Content-based classification of research articles: Comparing keyword extraction, BERT, and random forest classifiers

Cristina Arhiliuc[1] and Raf Guns[2]

[1] cristina.arhiliuc@uantwerpen.be, [2] raf.guns@uantwerpen.be
Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

Abstract

The classification of publications into disciplines has multiple applications in scientometrics – from contributing to further studies of the dynamics of research to allowing responsible use of research metrics. However, the most common ways to classify publications into disciplines are mostly based on citation data, which is not always available. Thus, we compare a set of algorithms to classify publications based on the textual data from their abstract and titles. The algorithms learn from a training dataset of Web of Science (WoS) articles that, after mapping their subject categories to the OECD FORD classification schema, have only one assigned discipline. We present different implementations of the Random Forest algorithm, evaluate a BERT-based classifier and introduce a keyword-based methodology for comparison. We find that the BERT classifier performs the best with an accuracy of 0.7 when trying to predict the discipline and an accuracy of 0.91 for the "real discipline" to be in top 3. Additionally, confusion matrices are presented that indicate that frequently the results of misclassifications are similar disciplines to "real" ones. We conclude that, overall, Random Forest-based methods are a compromise between interpretability and performance, being also the fastest to execute.

Introduction

This research presents a comparison between several approaches for research article classification into one or more disciplines according to the OECD Field of Research and Development (FORD) classification schema (OECD, 2015), based on textual data. For this, we are comparing several machine learning based models as well as a keyword-based one in which keywords are extracted from the articles and further assigned to disciplines. The approaches are evaluated by comparing their performance for a set of unseen articles.

The classification of publications into disciplines is important for many applications in scientometrics and science studies. For instance, classifications are needed for studying interdisciplinarity (Glänzel et al., 2021), for the responsible use of research metrics (Shu et al., 2020), for assessing future directions of research, and for the comparison of research impact across disciplines, as field-normalized scores depend on the correct classification. Although the journal classification models used by WoS and Scopus are convenient, accessible and robust, in cases such as multidisciplinary journals, they provide insufficient information and the individual papers frequently end up being improperly classified (Shu et al., 2019). Moreover, Shu et al. (2019) show that a significant proportion of articles in general are not published in journals of the same discipline. Consequently, a paper-level classification provides more accurate information about the discipline(s) to which a paper is associated. Multiple studies have proposed methods to accurately represent the contents of research publications.

There is a long tradition in scientometrics that uses citation-based data – references, direct citations, co-citations or bibliographic coupling – to infer disciplinary groups of publications. For instance, Waltman & van Eck (2012) are clustering publications based on direct citation relations between publications, and label the clusters based on terms extracted from the title and abstract of publications. Other recent studies include Ahlgren et al. (2020); Klavans & Boyack (2017). However, major citation databases do not cover the Social Sciences and Humanities well (Petr et al., 2021), and more local databases that aim to cover these fields better generally lack citation information. Since we aim to develop a method that can be applied to such local, SSH-oriented databases, we will focus on text-based methods.

Salatino et al. (2019) introduce the CSO Classifier, a classifier used to assign to a research paper multiple relevant research concepts from a pre-designed ontology for Computer Science (Salatino et al., 2018) by analyzing the text at both syntactic and semantic level. This approach comes as an improvement of the STM classifier formerly introduced by Osborne et al. (2016). However, these approaches only cover Computer Science and the development of a complete and clean ontology for another discipline would require a high amount of training data and subsequent reviews. This level of complexity is justified at a more granular level of classification, but might not be required for a discipline classification task where then number of classes is small in comparison.

Eykens et al. (2021) look into using two supervised machine learning algorithms – Multinomial Naïve Bayes and Gradient Boosting – for classifying articles from Social Sciences using textual data. The accuracy of both methods is below 50%, which is partially explained by the granularity of and proximity between the disciplines from Social Sciences. However, such accuracy might not be considered sufficient for a reliable classification of the articles.

Machine learning techniques have been tested in multiple studies on various classification schema. In their paper, Waltinger et al. (2011) are exploring a hierarchical classification of OAI metadata according to Dewey Decimal Classification (DDC) using Support Vector Machine (SVM) algorithm. Their method achieves a f_1 score of 0.81 for the multilabel classification over the 10 base classes, but for the deeper levels, only partial data is available. Other studies are also using DDC as schema for classification (Golub et al., 2018; Waltinger et al., 2011; Wang, 2009), but the results of classification at the first level (10 labels, one of which is “Science”) is usually not very useful and a good classification at second level (100 labels) requires a lot of data.

In their study, Weber et al. (2019) conclude that Neural Network based classifiers perform better than the classic machine learning algorithms, followed by Random Forest classifier on InCites data. However, their study doesn’t include the new transformer-based classifiers and statistical models.

Pech et al. (2022) use a similar keyword-based classification approach to the one introduced in this paper by identifying frequent terms for each subfield and using them to identify to which subfield each paper most likely belongs. However, they apply their method only for Physics and only at a subfield level, while the current study extends to all fields of science at a less granular level.

