glavaš and Vulić [19] has been originally developed for words and optimized for the properties of a word embedding space. We see retrofitting has a positive effect on Avg. FastText but a negative effect on SPECTER. SPECTER uses citation information and, therefore, its embedding space has different properties [12]. At the same time, SPECTER's citation information generally improves the performance of its generic and fine-tuned version compared to SciBERT. The poor performance of SciBERT is aligned with the results of related studies [53, 60], which show that document embeddings from BERT-based models are unsuited for the similarity search. Since we perform the similarity search based on static embeddings, each document needs to be independently encoded. While this is the case in Siamese-SciBERT, the non-Siamese Transformers (SPECTER and SciBERT) are fine-tuned in the sequence pair classification setting, i. e., a document pair is jointly encoded. As the results from [53] suggest, the joint encoding is superior for pairwise document classification approach. However, our results show the opposite in a similarity search setting. The independent encoding, as in the Siamese model, produces semantically similar documents embeddings with higher precision and recall.

Given the overall results, we consider Siamese-SciBERT as the best tested method to specialize embeddings. Nevertheless, we ask ourselves if the specialization effect depends on individual aspects. The most positive specialization effect can be observed for the *method* aspect, while the effect is less significant for *dataset*. We partially attribute the discrepancy in the specialization effect to training data availability, e. g., more samples for *method* than *dataset*. However, the effect is also due to the aspects being differently inherent in generic embeddings' similarity.

Bias in Generic Embeddings. The similarity of generic embeddings does not explicitly contain aspect information, i. e., we cannot attribute the document similarity to a specific aspect in which documents are similar. However, we can assume the aspects are implicitly part of the similarity. Thus, the similarity of generic embeddings would be denoted as a weighted sum $\sum_{a \in A} w_a * s_a$, where $A = \{task, method, dataset, \dots a_n\}$ is a set of aspects consisting of our three and an arbitrary number of other aspects. If the similarity of generic embeddings would evenly incorporate all aspects, all weights w_a should be equal. Still, our experiments suggest the aspects are not equally weighted. Table 4 reports an uneven intersection ratio among the recommendations. The *method*-specific recommendations have less overlap with the generic recommendation than the *dataset* or *task*-specific recommendations. Given that *task* has the most samples in the ground truth, we would have

expected a different outcome, e. g., more specialization concerning *task*. Therefore, $w_{method} < w_{task} < w_{dataset}$ likely holds true. Accordingly, the results indicate an implicit bias in the similarity of generic embeddings towards *dataset* and against *method*. Our qualitative analysis does not reject this finding. We hypothesize the bias is more likely to be caused by the corpus' characteristics than by the embedding methods themselves. Title and abstract of papers prominently mention tasks and datasets, whereas methodological details are of marginal importance, e. g., the t-SNE visualization in our example from Table 5.

Implications for Content-based Recommender Systems. Having this bias towards a single aspect indicates the generic embeddings present only a single view on the content of a document. Therefore, the conflation of meaning, which have been shown for word embeddings [8, 55], also exists for document embeddings. Consequently, a recommender system based on the generic embeddings is limited in the information needs that the system can address. Namely, those information needs that match with the single aspect, which is in our case the *dataset* aspect. Such a narrow focus on one information need hurts the diversity of the recommendations. In the literature [18, 49], the lack of diversity has been identified as a major issue of today's recommender systems. By changing the approach of representing documents, from generic to specialized embeddings, diverse information needs can be addressed even when user data is sparse. In the context of recommendations, our data does not allow a decisive statement on the relevancy of the generic or aspect-based recommendations since we primarily evaluate the similarity of research papers. We use similarity only as an approximation of relevance for specific information needs, i. e., interest in the task, method, or dataset of the presented research. To the best of our knowledge, a dataset that would allow a relevance-based evaluation of the Papers with Code corpus is not publicly available. Thus, further experiments involving user feedback are required to investigate the relevancy of aspect-based recommendations. Nonetheless, the recommendations from specialized embeddings can expose the implicit bias within the generic recommendations. Integrating the aspect information can improve research paper recommender systems as users would decide in which particular aspect they are interested. Thereby, tailored content-based recommendations are feasible even without user feedback. The aspect-based recommendation would increase the transparency of a recommender system since the system could provide explicit explanations on the aspects in that documents are related. Such explanations would also strengthen the trust in the recommendations as Kunkel et al. [37] demonstrate. Furthermore, diversity can be addressed through selection from multiple aspects. In a user interface (see [50]), one would not only display recommendations from a particular aspect but rather select one recommendation from each aspect, e. g., the top recommendation for *task*, *method*, and *dataset* (the items in the first row of Table 5).

