Automatically Detecting Scientific Political Science Texts from a Large General Document Index

1. Introduction

The specialized information service (FID) Political Science – Pollux[1] aims to offer political scientists subject-specific and fast access to special literature and information relevant to their research. Pollux obtains data from different providers, one of them is BASE[2] (Bielefeld Academic Search Engine), which indexes the metadata of all kinds of academically relevant resources. Many BASE entries have no domain classification; therefore, we developed a filtering approach which can detect entries pertinent to political science from the huge collection of BASE metadata. Additionally, publications relevant to political science also might be included in other scientific domains. State and University Library Bremen (SuUB) has used BASE filtering methods for their Electronic library (E-LIB) since 2012[3]. The core BASE filtering approach including a weighted keyword-based approach was developed in 2016 by the SuUB team for the Pollux project.

In this practical report, we describe the training procedure, training data and usage of two classification models which are part of one of the modules (soft filter) of our filtering approach (see Figure 1). Additionally, we provide an evaluation of both classification models and a weighted keyword-based approach (hard filter) on a specially designed test dataset.[4,5]

1.1 BASE

BASE is a search engine for academic web resources operated by Bielefeld University Library. The BASE provides over 340 million documents from different content providers and consists of various kinds of academic resources, e.g., journals, books digital collections, etc[6]. Many BASE records include a class number from the Dewey Decimal Classification (DDC). Certain content providers supply DDC numbers with their records, which are directly incorporated into the browsing system. Additionally,

1 POLLUX website. https://www.pollux-fid.de/, accessed 14.10.2024.
2 BASE search. https://www.base-search.net/, accessed 14.10.2024.
3 Blenkle, Martin; Ellis, Rachel; Haake, Elmar u.a.: Green Open Access im Bibliothekskatalog. Chancen & Risiken [WissKom 2016], Jülich 2016 (Schriften des Forschungszentrums Jülich Reihe Bibliothek / Library 22); Blenkle, Martin; Nölte, Manfred: Open Access Medien in den Bibliothekskatalog! Chancen&Risiken, Konferenzfolien, 2017.
4 Nina Smirnova acknowledges support by Deutsche Forschungsgemeinschaft (DFG) under grant number MA 3964/7-3, the Fachinformationsdienst Politikwissenschaft – Pollux. The core BASE-filtering pipeline was developed by the team of the State and University Library Bremen (SuUB): Martin Blenkle, Elmar Haake, Thore Christiansen, Marie-Saphira Flug, Lena Klaproth. The weighted keyword-based filtering approach was developed by Daniel Opitz. Michael Czolkoß-Hettwer contributed to the creation of the list of keywords for the political science domain. Nina Smirnova was founded by the European Union under the Horizon Europe OMINO – Overcoming Multilevel Information Overload (101086321, https://ominoproject.eu/). Views and opinions expressed are those of the authors alone and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them.
5 This article is based on the presentation 'Metadaten mit KI-Methoden filtern – die Nutzung von BASE im FID Politikwissenschaft' as part of the panel 'Open/Linked Data und/mit KI' on 6 June 2024 at the 112th BiblioCon in Hamburg.
6 BASE search. https://www.base-search.net/about/en/index.php, accessed 14.10.2024.

BASE employs automatic classification of documents[7]. The DDC is a system for organizing library contents by dividing all knowledge into ten categories, each assigned a range of 100 numbers. Aside from DDC numbers, BASE records contain many other important metadata elements, such as source, journal name, authors' names, journal name, etc.

1.2 Filtering Workflow

Developing approaches for filtering political science articles from a collection of unlabeled or multidisciplinary scientific articles is crucial for several reasons. Hołyst et al.[8] argue that excessive information constrains our ability to assess information and make optimal decisions. Therefore, excluding articles irrelevant to political science can enhance political scientists' time efficiency and enable focusing on the materials most pertinent to their work. Political science often intersects with other disciplines. Finding articles related to politics in other disciplines will provide political researchers with opportunities for interdisciplinary collaboration, enrich their perspectives and enhance the depth of their research. Political science is a dynamic field, with new events, policies, and trends continuously emerging. Timely access to relevant literature allows political researchers to remain abreast of the latest developments in the political field, enabling informed analyses and discussions.

