https://doi.org/10.51408/issi2025_001

Text-based Classification of All Social Sciences and Humanities Publications Indexed in the Flemish VABB Database

Cristina Arhiliuc[1], Raf Guns[2], Tim C. E. Engels[3]

[1] *cristina.arhiliuc@uantwerpen.be,* [2] *raf.guns@uantwerpen.be,* [3] *tim.engels@uantwerpen.be*
Centre for Research and Development Monitoring (ECOOM), University of Antwerp,
Middelheimlaan 2, 2020 Antwerp (Belgium)

Abstract

This research describes and evaluates a new methodology for classifying peer-reviewed publications based on the textual metadata available. The methodology is developed for application to the Flemish database for Social Sciences and Humanities (VABB-SHW) and could also be applied in similar databases. To build the classification model, we fine-tune the SSCI-SciBERT model with textual features of journal articles (journal titles, publication titles and abstracts) from Web of Science corresponding to the time period 2000-2022 that is covered by VABB-SHW. We experiment with different feature combinations to replicate the lack of abstracts or the publication channel for a proportion of publications in the target dataset. We conclude that the combined model, trained to handle various combinations of textual features, achieves similar results to feature(s)-specific models, while being more convenient to use. Then, to be able to apply the fine-tuned SSCI-SciBERT to the multilingual VABB-SHW dataset, we translate its data to English using gpt-4o-mini. As the VABB-SHW data is mostly unlabelled at the publication level and covers more publication types than the training dataset, we conduct a separate evaluation for the quality of the classification at the publication type level both by using the prior existing classification (for books and book chapters with generic names) and by comparing it with a manually classified sample of the data and evaluating the quality of the model classification. The model achieves a F1-score of 55% on the VABB-SHW test dataset, with publication type an impacting factor.

Introduction

The goal of this research is to propose a new method for the paper-level, text-based multilabel classification of research publications. The proposed approach is applied to the VABB-SHW database, which stores publications (co-)authored by researchers from Social Sciences and Humanities departments at Flemish universities. The models developed through this method can also be used for the classification of other scholarly and scientific texts. Moreover, this paper specifically examines how well data from journal articles transfers to other types of publications, namely conference proceedings, books, and book chapters.

National bibliographic databases have been created in several countries and regions to offer a comprehensive resource for studying and monitoring the research publications produced in a country or region (Sīle et al., 2018). Among other fields, such databases are especially relevant in the Social Sciences and Humanities. These fields are in their nature and research tradition more locally anchored and typically less well-covered in international citation indexes (Archambault et al., 2006; Sīle et al., 2017, 2018; Sivertsen, 2016; Sivertsen & Larsen, 2012). Although national bibliographic databases are usually more comprehensive, due to their local coverage they often lack citation information, which precludes classifying individual papers according to discipline making use of their positioning in the citation network as it is often done for paper-level classification using Web of Science (Perianes-Rodriguez & Ruiz-Castillo, 2017; Waltman & van Eck, 2012). Hence here we rely on natural language processing of the textual metadata of the publications to classify them to disciplines.

The VABB-SHW database has been implemented in 2008 to complement the Web of Science (WoS) data, which has a low coverage in the Social Sciences and Humanities with the purpose of implementing a fairer performance-based research funding system (Verleysen et al., 2014). The original classification in the database is an organizational classification, i.e. a classification that labels each paper with the discipline(s) of the unit(s) of its (Flemish) authors. This classification gives information about *who* writes the papers from the database. Later, a new classification has supplemented the organizational one: the cognitive channel-based classification that assigns to each publication the discipline(s) of the journal, conference proceeding, book or book series that it originates from (Guns et al., 2018). This cognitive classification provides information regarding where the publications written by Flemish SSH researchers are published. Finally, the paper-level classification presented in this paper supplements the existing two classifications and provides a more fine-grained classification of all the publications included in the VABB-SHW. It answers the question "what *disciplines* do the SSH researchers in Flanders contribute to?".

To train a model for the classification task, we require labelled data, which is not available at the publication level in the VABB-SHW database. Therefore, we use WoS data to train and evaluate different model configurations before applying them to our local database, relying on the classification of references of a paper in WoS to infer the final ground truth. While several studies have identified issues with the accuracy and consistency of WoS classifications (Aviv-Reuven & Rosenfeld, 2023; Milojević, 2020; Singh et al., 2020; Wang & Waltman, 2016) they also acknowledge that WoS remains one of the more reliable options for large-scale classification tasks. At an aggregate level, we consider it to provide a sufficiently robust foundation for this research. The WoS Science categories are mapped to an extended version of the OECD FoRD classification scheme (OECD, 2015) as this scheme is used for all classifications in the system.