One of the most challenging aspects of discipline classification is the evaluation, given that no golden standard exists and even subject experts may not always agree (Eykens et al., 2021; Salatino et al., 2019). With the rise of interdisciplinary studies (Morillo et al., 2003; Zhou et al., 2022) and the information flow between disciplines (Urata, 1990), the task of assigning a discipline to a publication becomes more complex as the barrier between disciplines becomes fuzzy. This is addressed in this research by using a controlled dataset both for training and testing with publications assigned to only one discipline. Although this doesn’t guarantee a correct initial assignment of the disciplines, it minimizes the error. The selection of this data is developed in Data section.

In summary, our study has the following aims: to develop keyword-based methods for supervised classification into disciplines and to compare them with established machine learning models. For this, we use a BERT-based classifier and Random Forest classifier using BoW embeddings and TF-IDF embeddings. The main advantage of a keyword-based method is its interpretability – the results of the classification of a publication can be easily explored and explained, contrary to, e.g., deep learning approaches. However, as we will see, the performance of keyword-based methods remains a challenge and for many purposes, ML models may still be preferable.

Data

For the purpose of training and evaluation of the results, this paper uses data from WoS (SCIE, SSCI, AHCI), years 2017-2021. A mapping between WoS subject categories (SC) and FORD has been applied to assign FORD disciplines to the journals from the collection. FORD consists of 5 areas of research (Natural Sciences, Engineering and Technology, Medical and Health Sciences, Agricultural Sciences, Social Sciences and Humanities), divided into 42 disciplines. It is worth mentioning that the mapping is a 1-to-n type of mapping where one FORD discipline may correspond to multiple SCs, but an SC has only one associated FORD discipline.

Only the journals that belong to one FORD discipline are retained. These journals have been sorted for each discipline according to their ranking in the WoS SC obtained using indicators like JIF. After that, in the order of their ranking, we have preserved all the journals until we reach 10 journals with more than 100 articles per discipline. This secures a coverage of at least 1000 publications per discipline and ensures a diversity in term of covered subjects since sorting by SC rank encourages a mix of journals from different SCs associated to the discipline. Multidisciplinary journals have been excluded as they would introduce noise in the training data. Data are collected from the in-house copy of WoS maintained by ECOOM KU Leuven. Only abstracts for articles and reviews are used. For “Other Natural Sciences”, “Nano-technology”, “Other Medical Sciences” no single-disciplinary journal examples have been extracted because of the low mapping to WoS subject categories, so they have been excluded from this classification. Additionally, we have also decided to separate “History and Archaeology” into two distinct disciplines: “History” and “Archaeology”, “Language and literature” into “Language and linguistics” (which has been excluded due to lack of examples) and “Literature”, and “Philosophy, ethics and religion” into “Philosophy and ethics” and “Religion”. Consequently, there are only 41 classes left.

Our classification model is trained on a dataset of 32800 abstracts – 800 for each discipline – and the evaluation is performed on a random sample of 8200 publications (200 per discipline). The balanced training and testing data allows reducing the errors introduced by handling unbalanced data, which is relevant for frequency-based methods.

Methods

Keyword-based method

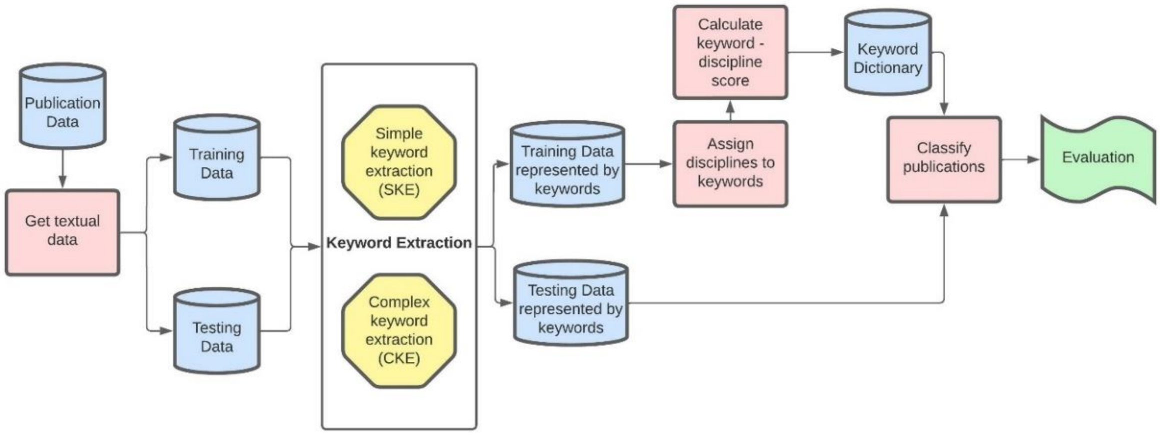
Publication Data
Get textual data
Training Data
Testing Data
Simple keyword extraction (SKE)
Keyword Extraction
Complex keyword extraction (CKE)
Training Data represented by keywords
Testing Data represented by keywords
Assign disciplines to keywords
Calculate keyword - discipline score
Keyword Dictionary
Classify publications
Evaluation

Figure 1. Schematic description of the methodology used for article classification

Figure 1 gives an overview of the methodology used for keyword-based discipline classification. First step is the abstract retrieval (see section Data). Next, the data is separated into training and test data by selecting 800 and 200 random publications for every discipline, without repetition (no publications from the training data are present in the testing data). Further, two methods are used for keyword extraction: simple keyword extraction and complex keyword extraction.