Scalability. Diversity and explainability are also covered by the pairwise multi-class classification approach [52]. However, the pairwise approach bears scalability constrains that would prevent recommender systems to be deployed in a production environment. Pairwise document classification requires large computational resources even for medium-sized corpora since aspect information need to be separately derived for all document pairs. To use the pairwise approach as a baseline, we introduced the candidate filtering but it still needs to perform 11.3M Transformer forward-passes while achieving only a lower performance compared to Siamese-SciBERT. Instead, our approach derives the aspect information during the encoding phase, which results in a linear complexity (118,146 forward-passes in our experiments). During the indexing of a new document, the system would only need to create n specialized embeddings instead of a single generic embedding. Thus, our approach's complexity is mainly bound to the number of aspects and not to the size of the document corpus as in pairwise classification (see Section 3.1). As a result, our approach is applicable for real-world recommender systems on commodity hardware. Our Web-based demo is one example for prototypical recommender system based on specialized document embeddings[2].

Interpretability. Aside from scalability, the specialized embeddings have additional advantages such as explainablility and interpretability. Each individual aspect-specific vector $\vec{d}_i^{(a_j)}$ could also be combined through concatenation into a single document vector $\vec{d}_i = [\vec{d}_i^{(a_1)}; \ldots; \vec{d}_i^{(a_n)}]$ for other downstream tasks. The aspect's dimensions could then facilitate the interpretability of the document vectors in similar fashion as Liao et al. [41] already demonstrated with sparse vectors. In the context of words, related approaches already exist. For example, Schwarzenberg et al. [64] project word vectors into a concept space in which the dimensions correspond to predefined concepts.

Alternative Approaches. Lastly, the question is whether comparable recommendations are also possible with alternative approaches such as query-sensitive similarity [68]. One could filter papers by a query, i. e., their respective aspect labels, and then perform a nearest neighbor search on the filtered papers' generic embeddings. However, the filtering depends on hard label assignments, e. g., papers need to have an identical task, method, or dataset to be considered. Papers only similar in a particular aspect would be excluded. In our example (Table 5), *Zhu et al. [78]* would have been excluded because its task is *emotion classification* related but not identical with *sentiment classification* as in the seed document. Moreover, the specialized embedding space allows dissimilarity search, e. g., considering papers with similarity above a certain threshold. This allows retrieving papers similar in their task but different in their method. The formulation of such queries could furthermore facilitate the discovery of analogies between research papers [9].

8 CONCLUSIONS

This paper introduces our approach of specialized document embeddings for aspect-based document similarity of research papers. Instead of considering each research paper as a single entity for document similarity, we incorporate multiple aspects in our approach, i. e., *task*, *method*, and *dataset*. Therefore, we move from a single generic representation to three specialized ones. We treat aspect-based similarity as a classical vector similarity problem in aspect-specific embedding spaces. Our approach contributes two major improvements over existing literature of aspect-based document similarity: In contrast to segment-level similarity [9, 28, 34], a document is not divided into segments which harms the coherence

of a document. Instead, we preserve the semantics of the whole document that are needed for a meaningful representation. Additionally, our approach is less resource intensive and achieves a higher precision and recall compared to the pairwise document classification baseline [52, 53]. The improved scalability allows the development a real-world recommender system, which we demonstrate with our demo[2].