This paper builds on our previous work in which we explored appropriate models for text-based classification of publications (Arhiliuc et al., 2024). On the basis of those previous findings, we select the SSCI-SciBERT model (Shen et al., 2022).

Throughout this paper we answer the following questions:

1. Which ground truth labelling strategy represents the data the best, while keeping the distribution of the number of labels to what we are currently expecting in our database?
2. Which strategy, accounting for the varying availability of distinct textual features, yields the best classification results?
3. How well does the knowledge extracted through model fine-tuning from WoS journal articles transfer to non-WoS articles and to other publication types?

In the following parts we first introduce the data, both the WoS data used for model fine-tuning for the classification task and the final application – the VABB-SHW data. Secondly, we explain the methodology and the evaluation procedure for the models. Thirdly, we present the results. We end with conclusion and discussion of the overall implications of this research and further work to be done.

Data description

This project uses two datasets. First, due to the unlabelled nature of the VABB-SHW data, Web of Science (WoS) data has been used to fine-tune the models for the task of classification of the scientific literature. Then, the pretrained models evaluated on the WoS data are applied on the local VABB-SHW database. This section describes the characteristics of both datasets.

WoS data

Web of Science is an international database that indexes peer-reviewed publications and provides extensive metadata. This includes publication titles, years, channels (e.g., journals, conference proceedings, books, or book series), disciplines (referred to as "science categories") assigned at the channel level, and citation information. In this study, we use data from three WoS indices - the Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), and Arts & Humanities Citation Index (AHCI).

We have previously run classification experiments on WoS data for the year 2022 (Arhiliuc et al., 2024). However, the disciplines from Social Sciences and Humanities were often underrepresented, which might have caused worse results overall for them. Due to the nature of the current classification task, where we aim to classify all publications written by researchers from Social Sciences and Humanities included in the VABB-SHW, it is essential to have a good coverage of those fields. We have therefore extended the dataset to include publications from multiple years in the range of years 2000-2022 represented in VABB-SHW. More precisely, we have extracted all the journal articles indexed in the Web of Science from the years 2002, 2006, 2010, 2014, 2018, and 2022. This dataset contains 7 973 222 publications.

The subject categories from WoS are then mapped to OECD FoRD categories (OECD, 2015) with an extension at the level of humanities: "History and archaeology" is split into "History" and "Archaeology", "Languages and literature" into "Languages and linguistics" and "Literature", and "Philosophy, ethics and religion" into "Philosophy and ethics" and "Religion" as this is the classification used in VABB-SHW. Three other disciplines are however excluded due to not being present in the mapping scheme: "Other natural sciences", "Other medical sciences", "Agricultural biotechnology". Additionally, multiple science categories marked as multidisciplinary in the WoS classification are mapped to "Multidisciplinary" discipline that is however less relevant when classifying at the publication level than at the channel level, so it will therefore not be used in this study.

VABB-SHW data

For this study, we are using the 14th edition of the VABB-SHW database that contains publications written by scholars from SSH departments from Flemish universities in the years 2000-2022, both peer-reviewed and not peer-reviewed. The metadata used for this research includes the publication title, abstract, and channel title. The channel depends on the publication type: journals for journal articles, conferences for conference proceedings papers, books for book chapters, and book series for authored or edited books.

Table 1 presents the specificities for each of the five publication types available in VABB-SHW. However, for the purpose of evaluation, they have been grouped into three groups based on their characteristics:

1. Journal articles are conference proceedings – are characterized by a higher availability of the abstracts in the database, more general channel titles and specific publication title
2. Books as author and books as editor ultimately represent the same entity type: books and have been grouped as such
3. Book chapters – can be both specific and general and normally make sense mostly in combination with the channel title.

Table 1. Availability of textual features and publication language for different publications types in VABB-SHW 14.

Type	*Count*	*Abstract*	*Channel title*	*Publication title*	*English*	*Dutch*
Journal articles	170 418	70 869 (41.59%)	170 340 (99.95%)	170 418 (100%)	**61.39%**	33.59%
Authored books	16 295	3 318 (20.36%)	5 838 (35.8%)	16 295 (100%)	25.04%	**64.04%**
Edited books	11 843	1 824 (15.40%)	6 297 (53.17%)	11 843 (100%)	45.88%	42.77%
Book chapters	74 071	9 212 (12.44%)	74 043 (99.96%)	74 071 (100%)	45.15%	41.77%
Conference proceedings papers	12 851	6 187 (48.14%)	12 849 (99.98%)	12 851 (100%)	**83.10%**	9.00%

Table 1 highlights several key characteristics of the available textual features in the dataset. The publication title is fully available across all publication types. The channel title (i.e., journal, conference proceedings, book, book series) is available, with some exceptions, for journal articles, conference proceedings papers and book chapters, but is less commonly available for books (both authored and edited). Abstracts, as previously mentioned, are primarily associated with journal articles and conference proceedings. Moreover, conference proceedings papers and journal articles are mostly in English, while books as author are mostly in Dutch and edited books and book chapters have similar numbers for English and Dutch with a smal share of publications in other languages. These differences may lead to variations in the quality of classification.