For the simple keyword extraction (SKE), the stop words and numbers have been removed from the abstract data and then the result has been lemmatized using the WordNetLemmatizer from the NLTK library[1]. Lemmatization consists in grouping together inflected forms of a word so they can be analysed as a single term, the word's lemma. It is useful for reducing the number of keywords generated by reducing variations of the same concept. From the obtained texts we extract all the remaining words as individual words and combinations of 2, 3, and 4 words (n-grams).

The complex keyword extraction (CKE) algorithm uses named entity recognition (NER) and noun phrases for the extraction of the keywords provided in the spaCy[2] library. Named-entity recognition (Mohit, 2014) is a natural language processing subtask that seeks to identify entities in a text. From the types of entities present in spaCy, we retain the names of people, events, nationalities or religious or political groups (NORP), organisations, countries, cities, states and other locations, products, works of art, law document names, and languages. Noun phrases are groups of two or more words that are based on a noun or pronoun surrounded by dependent words. For this task we have also added the nouns separately at the end. This approach is expected to provide a keyword list with less noise than the SKE method explained previously.
The keywords from the training data are used to build a mapping between the extracted keywords and disciplines with the corresponding score. This score has to consider both the presence of this keyword into this discipline compared to its presence into the other disciplines and the relevance of the keyword for the current discipline. We use the following notation: $k_{keyword \cap discipline}$ is the number of publications from the $discipline$ in which the $keyword$

[1] https://www.nltk.org/
[2] https://spacy.io/

appears, $k_{keyword}$ is the total number of publications in which the keyword appears, and $k_{discipline}$ the total number of publications in the discipline. Consequently, discipline relevance for the keyword – how relevant are the number of publications containing the keyword in the discipline compared to the total number of publications containing the keyword? – is calculated using a formula equivalent to precision in information retrieval (IR):

$$R_{discip}(discipline, keyword) = \frac{k_{keyword \cap discipline}}{k_{keyword}}$$

Additionally, the relevance of the keyword for the discipline – how relevant is the number of publications of the keyword in the discipline compared to the total number of publications in the discipline? – is calculated using a formula equivalent to recall in IR:

$$R_{kwd}(discipline, keyword) = \frac{k_{keyword \cap discipline}}{k_{discipline}}$$

The total score should reflect both these aspects. A keyword that is a good descriptor for a discipline should be both characteristic to the discipline (high R_{discip}) and appear in a significant number of publications from the discipline (high R_{kwd}). So, the final relevance score is calculated as the harmonic mean of the two relevance scores (F1 score):

$$R(discipline, keyword) = \frac{2 * R_{discip}(discipline, keyword) * R_{kwd}(discipline, keyword)}{R_{discip}(discipline, keyword) + R_{kwd}(discipline, keyword)}$$

After this calculation, all the existing combinations of keywords and disciplines have an assigned score.

For the classification, we calculate the sum of the keyword-discipline scores for all the combinations keyword-discipline with the keywords extracted from the publication. In mathematical terms:

$$Score(publication, discipline) = \sum_{keyword \in publication} R(discipline, keyword)$$

This score can be further used to decide on the most relevant discipline or disciplines for this publication. For the presentation of the results, this method will be referred to as KWBC (keyword-based classification), with simple keyword extraction (SKE) and complex keyword extraction (CKE) as variants.

The model is based on pattern identification using keywords, such that even if one keyword is misclassified or has a fuzzy classification (it is not clearly predominant in one discipline), the combination of multiple keywords helps distinguishing between them.

Random Forest Classifier

Additionally to the previous methods, we compare the results with a classic Random Forest Classifier (Breiman, 2001) trained on the encodings obtained both using TF-IDF and simple bag of words vectorizer. To avoid over-fitting, pruning was applied by limiting the minimum number of samples in a node for splitting to 20 and the minimum amount of samples in a leaf to 5.

Random Forest models are known to be very efficient and interpretable, however, in this implementation, the encodings consider the existence of the words and n-grams in the text, but not the context in which they appear, making the model not context-aware. They have extensively been used in classifying textual data (Chen et al., 2022; Eykens et al., 2021; Islam et al., 2019; Li et al., 2022) and are suitable for high dimensional noisy data (Islam et al., 2019).

BERT method

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based language representation model first introduced by Devlin et al. (2019). Since its launch, it has become a baseline for various natural language processing tasks, including text classification (González-Carvajal & Garrido-Merchán, 2021). The model pretrains bidirectional representations by considering both the context at the left and at the right of a given text, making it a context-aware model.

In this application of the method, the bert-base-uncased model[3] is used in combination with the provided script example provided by Hugging Face library[4] for the single discipline classification and an additional multi-label script created using the same library to identify the top 3 disciplines. The models are trained in 3 epochs with a learning rate of $5e^{-5}$.

Evaluation

For the evaluation, we are using a sample of our total data for evaluation that includes 200 publications from each FORD discipline (publications in training data are guaranteed not to be in test data). By design, the publications in the data have only one discipline assigned. Although the tested methods might classify some publications in another discipline than the one from WoS, the “right” one may still be in the top. Considering this, for evaluation purposes, both top-1 and top-3 disciplines according to the calculated score will be used. Publications will be considered correctly classified in top-1 if the assigned discipline is the same as in our data. Publications will be considered correctly classified in top-3 if the discipline from our data is present in top-3 disciplines assigned through classification to the publication. For evaluation purposes, the accuracy metric is used – the proportion of publications correctly classified.