The filtering approach for the BASE data collection comprises different modules to address entries with different metadata available as illustrated in Figure 1. Initially, the data is filtered according to parameters such as data type and source. Subsequently, in the second filter articles already categorized under the DDC system as political science (DDC 320-328) are selected. The third filter selects articles with a valid abstract. For articles lacking abstracts, a keyword-based filter (hard filter) is applied. Articles containing abstracts undergo a language filter, and those written in permissible languages are then subjected to a BERT-based classification model (soft filter). BERT (Bidirectional Encoder Representations Transformers) is a transformer-based language model designed to capture contextualized word embeddings by learning dynamic representations of words based on their surrounding context[9]. We apply additional classification to the BASE metadata for two primary reasons. First, according to BASE, only 9% of the entire dataset has an assigned DDC number[10]. Second, we are particularly interested in interdisciplinary articles, such as those originating from computer science but applied to political science, which are not always adequately captured by the existing BASE classification.

7 Waltinger, Ulli; Mehler, Alexander; Lösch, Mathias u.a.: Hierarchical Classification of OAI Metadata Using the DDC Taxonomy, in: Bernardi, Raffaella; Chambers, Sally; Gottfried, Björn u.a. (Hg.): Advanced Language Technologies for Digital Libraries, Bd. 6699, Berlin, Heidelberg 2011 (Lecture Notes in Computer Science), S. 29–40, https://doi.org/10.1007/978-3-642-23160-5_3, accessed 01.10.2024.

8 Protect our environment from information overload, in: Nature Human Behaviour 8 (3), 07.02.2024, S. 402–403, https://doi.org/10.1038/s41562-024-01833-8.

9 Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton u.a.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018, https://doi.org/10.48550/ARXIV.1810.04805, accessed 04.07.2024.

10 BASE search. https://www.base-search.net/Browse/Dewey, accessed 14.10.2024.

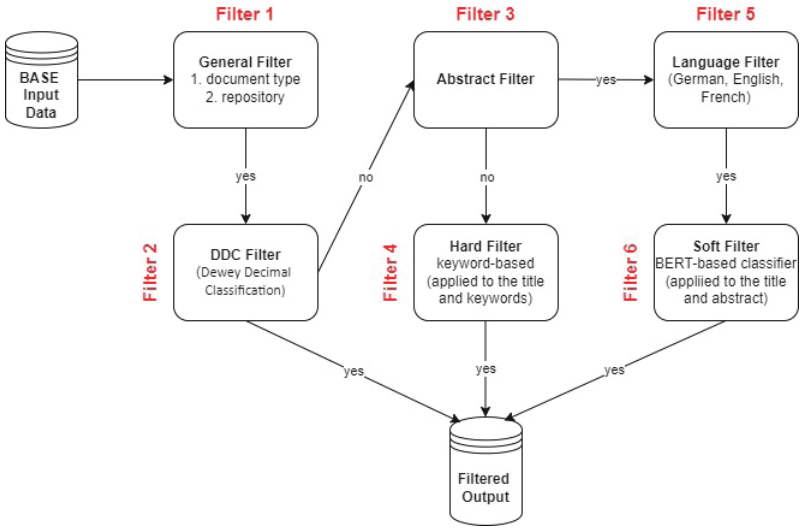

Figure 1: Pipeline for the BASE filtering approach

The two major modules of this approach are so-called hard and soft filters. The hard filter is a weighted keyword-based filter approach. This module is applied to the records which do not contain a full abstract, and therefore the filter is applied to the record's title and keywords if the latter are available. We developed a set of keywords which we believe to be effective for identifying data related to the political science domain. This set comprises 65 keywords primarily in English and German, also including root words which are similar for most Romance and Germanic languages, such as, e.g., 'politi' and 'ideolog'. Each keyword was manually assigned with a score. The score is assigned according to the keyword's relevancy for the domain, as Table 1 demonstrates. Figure 2 shows a simplified workflow for the hard filter. The filtering algorithm looks for the keywords in the metadata (in our case title and keywords) and returns a sum of scores of all found keywords. If this score is more or equal to 1, the article is marked as relevant to the domain.