Methodology

The current research has two main parts.

The first part uses the labelled WoS data to search for the right model structure and configuration to fit our classification requirements. The requirements are based on similar previous tasks and the characteristics of the VABB-SHW data: multilabel classification, in preponderantly one to three disciplines, able to provide optimal results based on the availability of the textual data representing a publication.

The second part focuses mainly on the VABB-SHW and covers the preparation of the VABB-SHW data for the classification, the application of the strategy designed in the first part and the evaluation of the classification.

Part 1: Model selection

Thresholds

Determining relevant ground-truth labels for the WoS data is fundamental for this research. The ground-truth classification for a specific publication is deduced from the distribution of disciplines in the reference list. However, this raises the question: what proportion of the references of a paper should be in a specific discipline to assume that the discipline is representative of the content of the paper?

In the ECOOM-Biblio-Antwerp team that is responsible for the maintenance and analysis of the VABB-SHW database, we have an annual task of manual classification at a journal, conference and book level. This is done to enrich the existing channel-based cognitive classification when no data regarding those channels has been automatically found in external sources. One of the guidelines for that task is limiting the number of disciplines to a maximum of three. Based on that, in a previous study of classification methods for journal articles (Arhiliuc et al., 2024), we have selected the threshold of 0.3 as most publications get classified in up to three disciplines with relatively few publications being classified in no discipline or more than five disciplines. In this study however, we aim on a more methodical analysis of the appropriate threshold that is going to happen in two steps:

1. Analysis of the distribution of the number of disciplines assigned to publications using thresholds varying from 0 to 1 with a 0.05 interval between them. The Multidisciplinary discipline to which multiple multidisciplinary science categories map, has been removed from this analysis, resulting in a few outliers having 0 disciplines even at threshold 0. The goal of this analysis is to select the thresholds that position most publications in 1 to 3 disciplines, which is what we are aiming for. More precisely, we are looking for the thresholds that have more than 90% of the publications in 1 to 3 disciplines to maximize the number of publications available for the creation of the train, validation and test datasets.

2. As a proxy of how representative the labels are of the data, train, validation and test datasets are created for each of the selected thresholds and then SSCI-

SciBERT is assigned with the task of classifying the publication into disciplines based on their abstracts. Small variations in the results among thresholds should not be viewed as significant as due to the variation in number of disciplines per publication for each threshold, the datasets are distinct among thresholds, which can have an impact on the result.

The optimal threshold is selected based on the distribution of number of disciplines and the F1-score on the test datasets in the second step.

Data partitioning

For all the experiments in this part, the train, validation and test datasets are selected to be as balanced as possible across disciplines given the multilabel nature of the classification. More specifically, we aim to select 500 examples per disciplines for the test dataset, 500 examples per discipline for the validation dataset and 10 000 examples for the train dataset if available.

To test various model configurations after the choice of the threshold (see *Thresholds*), we partitioned the data into separate train, validation, and test sets, ensuring no overlap of journals across the three sets to prevent leakage when using the journal names. Due to data availability challenges in certain Social Sciences and Humanities disciplines and the constraints of this partitioning approach, we prioritized maximizing the number of publications in the training set for underrepresented disciplines. To achieve this, we allocated publications from the least represented journals to the test and validation sets, avoiding the placement of journals with large numbers of examples in these smaller sets, where many examples would go unused. While this approach ensures an efficient use of available data for training, it reduces the randomness of partitioning.

Moreover, a second drawback of this method must be considered: by distributing journals among the three datasets, it is possible that no set fully captures the diversity of the disciplines, as distinct journals might focus on different aspects of the field.

Additionally, if the goal is to classify new publications, having a greater variety of journals in the training set could enhance classification quality, as the model benefits from learning discipline-specific patterns associated with that journal. Therefore, while datasets with no journal overlap across the three sets provide an opportunity to test how well the journal name represents a publication, ensuring a higher diversity of journals in the training set is a more effective approach to improving classification performance.

We provide our results for experiments on data partitioned with the constraint of distinct journals across datasets and without this constraint.