The complexity of designing and evaluating an algorithm for discipline classification is described by Zhang et al. (2022), who point out that only 27% of papers have the same minor field (e.g. WoS SCs, Dimension FoR4 groups) assigned using the Fields of Research classification (from Dimensions), Web of Science (WoS) subject categories and the subject classification provided by Springer, while at a macro level (OECD FORD major fields, FoR2 division in Dimensions, top-level subject areas) the “identical” (equivalent) groups are at 60-70%. It is important to understand that there is no “gold standard” for the evaluation and the discipline that a publication belongs to leaves space of interpretability. The evaluation is included here to give a general idea of the performance of the different methods; some of the sources of misclassification will be further discussed in the Results section.

Results

The three methods introduced in the Methods section have distinct advantages and drawbacks. The approach based on BERT is considering the context during the encoding, making it robust to polysemy. However, as most of neural network based approaches, it is difficult to interpret the results, which makes it like a black box. Comparatively, the keyword-based method is easy to interpret and, moreover, the generated mapping between keywords and disciplines can be

[3] https://huggingface.co/bert-base-uncased

[4] https://github.com/huggingface/transformers/blob/main/examples/pytorch/text-classification/run_glue.py

updated without starting from zero and can be used for other similar problems. Additionally, the mapping can be used to directly analyse the most relevant terms for a discipline. The Random Forest-based approaches are in-between. Although extracting information regarding how the classification is done is easier than with BERT classifier, it is still not that straightforward as with the keyword-based method. TF-IDF and Bag of Words encodings are frequency-based like the keyword-based methods. However, in contrast to keyword-based methods, they also consider the number of occurrences of the keywords in an article, while the keyword-based method is only checking their presence in a binary way.

Table 1 The results of article classification using various methods and textual data

Method	Top 1 accuracy	Top 3 accuracy
BERT	**0.70**	**0.91**
KWBC SKE Abstract	0.56	0.8
KWBC CKE Abstract	0.53	0.78
Random Forest + TF-IDF	0.60	0.83
Random Forest + Bag of Words	0.60	0.83

Table 1 presents the classification results. The accuracy score shows the proportion of publications correctly classified out of the total number of test samples. The numbers are obtained by averaging the results for 5 executions of the algorithm with distinct random samples of train and test data from the dataset. The machine learning based methods perform overall better for this classification task, with the BERT classifier being the clear lead for both the single label classification and for top 3 classification. Both Random Forest based methods perform similarly with the current configuration. However, during our experiments, it has been noticed that with more restrictive parameters for pruning, the model using TF-IDF performs slightly better, which suggests that it converges quicker to a good result. In term of execution time and simplicity of use, the Random Forest methods outperform the others. Unexpectedly, the keyword-based classification built on simple n-gram extraction offers better results than the entity extraction and noun phrases approach. We hypothesize that this is caused by the total number of keywords. It was observed that when the algorithm has a filter for the keywords added that only keeps keywords that appear in at least 0.1% (the most restrictive part, eliminates around 90% of the keywords) of the publications and in at most 95% of the publications, the accuracy is around 5% lower. Adding the title with a coefficient of 2 – i.e., doubling the importance of keywords extracted from the title – impacts the overall result negatively, which suggests that the addition of the title adds more noise to the score than relevant information.