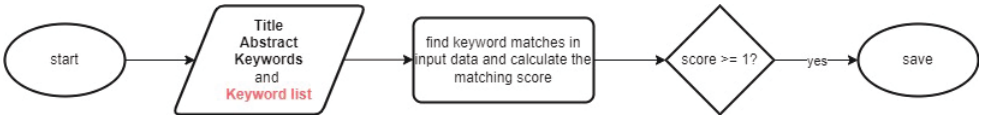

Figure 2: Hard filter simplified workflow

Table 1: Example of a keyword list for the hard filter

keyword	score
*politik	1
bürgerkrieg	0.6
policy	0.4

The soft filter uses a BERT-based classification model, trained to detect scientific articles from the political science domain. The model was finetuned on abstracts from scientific articles, therefore the filter is applied to the record's title and abstract.

2. Soft Filter

For our soft-filter approach, we trained two classification models: English and multilingual, which can detect texts from the political science domain. The English classification model is designed for the classification of articles written in English. The multilingual model extends the capabilities of the English model to accommodate additional languages. If your focus is solely on English-language articles, utilizing the English model is recommended due to its smaller size and faster processing speed compared to the multilingual model.

2.1 English Classification Model

Our English classification model to detect texts from the political science domain (SSciBERT_politics)[11] is based on the SSCI-SciBERT model [12]. SSCI-SciBERT was trained on the abstracts of social science research articles, therefore, we opted to use this model. SSciBERT_politics was fine-tuned by us using a dataset of 2,919 abstracts from scientific articles retrieved from the BASE and POLLUX collections.

The model was initially integrated into the soft filter of the BASE filtering pipeline in the POLLUX infrastructure (cf. Figure 1). Nevertheless, the model can be used separately with the Transformers library[13], as Listing 1[14] demonstrates. The Transformers library, developed by Hugging Face[15], is an open-source Python library designed for natural language processing (NLP) tasks.

11 The English classification model to detect texts from the political science domain. https://huggingface.co/kalawinka/SSciBERT_politics, acccessed 01.10.2024.
12 Shen, Si; Liu, Jiangfeng; Lin, Litao u.a.: SsciBERT. A pre-trained language model for social science texts, in: Scientometrics 128 (2), 02.2023, S. 1241–1263, https://doi.org/10.1007/s11192-022-04602-4.
13 Transformers library. https://huggingface.co/docs/transformers/en/index, accessed 01.10.2024.
14 Listings comprise code snippets written in Python and demonstrates the application usage of the models described in this practical report.
15 Huggingface. https://huggingface.co/, accessed 01.10.2024.

```
1. from transformers import
3. from transformers import pipeline
4.
5. tokenizer = AutoTokenizer.from_pretrained('kalawinka/SSciBERT_politics')
6. model = AutoModelForSequenceClassification.from_pretrained('kalawinka/SSciBERT_politics')
7. pipe = pipeline("text-classification", model=model, tokenizer = tokenizer, max_length=512,
 truncation=True)
8.
9. pipe('add scientific abstract')
```

Listing 1: Usage example of SSciBERT_politics with the Transformers library

Applying this model to the abstract of the practical report yields the output presented in Listing 2.

```
1. [{'label': 'multi', 'score': 0.9991981387138367}]
```