Choice of model configuration

As shown in Table 1, the resulting model should be able to work on different configurations of textual features. There are two possible approaches to achieve this. In the first approach, separate models could be built for each feature and combination of features. A meta-model would then determine, based on the textual data available for the instance to be classified, which of these models should be applied to achieve the best performance. In contrast, the second approach involves training a single unified model on various combinations of publication textual features. This unified model is designed to handle any combination of the three textual features as input. The second approach offers the advantage of being more compact and easier to use. However, it is assumed that the first approach might perform better on specific features since each model is exclusively trained on its corresponding configuration. In the results section, these two approaches will be compared, alongside the individual performance of each model.

The models will be evaluated using precision, recall, and the F1-score, which is the harmonic mean of precision and recall. While we have aimed to create a relatively balanced test set, perfect balance cannot be ensured in a multilabel scenario. As a result, we focus on macro scores (calculated as the average of class-wise metrics) rather than micro scores (calculated for the dataset as a whole). This ensures that performance is assessed at the level of individual disciplines, rather than being influenced by the potentially higher representation of certain disciplines in the dataset.

Part 2: Application to VABB-SHW

Translation

For this research, we opted to translate all non-English VABB-SHW publications into English to simplify the problem. We used the GPT-4o-mini model for this task. Although no studies have yet evaluated the quality of translation done by the GPT-4o-mini model, findings from the shared task in translation from the Workshop on Statistical Machine Translation (Kocmi et al., 2024) show promising results for its predecessor, GPT-4, positioning it as the top performing model for English-German translation quality (German is the language closest to Dutch from the list) based on human evaluation. Hendy et al., 2023 evaluated another one of its predecessors, GPT-3.5 (text-davinci-003), on translation tasks in comparison with other existing models and software. Some of the main conclusions are that translations produced by GPT are more fluent, achieving consistently lower perplexity and more non-monotonic, producing translations with longer range reordering. However, the authors have also noted that given that the models are not specialized in translation or trained on parallel texts in multiple languages, LLMs are less constrained in their faithfulness to the source text compared to translation-specialized models.

Nevertheless, we consider that for the task at hand a fluent, context-appropriate translation of the proposed text is sufficient to extract information regarding the discipline affiliation. Moreover, given previous comparisons of the GPT models on other tasks, we expect GPT-4o-mini to achieve superior results to its predecessors.

Evaluation of the classification

The methodology for evaluating the classification depends on the classification type. We combine automated testing with manual testing to estimate how reliable the database classification is at an individual publication level.

For evaluation of book classification, we use the existing classification based on international databases and manual classification at the level of book. In total, 53.01% of books already have a classification in the database .
A subset of the conference proceeding papers and the journal articles are classified manually by a member of our team with no prior access to the models' classification. The subset is selected based on previous classification experiments such that 0.10% of publications for each discipline are in the sample, but not less than five, in total 554 publications. The annotator has received a shuffled version of the data with no prior knowledge of how it has been selected.
A similar procedure is applied to a portion of the data for book chapters, with 0.30% of publications for each discipline included in the sample, again with a minimum of five publications per discipline, summing to 457 publications. This approach aims to preserve the supposed distribution of disciplines in the dataset, with a minimum representation for all.
Another part of the evaluation of book focuses on chapters with generic names, defined as instances where more than 15 book chapters share the same name. These chapters are expected to should be classified the same as the originating book and are excluded from the manual classification sample. The top 10 most frequent book chapter names are shown in Table 2. These names are typically variations of generic book sections (e.g., introductions, conclusions) or chapters about Belgium.

Table 2. Top 10 most frequent book chapter names.

Chapter title	*English translation*	*Count*
Introduction	Introduction	735
Inleiding	Introduction	235
Belgium	Belgium	205
Preface	Preface	148
Voorwoord	Foreword	125
Woord vooraf	Foreword	112
Foreword	Foreword	82
Conclusion	Conclusion	45
Préface	Preface	43
Ten geleide	Introduction	31

For this part of the analysis, since the test data partially reflects the discipline repartition in VABB-SHW for the specific publication type, we will focus on micro metrics (micro-precision, micro-recall, micro-F1). This approach aligns with our interest in evaluating the model's overall performance on the entire sample rather than its performance at the discipline level.

Results

Threshold analysis

As outlined in the methodology, the threshold selection is done in two steps: first, candidate thresholds are identified based on the distribution of labels, and second, the final threshold is selected for the model based on classification results with abstracts.

Table 3 presents the distribution of the number of labels for thresholds ranging from 0.0 (a discipline is assigned to a publication if any referenced publication is classified into that discipline in WoS) to 1.0 (a discipline is assigned only if all referenced publications are classified into that discipline in WoS). The thresholds 0.25 to 0.55 respect the constraint of having more than 90% of the publications into one to three labels, thus they are retained for further testing.