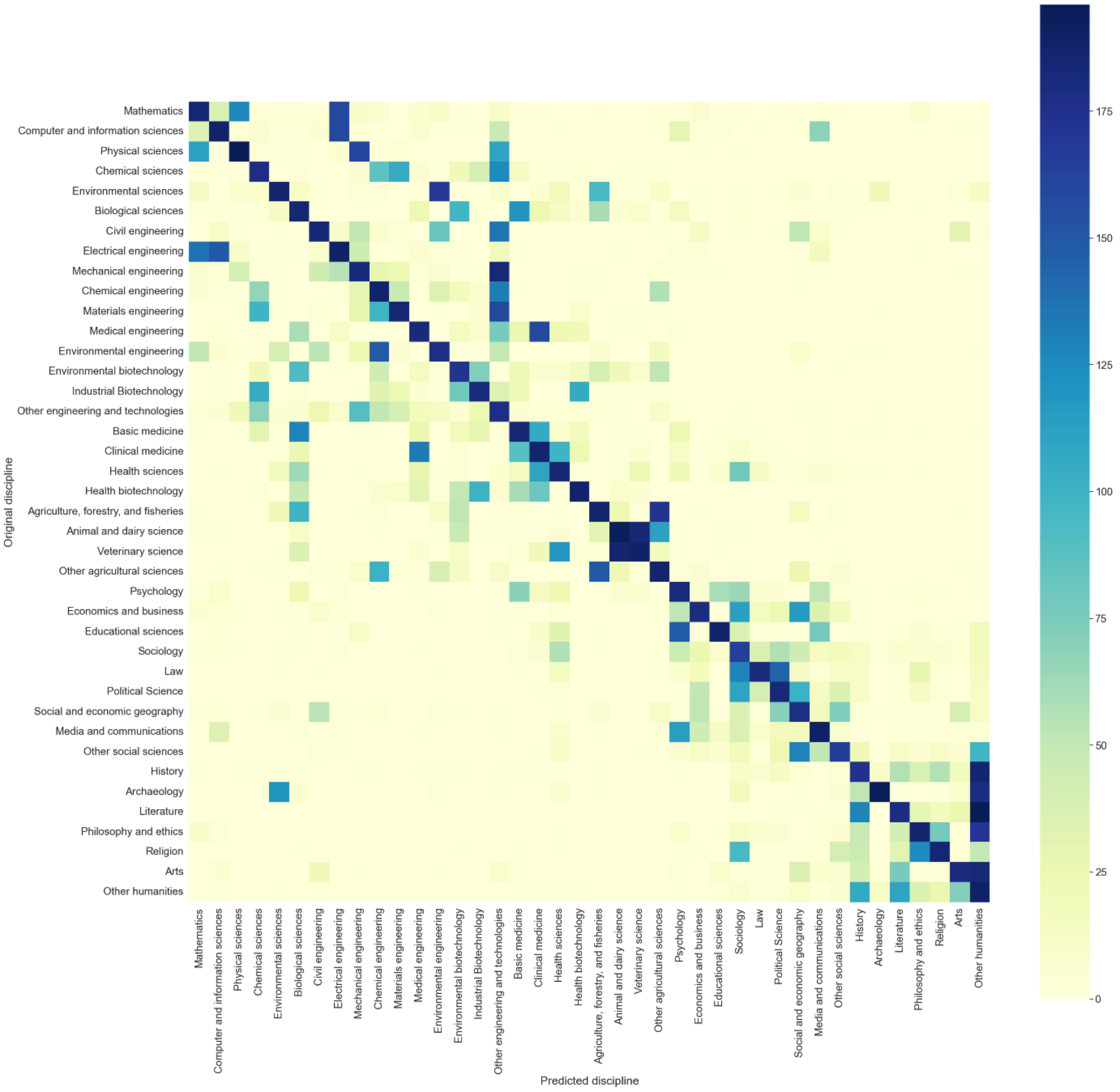

Figure 2. Confusion matrix between the "real" disciplines and the classification results using BERT classifier (top-3 classification)