In our empirical study, we compare and analyze three generic document embeddings, six specialized document embeddings and a pairwise classification baseline in the context of research paper recommendations. To the best of our knowledge, all applied specialization methods were, so far, used only to derive generic embeddings. Our evaluation is conducted on the newly constructed Papers with Code corpus containing more than 150,000 research papers. This Papers with Code corpus is unique for research on aspect-based document similarity as it contains manual annotations regarding different aspects of research papers. In our experiments, Siamese-SciBERT outperforms all other methods with 0.224 MAP for *task*-, 0.137 MAP for *method*-, and 0.235 MAP for *dataset*-specific recommendations. Our comparison between recommendations using generic and specialized embeddings indicates a tendency of generic recommendations being more similar regarding *dataset* than *method*. Thus, papers with a similar method are less likely to be recommended with these generic embeddings. Our approach of aspect-based document embeddings mitigates potential risks arising from implicit biases by making them explicit. This can, for example, be used for more diverse and explainable recommendations, e. g., by recommending documents for every aspect. The development of an aspect-based recommender system and its evaluation with user feedback is subject to future work.

ACKNOWLEDGEMENTS

The research presented in this article is partially funded by the German Federal Ministry of Education and Research (BMBF) through the projects QURATOR [59] (Unternehmen Region, Wachstumskern, no. 03WKDA1A) and PANQURA (no. 03COV03E).

REFERENCES

[1] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do Not Have Enough Data? Deep Learning to the Rescue! , 7383–7390 pages. https://doi.org/10.1609/aaai.v34i05.6233 arXiv:1911.03118

[2] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but though Baseline for Sentence Embeddings. In *5th International Conference on Learning Representations (ICLR 2017)*, Vol. 15. Toulon, France, 416–424.

[3] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2017. ANN-Benchmarks: A Benchmarking Tool for Approximate Nearest Neighbor Algorithms. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10609 LNCS. Springer, 34–49. https://doi.org/10.1007/978-3-319-68474-1_3 arXiv:1807.05614

[4] Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2011. A reflective view on text similarity. *International Conference Recent Advances in Natural Language Processing, RANLP* (2011), 515–520.

[5] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 3613–3618. https://doi.org/10.18653/v1/D19-1371 arXiv:1903.10676

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. arXiv:1607.04606 http://arxiv.org/abs/1607.04606

[7] Jane Bromley, J.W. Bentz, Leon Bottou, I. Guyon, Yann Lecun, C. Moore, Eduard Sackinger, and R. Shah. 1993. Signature verification using a Siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence* 7, 4 (1993).

[8] Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research* 63 (dec 2018), 743–788. https://doi.org/10.1613/jair.1.11259

[9] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (nov 2018), 1–21. https://doi.org/10.1145/3274300

[10] Jun Chen, Chaokun Wang, and Jianmin Wang. 2017. Modeling the intransitive pairwise image preference from multiple angles. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference* (2017), 351–359. https://doi.org/10.1145/3123266.3123285

[11] Jun Chen, Chaokun Wang, Jianmin Wang, Xiang Ying, and Xuecheng Wang. 2017. Learning the Personalized Intransitive Preferences of Images. *IEEE Transactions on Image Processing* 26, 9 (2017), 4139–4153. https://doi.org/10.1109/TIP.2017.2709941

[12] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[13] T. M. Cover and P. E. Hart. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the ACL*. ACL, Minneapolis, Minnesota, 4171–4186.

[15] Hai Ha Do, P. W.C. Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Systems with Applications* 118 (2019), 272–299. https://doi.org/10.1016/j.eswa.2018.10.003

[16] David Ellis, Jonathan Furner-Hines, and Peter Willett. 1993. Measuring the Degree of Similarity Between Objects in Text Retrieval Systems. *Perspectives in Information Management* 3, 2 (1993), 128–149.

[17] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1606–1615. https://doi.org/10.3115/v1/N15-1184

[18] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proc. of the fourth ACM Conf. on Recommender Systems*. ACM Press, New York, New York, USA, 257.

[19] Goran Glavaš and Ivan Vulić. 2018. Explicit Retrofitting of Distributional Word Vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 37. Association for Computational Linguistics, Stroudsburg, PA, USA, 34–45.

[20] Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. 2020. Recurrent Chunking Mechanisms for Long-Text Machine Reading Comprehension. In *Proc. of the 58th Annual Meeting of the Assoc. for Computational Linguistics*. ACL, Stroudsburg, PA, USA, 6751–6761.