Table 3. Distribution of the number of disciplines per publication across different thresholds.

Threshold	*0 labels*	*1 label*	*2 labels*	*3 labels*	*4 labels*	*5+ labels*	*Share with 1-3 labels*
0.0	1 932	719 570	918 916	1 151 588	1 220 675	3 960 541	34.99
0.05	1 935	960 321	1 278 720	1 717 112	1 504 943	2 510 191	49.62
0.1	1 954	1 396 688	1 851 380	2 074 124	1 408 808	1 240 268	66.75
0.15	2 088	1 833 008	2 325 294	2 102 251	1 123 273	587 308	78.52
0.2	2 896	2 363 844	2 746 408	1 886 007	750 042	224 025	87.75
0.25	6 204	2 928 016	2 987 144	1 520 333	448 764	82 761	**93.26**
0.3	14 186	3 472 914	3 025 206	1 153 258	270 289	37 369	**95.96**
0.35	43 491	4 121 897	2 858 108	787 419	148 342	13 965	**97.42**
0.4	110 930	4 789 151	2 486 922	497 389	82 194	6 636	**97.49**
0.45	217 847	5 273 439	2 087 337	337 452	52 578	4 569	**96.55**
0.5	481 774	5 767 008	1 517 250	180 793	25 125	1 272	**93.63**
0.55	692 572	5 894 650	1 233 191	133 630	18 125	1 054	**91.07**
0.6	1 060 689	5 922 292	893 183	84 780	11 544	734	86.54
0.65	1 412 112	5 804 604	686 108	61 275	8 475	648	82.17
0.7	1 869 712	5 579 591	480 314	37 771	5 398	436	76.48
0.75	2 361 706	5 248 435	334 441	24 724	3 564	352	70.33
0.8	2 811 153	4 895 797	245 374	18 007	2 577	314	64.71
0.85	3 234 406	4 533 138	188 938	14 419	2 021	300	59.41
0.9	3 723 689	4 095 108	140 776	11 678	1 684	287	53.27
0.95	4 172 587	3 672 019	116 039	10 716	1 577	284	47.64
1.0	5 282 625	2 669 322	21 248	27	0	0	33.75

Table 4 shows the macro scores for each threshold on the threshold's test data. With the exception of 0.25, the values tend to peak at 0.5 and then start going down. Threshold 0.25 is notable for its higher representation for Other Humanities and Health Biotechnology, which are otherwise significantly underrepresented for the other thresholds and often with a F1-score of 0. Excluding these two disciplines would result in similar values between 0.25 and 0.5.

For the next part, the results with the 0.5 threshold are presented. However, for the classification analysis of VABB-SHW, the results with both models are tested to reverify which is the more accurate model.

Table 4. Classification results for different thresholds.

Threshold	0.25	0.30	0.35	0.40	0.45	0.50	0.55
Macro recall	82.94%	75.90%	75.16%	76.99%	76.09%	76.46%	76.92%
Macro precision	73.34%	71.52%	71.73%	71.28%	71.33%	72.81%	71.28%
Macro F1-score	**76.48%**	73.33%	73.15%	73.78%	73.38%	**74.27%**	73.72%

Results for WoS data

First, we evaluate the impact of journal names (channel title) on the quality of the classification. This is done by using distinct journals for the train, validation and test dataset as explained in the Methodology section. The results are presented in Table 5.

The journal name is a poor predictor of the discipline of the publication (7.44 % macro F1-score) and the increase in the quality of prediction when the journal name is added to the article title is insignificant (59.80% macro F1-score for title only and 60.38% for title and journal title). There is in fact a decrease when the journal name is added to the abstract (66.92% macro F1-score for abstract only and 65.95% with abstract and journal title). This result is not surprising given that when only the journal name is used as a feature to predict publication classification, the same journal can have different classifications assigned in the train dataset as the entity classified is the publication, not the journal.

Therefore, as mentioned in the Methodology section, to increase the variety of publications in a discipline, we have decided that for final model selection we ignore this restriction and allow publications from the same model to be present in the train, validation and test database. Modelled like this, the problem is a more realistic representation of the general classification problem studied in this research that should not exclude the benefit given by the presence of the journal in the train database.

Table 5. Classification results for train, validation and test datasets containing distinct journals.