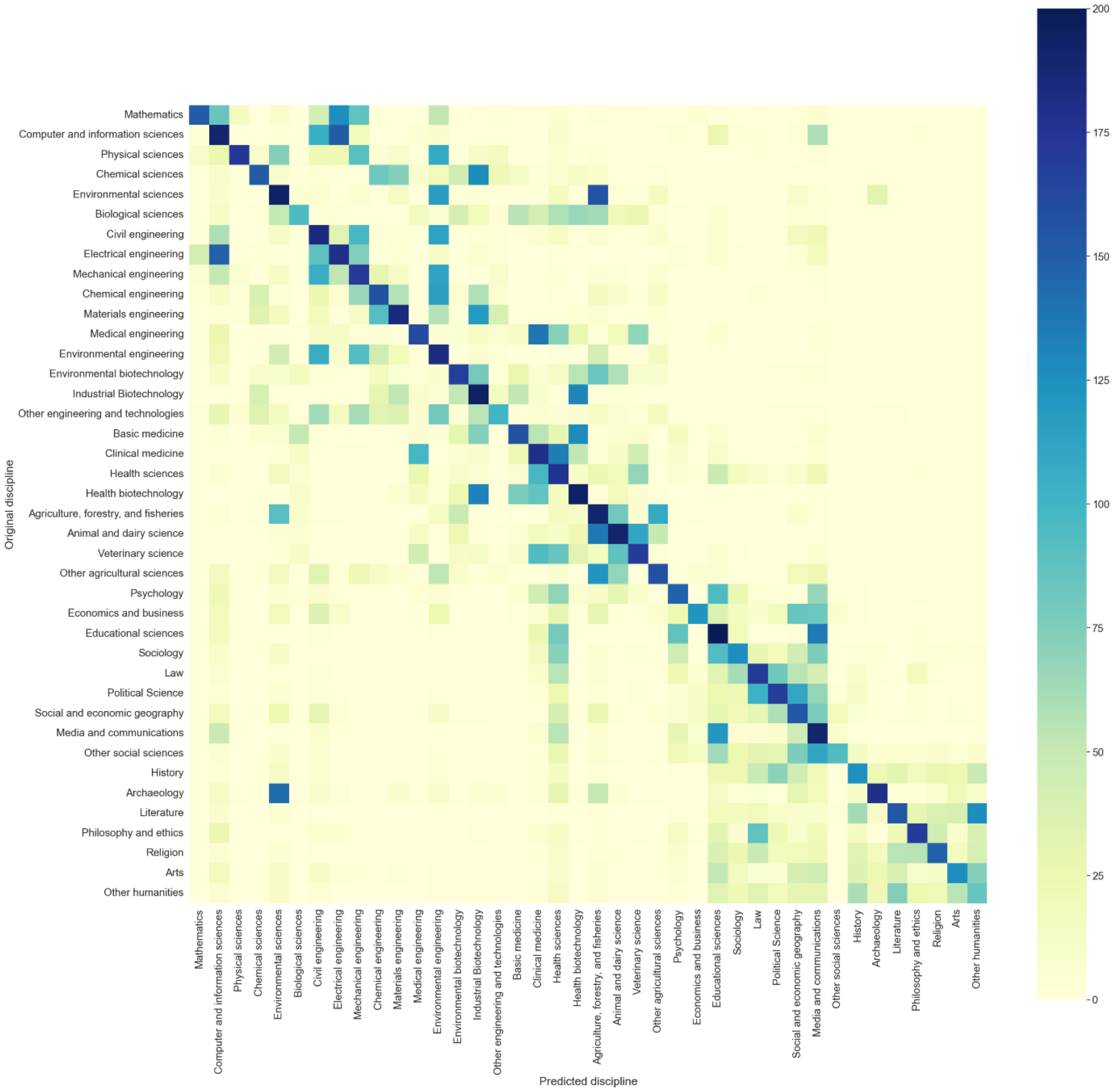

Figure 3. Confusion matrix between the "real" disciplines and the classification results using KWBC SKE Abstract method (top-3 classification)

Figure 2 and

Figure 3 show for each discipline the distribution of the predicted disciplines for top 3 using the KWBC SKE Abstract method and the BERT Classifier. The view using Random Forest on TF-IDF vectors is similar to KWBC SKE and are put in the Annex as Figure A2. Figures A1, A3, and A4 in the annex are the same statistics for top 1. As expected given the accuracy of the models, for almost all publications the BERT classifier has the correct class in the top 3 (dark colour for the diagonal). However, there are some discipline pairs the classifier is often unable to distinguish between. These include, among others, "Basic medicine" and "Biological Sciences", "Environmental biotechnology" and "Environmental engineering", "Veterinary Sciences" and "Animal and dairy sciences". It is also noticeable that compared to the keyword-based method, the list of the top 3 predicted disciplines for a publication from a given discipline using BERT classifier is more stable – there are less values on the confusion matrix between 50 and 100 and more high values and low values, meaning that most of the times the same disciplines are together in top 3. For the keywords-based method, the number of publications wrongly classified is correlated with the sum of the score of all the keywords for a discipline. The disciplines that have a higher sum are, on average, better classified. This puts the disciplines that are often applied in other areas ("Computer and information sciences", "Mathematics", etc.) and the ones that are very general or have unclear boundaries ("Other humanities", "Other social sciences", "Other engineering and technologies", etc.) at a disadvantage.

A table with the top 10 most relevant keywords per discipline using different keyword-based methods can be found in the Annex. The results are consistent across different random samples. Moreover, the top keywords seem to be keywords that are generally considered related to the discipline. However, even for some of the top keywords, the score for the discipline is quite low. This is mostly due to those keywords being frequently used in publications from other disciplines too, despite the fact that R_{discip} corrects for this. For example the keyword "art", which has the highest score for Other humanities, is present in 38 other disciplines, with a total count of 488. We note, furthermore, that it has a low score compared to top scores for other disciplines, and that it is more dominant in "Arts" (134 publications) than "Other humanities" (82 publications). This applies to most of the top keywords from Other humanities, which results in this discipline being frequently misclassified. Most of the highly misclassified disciplines in top 1 follow a similar pattern. Theoretically, having the top keywords have the highest score for the studied discipline is not a requirement because the algorithm is considering the sum of keywords' scores, which could compensate for some keywords having high scores in other disciplines. However, if there are multiple keywords having high scores in the same disciplines, it could lead to a higher rate of misclassification.

But, outside of the goal of improving model's accuracy, the confusion between disciplines is not an error, but a representation of the fuzzy boundaries between them and the difficulty of operationalizing the notion of a discipline – a natural result (Sugimoto & Weingart, 2015).

Discussion and conclusion

This article compares several text classification methods and introduces a text-based publication classification method based on keyword extraction. The approach that offers the best results for our data is the one using BERT for classification, however it lacks interpretability for now. The Random Forest approaches offer more interpretability and a higher score than the keyword-based approach. All the methods perform significantly better when checking if the "correct" discipline is in top 3, but BERT outperforms the other methods for both single class classification achieving an accuracy score of 70% and top-3 classification (91%). However, given the increasing level of interdisciplinarity and transdisciplinarity, the presence of the "correct" discipline in the top 3 predicted disciplines is more indicative of the quality of the classifier. Moreover, the usual disciplines that are predicted instead of the "correct" one are typically closely related fields of research (e.g. Predicting Clinical Medicine for Health Sciences). The top 3 predicted disciplines encapsulate better the lack of clearly determined limits between disciplines and allow more flexibility.

The classification of the publications using extracted keywords is highly dependent on the keyword extraction method. Counterintuitively, although the complex method that extracts noun phrases and entities generates a cleaner list of keywords, the results obtained using this method are worse than the ones from extracting all the words and n-grams. In our experiments, the number of extracted keywords has a higher impact than the "cleanliness" of the keyword list. Additionally, the algorithm could be further improved by exploring synonym analysis, considering the relevance of a keyword in a publication (e.g. using frequency).

The current study could be further improved in various directions:
Firstly, although the current research has not been conducted on interdisciplinary data, the presented models could be applied on interdisciplinary publications as each model calculates a score for each class, which can allow – using a threshold – the selection of multiple disciplines. Thus, further work is required to design a methodology for the identification of interdisciplinary publications and testing the models on them.