[21] Nelson Goodman. 1972. Seven strictures on similarity. *Problems and Projects* (1972).

[22] Rahul Gupta. 2019. Data Augmentation for Low Resource Sentiment Analysis Using Generative Adversarial Networks. In *ICASSP 2019 - 2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 2019-May. IEEE, 7380–7384.

[23] Rahul Gupta, Saurabh Sahu, Carol Espy-Wilson, and Shrikanth Narayanan. 2018. Semi-Supervised and Transfer Learning Approaches for Low Resource Sentiment Classification. In *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 2018-April. IEEE, 5109–5113.

[24] Jialong Han, Yan Song, Wayne Xin Zhao, Shuming Shi, and Haisong Zhang. 2018. hyperdoc2vec: Distributed Representations of Hypertext Documents. In *Proc. of the 56th Annual Meeting of the Assoc. for Computational Linguistics*, Vol. 1. ACL, Stroudsburg, PA, USA, 2384–2394.

[25] Zellig S. Harris. 1954. Distributional Structure. *WORD* 10, 2-3 (aug 1954), 146–162. https://doi.org/10.1080/00437956.1954.11659520

[26] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. (may 2017). arXiv:1705.00652

[27] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2004), 168–177. https://doi.org/10.1145/1014052.1014073

[28] Ting-Hao 'Kenneth' Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Reliably Annotating Research Aspects on 10,000+ CORD-19 Abstracts Using a Non-Expert Crowd. (2020). arXiv:2005.02367

[29] Sven Husmann, Antoniya Shivarova, and Rick Steinert. 2020. Company classification using machine learning. *arXiv 2004.01496* (mar 2020). arXiv:2004.01496 http://arxiv.org/abs/2004.01496

[30] Karen Sparck Jones. 1973. Index term weighting. *Information Storage and Retrieval* 9, 11 (1973). https://doi.org/10.1016/0020-0271(73)90043-0

[31] Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. AxCell: Automatic Extraction of Results from Machine Learning Papers. In *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Online, 8580–8594.

[32] Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2020. Adversarial Training for Aspect-Based Sentiment Analysis with BERT. (2020). arXiv:2001.11316

[33] Maryam Karimzadehgan, Cheng Xiang Zhai, and Geneva Belford. 2008. Multi-aspect expertise matching for review assignment. *International Conference on Information and Knowledge Management, Proceedings* (2008), 1113–1122. https://doi.org/10.1145/1458082.1458230

[34] Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. 2018. Citation Recommendation Using Distributed Representation of Discourse Facets in Scientific Articles. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. ACM, New York, NY, USA, 243–251.

[35] David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 2022. A Domain-adaptive Pre-training Approach for Language Bias Detection in News. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (2022-06-20). Kø"ln. Accepted for publication.

[36] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems – A survey. *Knowledge-Based Systems* 123 (2017), 154–162.

[37] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proc. of the 2019 CHI Conf. on Human Factors in Computing Sys.* ACM, New York, NY, USA, 1–12.

[38] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning* 32 (2014), 1188–1196.

[39] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[40] Jiwei Li and Dan Jurafsky. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding?. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1722–1732.

[41] Keng-te Liao, Pochun Chen, Kuansan Wang, and Shou-de Lin. 2020. Explainable and Sparse Representations of Academic Articles for Knowledge Exploration. In *Proceedings of the 28th International Conference on Computational Linguistics*. 6207–6216.

[42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[43] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou. 2011. Multi-aspect sentiment analysis with topic models. *Proceedings - IEEE International Conference on Data Mining, ICDM* (2011), 81–88. https://doi.org/10.1109/ICDMW.2011.125

[44] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Vol. 16. Cambridge University Press, Cambridge. 100–103 pages. https://doi.org/10.1017/CBO9780511809071

[45] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning Attitudes and Attributes from Multi-aspect Reviews. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 1020–1025.

[46] Hardik Meisheri and Harshad Khadilkar. 2018. Learning representations for sentiment classification using Multi-task framework. In *Proc. of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. ACL, Stroudsburg, PA, USA, 299–308.