Model data	Macro precision	Macro recall	Macro F1
Abstract only	72.72%	63.91%	66.92%
Channel title + Abstract	71.33%	63.32%	65.95%
Title only	67.79%	55.74%	59.80%
Channel title + Title	68.30%	56.24%	60.38%
Channel title	32.17%	4.47%	7.44%

In general, the discrepancy between the results of the predictions when allowing (Table 6) publications from the same journals in train, validation and test set – whether or not the journal name is used as a feature – compared to when the publications in the three sets come from distinct journals (

Table 5) point towards journal specialization resulting in publications from the journals in the train set being a worse representation of the ones in the test and validation sets when they come from other journals from that discipline.

Table 6. Classification results for the train, validation and test datasets selected with no restriction at the level of journal. (-) marks the models that would not be used for the final classification.

Model data	Macro precision	Macro recall	Macro-F1	Rank
Abstract only	76.94%	72.03%	74.12%	5 (-)
Title only	72.79%	62.87%	66.94%	6
Title + Abstract	76.94%	73.07%	74.77%	3
Channel title + Abstract	77.97%	75.67%	76.63%	2 (-)
Channel title + Title	77.11%	72.28%	74.33%	4
Channel title + Title + Abstract	78.38%	75.81%	76.91%	1
Combined	77.08%	71.78%	74.04%	

Table 6 shows the results of the classification when no restrictions are applied on the channel of the classified article. The table includes results for individual features, combinations of features, and a combined model. The combined model is trained on the merged training data from the other experiments, meaning it includes examples with only abstracts, examples with both abstract and title, examples with only the title, and so on.
A meta-model would need to address 4 possible combinations of features in the VABB-SHW dataset: all the features are available, only the title and the channel title are available, only the title and the abstract are available, and only the title is available. The results in Table 6 indicates that the model that is trained on all the available features should be used for all the scenarios.
Since the training, validation and test datasets for all the previous models consists of the same articles, but with different textual features put forward, the combined data is six times larger than the individual datasets. It includes the same articles six times, but represented by distinct features or combinations of features. The next experiment aims to determine whether building a single model capable of classifying data with different structures results in any loss of prediction quality.
To properly evaluate the combined model, its performance must be tested on the individual test datasets to assess whether it underperforms or overperforms compared to models specialized for specific features or feature combinations. Table 7 presents these results, showing that variations in the F1 score are not significant to conclude that the combined model performs better or worse than the models specialized on a feature or a group of features.
Based on these findings, we focus our further analysis on the combined model, as it can be applied to the VABB-SHW dataset as a whole, even in cases where certain features are missing.

Table 7. Classification results for the combined model when tested on individual features and feature combinations.

Test data	Macro-F1	Comparison Macro
Abstract only	74.21%	+ 0.09%
Title only	67.24%	+ 0.30%
Title + Abstract	74.49%	- 0.28%
Channel title + Abstract	76.51%	- 0.12%
Channel title + Title	74.14%	-0.19%
Channel title + Title + Abstract	76.56%	-0.35%

Results for VABB-SHW

Table 8 shows the results for all the available labelled datasets originating from the VABB-SHW dataset.

Table 8. Classification results for the available labelled VABB-SHW datasets.

		Threshold 0.5			*Threshold 0.25*		
Test set	*Nb. Pub.*	*Micro Precis.*	*Micro Recall*	*Micro F1-score*	*Micro Precis.*	*Micro Recall*	*Micro F1-score*
Manual journal articles and conference proceedings	554	50.25%	58.25%	53.96%	42.65%	65.26%	51.59%
Manual book chapters	457	56.31%	60.11%	58.15%	51.27%	67.91%	58.43%
Book chapters with generic names	339	51.92%	51.92%	51.92%	47.82%	57.14%	52.07%
Books (from previous classification)	14 916	55.14%	55.41%	55.27%	51.63%	61.98%	56.33%
Total	16 266	54.90%	55.59%	55.25%	51.11%	62.19%	56.11%

For the manual classification, and book classification datasets, the results consistently achieve an F1-score of 54–58%. However, book chapters with generic names score lower, likely due to the noise introduced by the chapter name and the overall lack of sufficient textual data for accurate classification. When comparing the 0.25 threshold with the 0.5 threshold, the former gains in recall but loses in precision. This is because the 0.25 threshold predicts a larger number of labels.

To further understand the classification results, Table 9 presents the outcomes for the top 10 most represented disciplines in the total VABB-SHW test dataset (the combination of all test datasets for VABB-SHW), including the results of the combined model on the WoS test data. Disciplines that are easily identified in VABB-SHW, such as Law and Language and Linguistics also achieve good results on WoS data. In contrast, History, Art, and Sociology underperform on both test datasets, with Sociology proving particularly challenging for the model to classify accurately.