Secondly, different approaches for evaluation can be considered like calculating the score difference between the first classified discipline and the "correct" one or using existing information retrieval methods like discounted cumulative gain. This allows the classifier to punish less a wrong classification where multiple disciplines, along with the "correct" one have similar scores – a probable interdisciplinary research. Moreover, given that the approach creates a discipline-keyword mapping that characterizes the vocabulary used in different domains, it can be tested for the classification of other types of scientific texts, applied to identify the dynamics of keywords and topics in a discipline over time, to follow the apparition of new keywords if the model is made incremental.

Thirdly, another possibility for research concerns the multilingual aspect of research. Currently, the model focuses on English-language publications. We aim to extend it to (at least) Dutch and French, the predominant languages next to English for the research in Social Sciences and Humanities in Flanders, which is the main focus of our group. Additionally, the multilingual approach could be further used to confirm the generality of the approaches, which can be tested on data from different sources with different specificities.

Lastly, we aim to explore how the use of full-text would affect the results of the classification. This could result in more information about the content and hence give a better classification but it may also introduce more noise and decrease accuracy. For the current research it was considered that abstracts summarize well the essence of the article and are more generally available.

A limitation of our current approach relates to the training and test data collection: since the discipline of a publication is identified through the journal in which it has been published, our models are trying to identify the patterns present in the publications from those journals. Using this for training introduces some noise as not all the publications from a journal in a certain discipline are the best representatives of that discipline. In reality, what we are predicting is the probability that a publication belongs to a journal from a given discipline.

In conclusion, this article has compared several methods for the classification of scientific articles and for the model that has performed the best is BERT with 70% accuracy for top 1 and 91% accuracy for top 3 predictions. However, this method has the disadvantage of being more difficult to interpret, although research is conducted into facilitating the interpretability of neural network based models (Hao et al., 2021; Räuker et al., 2023). The two other studied group of approaches – Random Forest and keyword-based – are frequency-based approaches that are much more interpretable, but comparatively less accurate. We conclude that, overall, Random Forest-based methods are a compromise between interpretability and performance, being also the fastest to execute.

References

Ahlgren, P., Chen, Y., Colliander, C., & van Eck, N. J. (2020). Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of PubMed publications. *Quantitative Science Studies*, *1*(2), 714–729. https://doi.org/10.1162/qss_a_00027

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, *59*(2), 102798. https://doi.org/10.1016/j.ipm.2021.102798

Eykens, J., Guns, R., & Engels, T. C. E. (2021). Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches. *Quantitative Science Studies*, *2*(1), 89–110. https://doi.org/10.1162/qss_a_00106

Glänzel, W., Thijs, B., & Huang, Y. (2021). Improving the precision of subject assignment for disparity measurement in studies of interdisciplinary research. *FEB Research Report MSI_2104*, 1–12.

Golub, K., Hagelbäck, J., & Ardö, A. (2018). *Automatic classification using DDC on the Swedish Union Catalogue*. 4–16. http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-78378

González-Carvajal, S., & Garrido-Merchán, E. C. (2021). *Comparing BERT against traditional machine learning text classification* (arXiv:2005.13012). arXiv. https://doi.org/10.48550/arXiv.2005.13012

Hao, Y., Dong, L., Wei, F., & Xu, K. (2021). Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(14), Article 14. https://doi.org/10.1609/aaai.v35i14.17533

Islam, M. Z., Liu, J., Li, J., Liu, L., & Kang, W. (2019). A Semantics Aware Random Forest for Text Classification. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1061–1070. https://doi.org/10.1145/3357384.3357891

Klavans, R., & Boyack, K. W. (2017). Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998. https://doi.org/10.1002/asi.23734

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, *13*(2), 31:1-31:41. https://doi.org/10.1145/3495162

Mohit, B. (2014). Named Entity Recognition. In I. Zitouni (Ed.), *Natural Language Processing of Semitic Languages* (pp. 221–245). Springer. https://doi.org/10.1007/978-3-642-45358-8_7

Morillo, F., Bordons, M., & Gómez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology*, *54*(13), 1237–1249. https://doi.org/10.1002/asi.10326

OECD. (2015). *Frascati Manual 2015*. OECD Publishing. https://doi.org/10.1787/9789264239012-en

Osborne, F., Salatino, A., Birukou, A., & Motta, E. (2016). Automatic Classification of Springer Nature Proceedings with Smart Topic Miner. In P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, & Y. Gil (Eds.), *The Semantic Web – ISWC 2016* (pp. 383–399). Springer International Publishing. https://doi.org/10.1007/978-3-319-46547-0_33

Pech, G., Delgado, C., & Sorella, S. P. (2022). Classifying papers into subfields using Abstracts, Titles, Keywords and KeyWords Plus through pattern detection and optimization procedures: An application in Physics. *Journal of the Association for Information Science and Technology*, *73*(11), 1513–1528. https://doi.org/10.1002/asi.24655

Petr, M., Engels, T. C. E., Kulczycki, E., Dušková, M., Guns, R., Sieberová, M., & Sivertsen, G. (2021). Journal article publishing in the social sciences and humanities: A comparison of Web of Science coverage for five European countries. *PLOS ONE*, *16*(4), e0249879. https://doi.org/10.1371/journal.pone.0249879

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). *Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks* (arXiv:2207.13243). arXiv. https://doi.org/10.48550/arXiv.2207.13243

Salatino, A. A., Osborne, F., Thanapalasingam, T., & Motta, E. (2019). The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles. In A. Doucet, A. Isaac, K. Golub, T.