Listing 2: Example of SSciBERT_politics output

The model was fine-tuned utilizing the Transformers library and using the AdamW optimizer[16]. Evaluation was done at the end of each epoch. The total number of training epochs to perform was set to 4. The initial learning rate was set to 5e-5. The training was performed using one NVIDIA A40 GPU. Listing 3 demonstrates the training parameters used to fine-tune the SSciBERT_politics model.

```
1. training_args = TrainingArguments(output_dir=output_dir,
2.                                   evaluation_strategy="epoch",
3.                                   num_train_epochs = 4,
4.                                   )
```

Listing 3: Training parameters for the finetuning of SSciBERT_politics with the Transformers library

2.1.1 Training Data

Due to a lack of manually annotated data, we employed a semi-automated method to generate the labelled training dataset. Figure 3-A illustrates the distribution of articles across the training, test, and validation corpora. Accordingly, the test and validation datasets each contain 20 % of the data, while the training corpus comprises 60 % of the total data.

Data from the BASE and POLLUX collections were utilized to construct the training corpus. To acquire scientific abstracts within the domain of political science, we utilized both POLLUX and BASE collections. We compiled a list of political science journals in English. Afterwards, we collected articles from the POLLUX collection published in these journals containing an abstract and published after

16 AdamW optimizer. https://huggingface.co/docs/bitsandbytes/main/en/reference/optim/adamw, accessed 01.10.2024.

2018. Next, we collected articles from the BASE collection adhering to specific criteria. Only articles with the types article or review, DDC classification 320-328 (Political science), written in English, containing an abstract and published after 2018 was selected. To collect data from other (not political science) domains, we utilized only articles from the BASE collection[17]. We sampled articles which correspond to the following criteria. Only articles with the types article or review, DDC classification not 320-328 (Political science), written in English containing an abstract and published after 2015 were selected. Additionally, we designed our query to restrict the selection only to articles which do not contain the keyword 'politi*' in title, keywords or abstracts. Figure 3-B shows the distribution of scientific domains in the 'multi' category[18]. Figure 3-C shows the distribution of sources in the 'politics' category. After collecting the data from both collections, we verified the language of the abstracts to ensure they were in English and excluded abstracts containing fewer than 20 words.

17 The classification model requires data containing all relevant classes to ensure the model can learn discriminative features to distinguish between the categories.