[47] Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic Specialization of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *Transactions of the Association for Computational Linguistics* 5 (dec 2017), 309–324. https://doi.org/10.1162/tacl_a_00063 arXiv:1706.00374

[48] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. 2022. VITALITY: Promoting Serendipitous Discovery of Academic Literature with Transformers & Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (jan 2022), 486–496. https://doi.org/10.1109/TVCG.2021.3114820 arXiv:2108.03366

[49] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proc. of the 23rd Int. Conf. on World Wide Web*. ACM Press, New York, New York, USA, 677–686.

[50] Malte Ostendorff. 2020. Contextual Document Similarity for Content-based Literature Recommender Systems. (2020). https://doi.org/10.48550/ARXIV.2008.00202

[51] Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. *ArXiv* abs/2202.06671 (2022).

[52] Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based Document Similarity for Research Papers. In *Proc. of the 28th Int. Conf. on Computational Linguistics (COLING 2020)*. https://doi.org/10.18653/v1/2020.coling-main.545

[53] Malte Ostendorff, Terry Ruas, Moritz Schubotz, Georg Rehm, and Bela Gipp. 2020. Pairwise Multi-Class Document Classification for Semantic Relations between Wikipedia Articles. In *Proc. of the 2020 ACM/IEEE Joint Conf. on Digital Libraries (JCDL '20)*. https://doi.org/10.1145/3383583.3398525

[54] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proc. of the 2018 Conf. of the North American Chapter of the ACL*. ACL, Stroudsburg, PA, USA, 2227–2237.

[55] Mohammad Taher Pilehvar and Nigel Collier. 2016. De-Conflated Semantic Representations. In *Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1680–1690. https://doi.org/10.18653/v1/D16-1174

[56] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proc. of the 8th Int. Workshop on Semantic Evaluation (SemEval 2014)*. ACL, Stroudsburg, PA, USA, 27–35.

[57] Jason Portenoy, Marissa Radensky, Jevin West, Eric Horvitz, Daniel Weld, and Tom Hope. 2021. *Bursting Scientific Filter Bubbles: Boosting Innovation via Novel Author Discovery*. Vol. 1. Association for Computing Machinery. https://doi.org/10.1145/3491102.3501905 arXiv:2108.05669

[58] Mehdi Regina, Maxime Meyer, and Sébastien Goutal. 2020. Text Data Augmentation: Towards better detection of spear-phishing emails. (2020), 1–31. arXiv:2007.02033

[59] Georg Rehm, Peter Bourgonje, Stefanie Hegele, Florian Kintzel, Julián Moreno Schneider, Malte Ostendorff, Karolina Zaczynska, Armin Berger, Stefan Grill, Sören Räuchle, Jens Rauenbusch, Lisa Rutenburg, André Schmidt, Mikka Wild, Henry Hoffmann, Julian Fink, Sarah Schulz, Jurica Seva, Joachim Quantz, Joachim Böttger, Josefine Matthey, Rolf Fricke, Jan Thomsen, Adrian Paschke, Jamal Al Qundus, Thomas Hoppe, Naouel Karam, Frauke Weichhardt, Christian Fillies, Clemens Neudecker, Mike Gerber, Kai Labusch, Vahid Rezanezhad, Robin Schaefer, David Zellhöfer, Daniel Siewert, Patrick Bunk, Lydia Pintscher, Elena Aleynikova, and Franziska Heine. 2020. QURATOR: Innovative Technologies for Content and Data Curation. In *Proceedings of QURATOR 2020 – The conference for intelligent content solutions*, Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, and Lydia Pintscher (Eds.). Berlin, Germany. CEUR Workshop Proceedings, Volume 2535. 20/21 January 2020.

[60] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *The 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*. arXiv:1908.10084 http://arxiv.org/abs/1908.10084

[61] Terry Ruas, Charles P. H. Ferreira, William Gorsky, Fabrício O. França, and Débora M. R. Medeiros. 2020. Enhanced word embeddings using multi-semantic representation through lexical chains. *Information Sciences* 532 (2020), 16 –32. https://doi.org/10.1016/j.ins.2020.04.048

[62] Terry Ruas, William Gorsky, and Akiko Aizawa. 2019. Multi-sense embeddings through a word sense disambiguation process. *Expert Systems with Applications* 136 (2019), 288 – 303. https://doi.org/10.1016/j.eswa.2019.06.026

[63] Gerard Salton. 1963. Associative Document Retrieval Techniques Using Bibliographic Information. *J. ACM* 10, 4 (Oct. 1963), 440–457.