Economics and Business, Philosophy and Ethics, and Political Science are notable cases. While these disciplines perform well on WoS data, they underperform on VABB-SHW data. This discrepancy may indicate that the training data does not adequately represent these disciplines as they appear in VABB-SHW. Alternatively, given that the book dataset is the largest in the test datasets, the definition of these disciplines, as inferred from journal articles, may not translate well to other publication types.

To investigate this further, Table 10 presents the results for these disciplines in the individual test datasets. The findings for manually annotated datasets outperform those for the total test set. Additionally, differences between the dataset containing journal articles and conference proceedings and the one with book chapters suggest that publication types significantly impact classification performance. Furthermore, the differences between the manually annotated datasets and the rest may also be, at least in part, due to variations in the annotation methodology across datasets.

Table 9. Classification results for top 10 disciplines based on the frequency in the total test set for VABB-SHW, threshold 0.5.

Discipline	*# instances in combined test set*	*Precision*	*Recall*	*F1-score*	*F1-score for WoS data*
Law	3 532	89.47%	79.16%	84.00%	89.34%
Literature	1 673	52.35%	67.12%	58.83%	67.60%
History	1 524	37.06%	54.00%	43.95%	62.24%
Sociology	1 395	52.60%	21.79%	30.82%	48.70%
Languages and linguistics	1 344	85.33%	61.021%	71.15%	83.75%
Economics and business	1 262	67.73%	45.56%	54.48%	74.26%
Art	1 217	58.12%	55.88%	56.98%	66.61%
Religion	1 199	78.59%	48.37%	59.89%	70.95%
Political Science	1 041	43.35%	56.00%	48.87%	75.53%
Philosophy and ethics	914	46.64%	56.24%	50.99%	77.54%

Table 10. Classification results for Economics and business, Political Science and Philosophy and ethics across different VABB-SHW test datasets.

Dataset	*Metric*	*Economics and business*	*Political Science*	*Philosophy and ethics*
Manual journal articles and conference proceedings	*# instances*	61	39	17
	F1-score	66.10%	54.55%	57.89%
Manual book chapters	*# instances*	57	43	25

	F1-score	58.59%	55.56%	74.51%
Book chapters with generic names	*# instances*	31	13	25
	F1-score	41.86%	43.90%	51.52%
Books (from previous classification)	*# instances*	1 113	946	847
	F1-score	53.81%	48.57%	50.19%
WoS data	*F1-score*	74.26%	75.53%	77.54%
Total VABB-SHW test data	*# instances*	1 262	1 041	914
	F1-score	54.48%	48.87%	50.99%

Conclusion

This research presents a methodology for classifying publications from local databases based solely on textual information. We divided the analysis into two parts: one focused on building the model, and the other on applying it to classify the publications included in the Flemish database for Social Sciences and Humanities (VABB-SHW).

In the first part, we investigated which ground truth strategy best represents the data while maintaining an optimal number of disciplines per publication. The range for the optimal threshold was narrowed to 0.25–0.55. Based on classification results across various thresholds, we selected the 0.5 threshold for further analysis of how to address the availability of different textual features. However, given the promising results of the 0.25 threshold, it was also considered for the VABB-SHW data.

Additionally, we evaluated two strategies to address the potential lack of certain textual features in the VABB-SHW data. The first strategy involved using a meta-model that selects among feature-specific models, while the second proposed a single model trained on various textual features and feature combinations to handle varied input. The results showed similar performance for both strategies, and we opted for the combined model due to its ease of application.

When analyzing the classification results on VABB-SHW, we observed significantly worse performance on the VABB-SHW test dataset compared to the WoS test dataset. One identified factor contributing to this discrepancy is the publication type.

Discussions and Limitations

Other factors, such as the availability of textual features, translation errors, local terminology, and specific topics, may also contribute to the observed discrepancies between the results for VABB-SHW and WoS. We have currently not yet explored these aspects in detail but this could provide valuable insights in future research.

While this research has provided overall metrics for classification performance, it has not qualitatively analysed the nature of the classification errors. Future work could involve examining disciplines that are frequently misclassified and investigating whether errors stem from true misclassification or differences in interpretation. Given the absence of an incontestable ground truth for discipline classification and the fact that some publications lie at the intersection of multiple disciplines, some errors may involve such borderline cases.

This study has certain limitations that should be considered while interpreting the results. First, the methodology relies on the classification of references in WoS to infer the final ground truth. Consequently, the model is trained to predict the disciplines associated with the journals most commonly cited by the publication, using this as a proxy for the discipline of its content.

Secondly, we assume that the selected classification scheme accurately represents the underlying structure of the data and that the model can effectively learn to distinguish each discipline based on the provided examples. However, this assumption has not yet been empirically tested, as the classification scheme was chosen based on its alignment with other types of classification in the database rather than its specific suitability for the data.