18 One article can belong to multiple scientific domains.

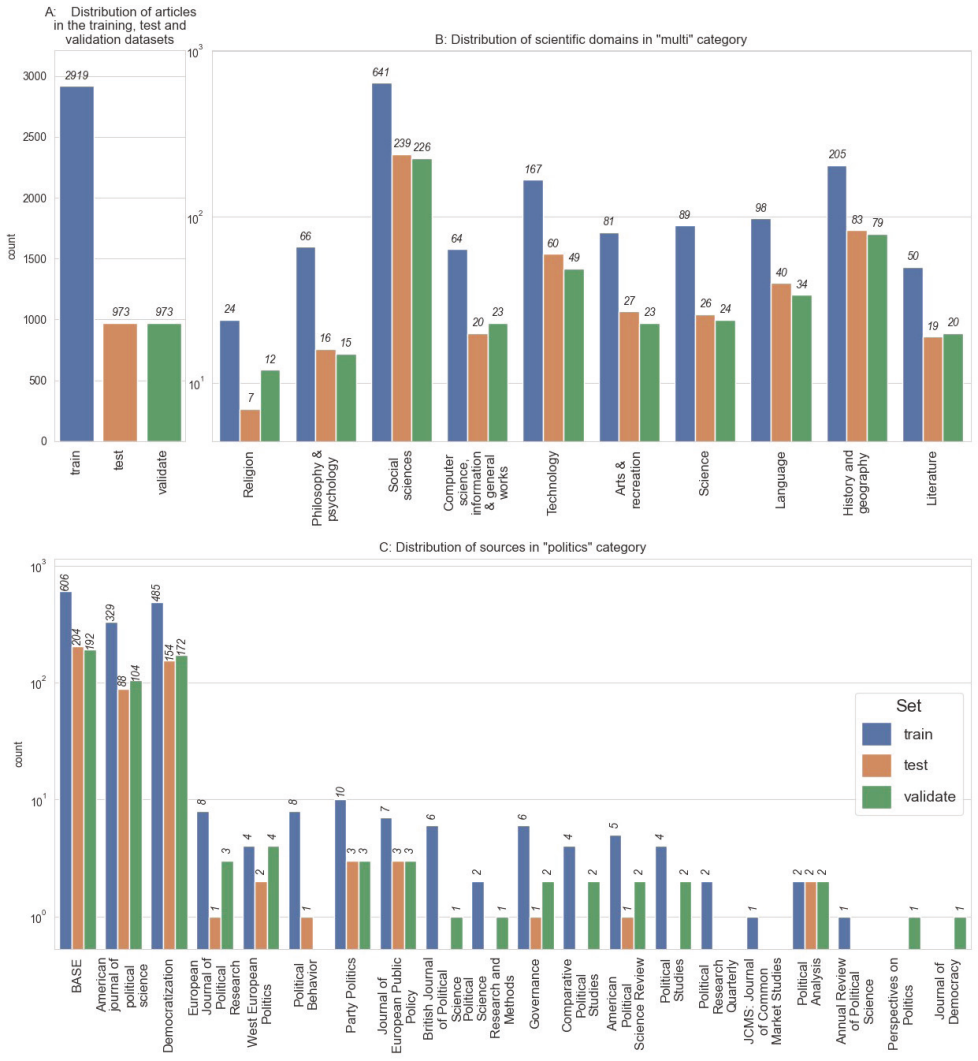

Figure 3: Distribution of training data in English corpus

2.2 The Multilingual Classification Model

The multilingual classification model to detect texts from the political science domain (bert-base-ml-politics)[19] is based on the BERT multilingual base model [20]. The BERT multilingual base model was trained on multilingual Wikipedia data (102 languages). As SSCI-SciBERT was trained only for English it was not suitable for the fine-tuning with multilingual data. Therefore, we utilized a multilingual language representation model for the fine-tuning.

The bert-base-ml-politics model was fine-tuned using a dataset of 14,178 abstracts from scientific articles retrieved from the BASE and POLLUX collections of scientific articles. Abstracts from scientific articles in 3 languages (English, German and French) were selected for the training.

The model is integrated into the soft filter of the BASE filtering pipeline in the POLLUX infrastructure. Nevertheless, the model can be used separately with the Transformers library, as Listing 4 demonstrates.

```
1. from transformers import AutoTokenizer, AutoModelForSequenceClassification
2. from transformers import pipeline
3.
4. tokenizer = AutoTokenizer.from_pretrained('kalawinka/bert-base-ml-politics')
5. model = AutoModelForSequenceClassification.from_pretrained('kalawinka/bert-base-ml-politics')
6. pipe = pipeline("text-classification", model=model, tokenizer = tokenizer, max_length=512,
   truncation=True)
7.
8. pipe('add scientific abstract')
```

Listing 4: Usage example of bert-base-ml-politics with the Transformers library

Applying this model to the abstract of this practical report produces the output illustrated in Listing 5.

```
1. [{'label': 'politics', 'score': 0.9972042441368103}]
```

Listing 5: Example of bert-base-ml-politics output

As SSciBERT_politics, the model was finetuned using the Transformers library as Listing 6 demonstrates. The AdamW algorithm was applied to minimise the loss function. Evaluation was done at the end of each epoch. The total number of training epochs to perform was set to 3. The initial learning rate was set to 5e-5. The training was performed using one NVIDIA A40 GPU. Training time comprises 52,20334 minutes.