[64] Robert Schwarzenberg, Lisa Raithel, and David Harbecke. 2019. Neural Vector Conceptualization for Word Vector Space Interpretation. In *Proc. of the 3rd Workshop on Evaluating Vector Space Representations*. ACL, Stroudsburg, PA, USA, 1–7.

[65] Fumin Shen, Chunhua Shen, Qinfeng Shi, Anton van den Hengel, and Zhenmin Tang. 2013. Inductive Hashing on Manifolds. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1562–1569. https://doi.org/10.1109/CVPR.2013.205 arXiv:1303.7043

[66] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 1166–1177. https://doi.org/10.18653/v1/2021.findings-emnlp.101

[67] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *Proc. of the 24th Int. Conf. on World Wide Web*. ACM Press, New York, New York, USA, 1067–1077.

[68] Anastasios Tombros and C. J. Van Rijsbergen. 2001. Query-Sensitive similarity measures for the calculation of interdocument relationships. *International Conference on Information and Knowledge Management, Proceedings* (2001), 17–24. https://doi.org/10.1145/502586.502589

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proc. of the 31st Int. Conf. on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. 6000–6010.

[70] Jan Philip Wahle, Nischal Ashok, Terry Ruas, Norman Meuschke, Tirthankar Ghosal, and Bela Gipp. 2022. Testing the Generalization of Neural Language Models for COVID-19 Misinformation Detection. In *Information for a Better World: Shaping the Global Future*, Malte Smits (Ed.). Vol. 13192. Springer International Publishing, Cham, 381–392. https://doi.org/10.1007/978-3-030-96957-8_33 Series Title: Lecture Notes in Computer Science.

[71] Jan Philip Wahle, Terry Ruas, Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2022. Identifying Machine-Paraphrased Plagiarism. In *Information for a Better World: Shaping the Global Future*, Malte Smits (Ed.). Vol. 13192. Springer International Publishing, Cham, 393–413. https://doi.org/10.1007/978-3-030-96957-8_34 Series Title: Lecture Notes in Computer Science.

[72] Jan Philip Wahle, Terry Ruas, Norman Meuschke, and Bela Gipp. 2021. Are Neural Language Models Good Plagiarists? A Benchmark for Neural Paraphrase Detection. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, Champaign, IL, USA, 226–229. https://doi.org/10.1109/JCDL52503.2021.00065 tex.ids= WahleRMG21 arXiv: 2103.12450.

[73] Jan Philip Wahle, Terry Ruas, Norman Meuschke, and Bela Gipp. 2021. Incorporating Word Sense Disambiguation in Neural Language Models. *CoRR* abs/2106.07967 (2021). arXiv:2106.07967 https://arxiv.org/abs/2106.07967

[74] Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional BERT Contextual Augmentation. In *Lecture Notes in Computer Science*, Vol. 11539 LNCS. 84–95.

[75] Rong Xiang, Yunfei Long, Mingyu Wan, Jinghang Gu, Qin Lu, and Chu-ren Huang. 2020. Affection Driven Neural Networks for Sentiment Analysis. In *Proc. of the 12th Language Resources and Evaluation Conf.* European Language Resources Association, Marseille, France, 112–119.

[76] Rifat Zahan, Ian McQuillan, and Nathaniel Osgood. 2018. DNA Methylation Data to Predict Suicidal and Non-Suicidal Deaths: A Machine Learning Approach. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 363–365. https://doi.org/10.1109/ICHI.2018.00057

[77] Yuebing Zhang, Duoqian Miao, and Jiaqi Wang. 2019. Hierarchical Attention Generative Adversarial Networks for Cross-domain Sentiment Classification. (2019). arXiv:1903.11334

[78] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. 2018. Emotion Classification with Data Augmentation Using Generative Adversarial Networks. In *Lecture Notes in Computer Science*, Vol. 10939 LNAI. 349–360.