Thirdly, the evaluation was conducted on a small sample of VABB-SHW publications, which may not fully capture the diversity of the dataset, especially for journal articles, conference proceedings, and book chapters. Expanding this sample in future research would provide a more comprehensive understanding.

Fourthly, the data for non-SSH disciplines in VABB-SHW consists of publications (co-)authored by Flemish researchers from SSH departments. As a result, this content may deviate slightly from the typical literature in those fields. Exploring this aspect further could shed light on its potential impact.

Finally, the study assumes that disciplines are static over time, which has been shown by previous research (Manning, 2020; Zhou et al., 2022) to be an oversimplification. While the time dimension was not explicitly accounted for in this analysis, its potential influence represents an interesting direction for future exploration.

Acknowledgments

We want to thank our colleague, Eline Vandewalle, for the manual annotation of the data.

References

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingrasb, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329–342. https://doi.org/10.1007/s11192-006-0115-z

Arhiliuc, C., Guns, R., Daelemans, W., & Engels, T. C. E. (2024). Journal article classification using abstracts: A comparison of classical and transformer-based machine learning methods. *Scientometrics*. https://doi.org/10.1007/s11192-024-05217-7

Aviv-Reuven, S., & Rosenfeld, A. (2023). A logical set theory approach to journal subject classification analysis: Intra-system irregularities and inter-system discrepancies in Web of Science and Scopus. *Scientometrics*, *128*(1), 157–175. https://doi.org/10.1007/s11192-022-04576-3

Guns, R., Sīle, L., Eykens, J., Verleysen, F. T., & Engels, T. C. E. (2018). A comparison of cognitive and organizational classification of publications in the social sciences and humanities. *Scientometrics*, *116*(2), 1093–1111. https://doi.org/10.1007/s11192-018-2775-x

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation* (arXiv:2302.09210). arXiv. https://doi.org/10.48550/arXiv.2302.09210

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., … Zouhar, V. (2024). Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 1–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.1

Manning, P. (2020). Disciplines and Their Evolution. In P. Manning (Ed.), *Methods for Human History: Studying Social, Cultural, and Biological Evolution* (pp. 83–90). Springer International Publishing. https://doi.org/10.1007/978-3-030-53882-8_8

Milojević, S. (2020). Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines. *Quantitative Science Studies*, *1*(1), 183–206. https://doi.org/10.1162/qss_a_00014

OECD. (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/science-and-technology/frascati-manual-2015_9789264239012-en

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2017). A comparison of the Web of Science and publication-level classification systems of science. *Journal of Informetrics*, *11*(1), 32–45. https://doi.org/10.1016/j.joi.2016.10.007

Shen, S., Liu, J., Lin, L., Huang, Y., Zhang, L., Liu, C., Feng, Y., & Wang, D. (2022). SsciBERT: A pre-trained language model for social science texts. *Scientometrics*. https://doi.org/10.1007/s11192-022-04602-4

Sīle, L., Guns, R., Sivertsen, G., & Engels, T. (2017). *European Databases and Repositories for Social Sciences and Humanities Research Output*. https://doi.org/10.6084/M9.FIGSHARE.5172322

Sīle, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T. C. E., Arefiev, P., Dušková, M., Faurbæk, L., Holl, A., Kulczycki, E., Macan, B., Nelhans, G., Petr, M., Pisk, M., Soós, S., Stojanovski, J., Stone, A., Šušol, J., & Teitelbaum, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation*, *27*(4), 310–322. https://doi.org/10.1093/reseval/rvy016

Singh, P., Piryani, R., Singh, V. K., & Pinto, D. (2020). Revisiting subject classification in academic databases: A comparison of the classification accuracy of Web of Science, Scopus & Dimensions. *Journal of Intelligent & Fuzzy Systems*, *39*(2), 2471–2476. https://doi.org/10.3233/JIFS-179906

Sivertsen, G. (2016). Patterns of internationalization and criteria for research assessment in the social sciences and humanities. *Scientometrics*, *107*(2), 357–368. https://doi.org/10.1007/s11192-016-1845-1

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, *91*(2), 567–575. https://doi.org/10.1007/s11192-011-0615-3

Verleysen, F., Ghesquière, P., & Engels, T. (2014). *The objectives, design and selection process of the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB-SHW)*.

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347–364. https://doi.org/10.1016/j.joi.2016.02.003

Zhou, H., Guns, R., & Engels, T. C. E. (2022). Are social sciences becoming more interdisciplinary? Evidence from publications 1960–2014. *Journal of the Association for Information Science and Technology*, *73*(9), 1201–1221. https://doi.org/10.1002/asi.24627