19 The multilingual classification model to detect texts from the political science domain. https://huggingface.co/kalawinka/bert-base-ml-politics, accessed 01.10.2024.
20 Devlin u.a.: BERT.

```
1. training_args = TrainingArguments(output_dir=output_dir,
2.                                   evaluation_strategy = "epoch",
3.                                   num_train_epochs = 3,
4.                                   )
```

Listing 6: Training parameters for the finetuning of bert-base-ml-politics with the Transformers library

2.2.1 Training Data

The collection of the training data was similar to the one described in Section 2.1.1. Data from the BASE and POLLUX collections were utilized to construct the training corpus. A semi-automated method was employed to generate the labelled training dataset.

To obtain scientific abstracts from the political science domain we used both POLLUX and BASE collection. We created a list of political science journals from German-, English- and French-speaking countries. Afterwards, we collected articles from the POLLUX collection published in these journals containing an abstract and publication date. In the next step, we collected articles from the BASE collection corresponding to the following criteria. Only articles with the types article or review, DDC classification 320-328 (Political science), written in English, German or French, containing an abstract were selected. To collect data from other (not political science) domains, we utilized only articles from the BASE collection. We sampled articles which correspond to the following criteria. Only articles with the types article or review, DDC classification not 320-328 (Political science), written in English, German or French, containing an abstract were selected. Additionally we designed our query to restrict the selection only to articles, which do not contain the keyword 'politi*' in title, keywords or abstracts. Figure 4-C shows the distribution of scientific domains in the 'multi' category[21]. Figure 4-D shows the distribution of sources in the 'politics' category. After collecting the data from both collections, we verified the language of the abstracts to ensure they were in English, German, or French and excluded abstracts containing fewer than 20 words.

Figure 4-A shows the distribution of articles in the training, test and validation corpora. Thus, test and validation datasets contain 20% each and the training corpus comprises 60% of all data. Figure 4-B shows the distribution of languages in the training data. English is a prevailing language, as we encountered difficulties locating sufficient data in German and French. Furthermore, our training data is limited to these languages because we were unable to find adequate training data in other languages that met our criteria. Specifically, we required data that included both a title and a valid abstract. Many multilingual entries contained abstracts only in English. Additionally, for the BASE dataset, we required entries with the DDC classification.

21 One article can belong to multiple scientific domains.

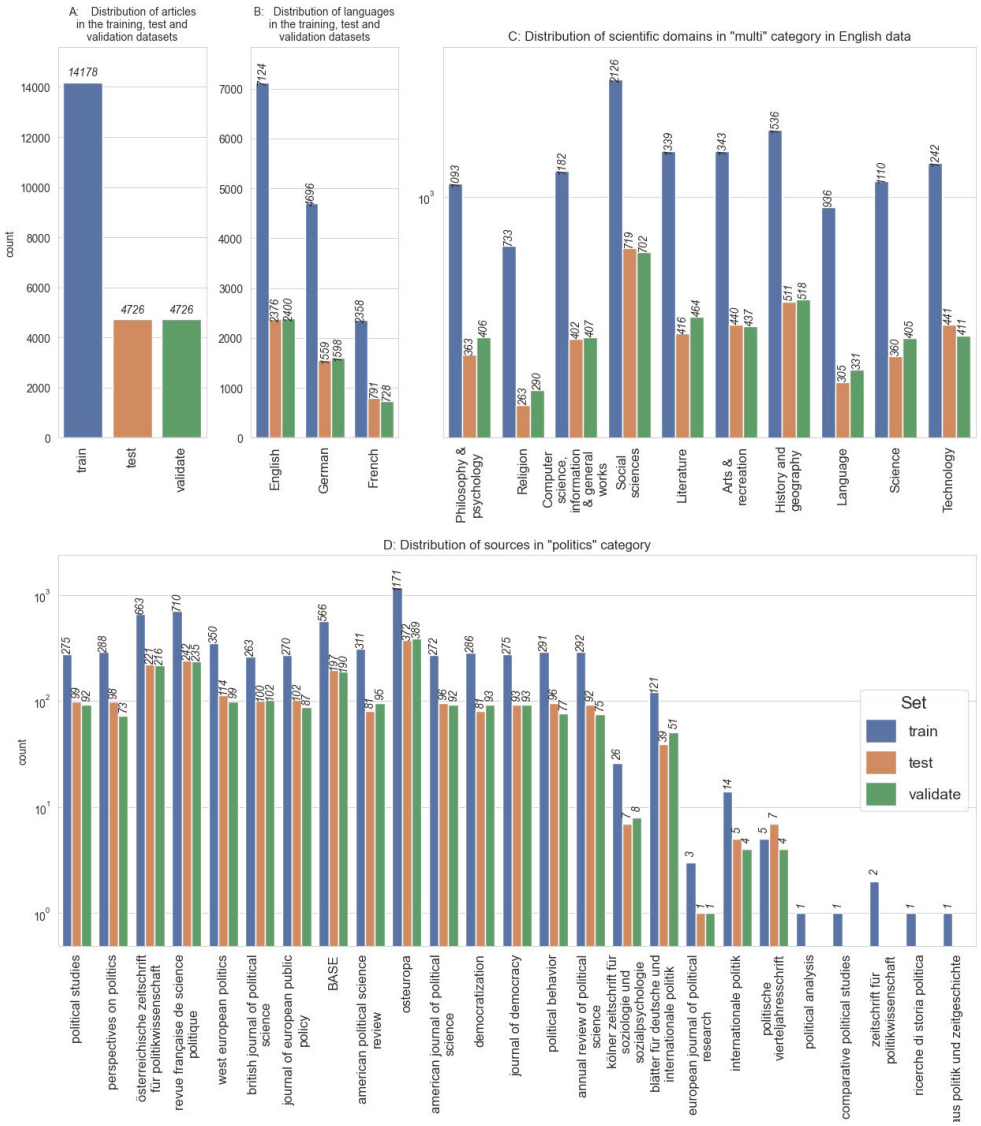

Figure 4: Distribution of training data in the multilingual corpus

3. Evaluation

We evaluated keyword-based approach and classification models using an annotated dataset of 4,726 abstracts from different scientific domains. The weighted keyword-based approach was tested in two modes: applied to title, keywords and abstract, and applied only to title and keywords. The SSciBERT_politics model was evaluated on a separate dataset consisting of 973 abstracts written in English.

Table 2 shows an evaluation[22] of the approaches by class for all languages. Keyword-based approach, applied to title, abstract and keywords showed a high accuracy of 0,87, as well as a high F1-score for both classes: multi and politics. The keyword-based approach applied only to titles and keywords showed a slightly deteriorating performance of 0,8. Class politics has high precision and low recall, at the same time class multi has high recall and low precision[23].

Both BERT-based classifiers showed high total accuracy and F1-scores for all classes. The bert-base-ml-politics showed higher accuracy (0.978) than the SSciBERT_politics model (0.9).

Table 2: Evaluation by class

approach	label	precision	recall	f1-score	support	accuracy
keyword filter (title, abstract, keywords)	politics	0.871	0.846	0.859	2143	**0.874**
	multi	0.876	0.896	0.886	2583	
keyword filter (title, keywords)	politics	***0.947***	***0.593***	0.729	2143	**0.800**
	multi	***0.742***	***0.973***	0.842	2583	
bert-base-ml-politics	politics	0.975	0.978	0.976	2143	**0.978**
	multi	0.981	0.979	0.980	2583	
SSciBERT_politics	politics	0.889	0.902	0.895	460	**0.900**
	multi	0.911	0.899	0.905	513	

Table 3 presents the evaluation results by class and language, which align with the findings in Table Table 2. The bert-base-ml-politics model demonstrates the best performance across all languages. Conversely, the keyword-based approach applied to titles and keywords exhibits low performance for French, with an F1-score of 0.343. However, French had the least amount of training and evaluation data within the entire corpus, potentially introducing bias into the evaluation for all algorithms.

22 We used precision, recall, f-1 score and accuracy as performance evaluation metrics. Precision evaluates the classifier's ability not to label a negative sample as positive. Recall evaluates the classifier's ability to identify all relevant positive instances within the dataset. Support represents the frequency of each class in the reference dataset. The f-1 score is a harmonic mean of precision and recall, providing a balanced measure of both metrics. The precision, recall and f-1 score were computed using the Scikit-learn library (https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.precision_recall_fscore_support.html). Accuracy reflects the proportion of all correct classifications, whether positive or negative and was calculated using the Scikit-learn library (https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.accuracy_score.html).

23 Class politics describes records from the political science domain. Class multi comprises records from other not political science domains.

Table 3: Evaluation by class and language

approach	label	language	precision	recall	f1-score	support
keyword filter (title, abstract, keywords)	politics	English	0.931	0.839	0.883	1212
	multi		0.848	0.936	0.890	1164
	politics	German	0.843	0.861	0.852	783
	multi		0.856	0.838	0.847	776
	politics	French	0.647	0.831	0.728	148
	multi		0.958	0.896	0.926	643
keyword filter (title, keywords)	politics	English	0.976	0.568	0.718	1212
	multi		0.686	0.985	0.809	1164
	politics	German	0.943	0.699	0.803	783
	multi		0.759	0.957	0.847	776
	politics	French	0.625	0.236	0.343	148
	multi		0.846	0.967	0.903	643
bert-base-ml-politics	politics	English	0.989	0.993	0.991	1212
	multi		0.992	0.989	0.991	1164
	politics	German	0.952	0.958	0.955	783
	multi		0.957	0.951	0.954	776
	politics	French	0.979	0.959	0.969	148
	multi		0.991	0.995	0.993	643

4. Conclusion

In general, the BERT-based classifier demonstrated the highest accuracy, but at the same time the highest processing time, approximately 4.5 seconds (on a CPU) to process a single abstract. Conversely, the keyword filter, although slightly less accurate, exhibited significantly faster performance compared to the classification model (less than 1 second to process a single abstract). The efficacy of the keyword filter improves when applied to all available metadata.

Both methodologies are applicable for filtering political science data. The keyword-based approach is particularly advantageous when dealing with large datasets and limited computational resources for running complex models. Furthermore, this approach is also suitable for processing articles with incomplete metadata.

The main limitation of the method is a lack of manually annotated training data. To address this issue, we applied a semi-automated labelling approach. The proposed filtering approach can be applied for filtering metadata from other scientific domains and therefore improve the overview of the domain-related literature and facilitate efficiency in research.

In the previous version of our filtering pipeline for the BASE dataset, we employed an English classification model. Using this pipeline, we filtered 2,318,541 records related to the political science domain from a total of 96,615,560 records provided by BASE. Of these, 68,307 articles already had an assigned Dewey Decimal Classification (DDC). The remaining articles were filtered based on additional metadata features, such as article type, repository and language. From the remaining 24,815,977 records, 1,751,037 were filtered using a keyword-based approach, and 499,197 (from the remaining 23,064,940 records) were filtered using the English classification model.

The multilingual classification model has been integrated into the updated version of our filtering pipeline for the BASE dataset. For several reasons, we expect it to deliver more relevant records than the previous version. First, the pipeline is now adapted for languages other than English. Second, our tests indicate that the multilingual classification model achieves higher overall accuracy than the English classification model. Lastly, we have expanded the keyword list in this version compared to the previous version.

Nina Smirnova, GESIS – Leibniz Institute for the Social Sciences, Department: Knowledge Technologies for the Social Sciences (KTS), Cologne, Germany, ORCID: https://orcid.org/0000-0002-3177-3554

Citable Link (DOI): https://doi.org/10.5282/o-bib/6061