Aalberg, & A. Jatowt (Eds.), *Digital Libraries for Open Knowledge* (pp. 296–311). Springer International Publishing. https://doi.org/10.1007/978-3-030-30760-8_26

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Osborne, F., & Motta, E. (2018). The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas. In D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, & E. Simperl (Eds.), *The Semantic Web – ISWC 2018* (Vol. 11137, pp. 187–205). Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_12

Shu, F., Julien, C.-A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, *13*(1), 202–225. https://doi.org/10.1016/j.joi.2018.12.005

Shu, F., Ma, Y., Qiu, J., & Larivière, V. (2020). Classifications of science and their effects on bibliometric evaluations. *Scientometrics*, *125*(3), 2727–2744. https://doi.org/10.1007/s11192-020-03701-4

Sugimoto, C. R., & Weingart, S. (2015). The kaleidoscope of disciplinarity. *Journal of Documentation*, *71*(4), 775–794. https://doi.org/10.1108/JD-06-2014-0082

Urata, H. (1990). Information flows among academic disciplines in Japan. *Scientometrics*, *18*(3), 309–319. https://doi.org/10.1007/BF02017767

Waltinger, U., Mehler, A., Lösch, M., & Horstmann, W. (2011). Hierarchical Classification of OAI Metadata Using the DDC Taxonomy. In R. Bernardi, S. Chambers, B. Gottfried, F. Segond, & I. Zaihrayeu (Eds.), *Advanced Language Technologies for Digital Libraries* (pp. 29–40). Springer. https://doi.org/10.1007/978-3-642-23160-5_3

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, J. (2009). An extensive study on automated Dewey Decimal Classification—Wang—2009—Journal of the American Society for Information Science and Technology—Wiley Online Library. *Journal of the American Society for Information Science and Technology*. https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.21147?casa_token=X0QcwDzCFQEAAAAA%3Ar8eZ-P7P9RkCZ1dKWVKBU0liIXmDugfb5cm4SQEm38muSoRTL9tfyzW60i2bjCLc_ZDkjziygt86xIh5

Weber, T., Kranzlmüller, D., Fromm, M., & de Sousa, N. T. (2019). Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research. *ArXiv:1910.09313 [Cs, Stat]*. http://arxiv.org/abs/1910.09313

Zhang, L., Sun, B., Shu, F., & Huang, Y. (2022). Comparing paper level classifications across different methods and systems: An investigation of Nature publications. *Scientometrics*. https://doi.org/10.1007/s11192-022-04352-3

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

Appendix

Table A2 Top 10 keywords as identified using the keyword method

Discipline	Top 10 keywords SKE	Top 10 keyword CKE
Agriculture, forestry, and fisheries	soil, forest, plant, tree, crop, leaf, organic, specie, ecosystem, carbon	soil, soils, trees, forests, plants, species, carbon, ecosystems, yield, forest
Animal and dairy science	diet, feed, fed, dietary, weight, intake, broiler, supplementation, body weight, kg	diet, diets, kg, intake, supplementation, weight, treatments, fed, digestibility, gain

Archaeology	archaeological, site, excavation, settlement, late, archaeology, century, ancient, bronze, dating	site, sites, the site, excavations, bc, archaeology, remains, period, century, region
Art (arts, history of arts, performing arts, music)	music, art, musical, artist, contemporary, aesthetic, cultural, composer, artistic, theater	music, century, works, artists, this article, article, art, architecture, project, history
Basic medicine	Inhibitor, compound, potent, mouse, cell, drug, receptor, therapeutic, derivative, cancer	inhibitors, compounds, cells, mice, inhibition, activation, compound, inhibitor, activity, drugs
Biological sciences	specie, gene, genetic, genome, protein, evolutionary, sequencing, molecular, cell, genomic	species, genes, proteins, rna, cells, mechanisms, interactions, infection, expression, we
Chemical engineering	separation, particle, adsorption, liquid, gas, catalyst, flow, industrial, removal, reactor	particles, membrane, selectivity, temperature, adsorption, this work, separation, membranes, concentration, catalyst
Chemical sciences	reaction, ligand, synthesis, electrochemical, metal, synthesized, ion, electrode, bond, catalyst	reaction, complexes, synthesis, ions, ligands, reactions, ligand, compounds, molecules, yields
Civil engineering	concrete, steel, load, traffic, building, seismic, stiffness, finite element, reinforced, strength	concrete, buildings, strength, tests, specimens, stiffness, loading, beams, columns, capacity
Clinical medicine	patient, ci, cohort, outcome, confidence interval, clinical, disease, trial, median surgery	patients, ci, disease, surgery, therapy, outcomes, conclusions, risk, interval, mortality
Computer and information sciences	algorithm, propose, computing, optimization, cloud, network, datasets, proposed, user, task	algorithm, algorithms, theart, art, applications, problem, tasks, computing, datasets, internet
Earth and related environmental sciences	sediment, ocean, basin, record, deposit, climate, sea, marine, rock, ice	basin, deposits, record, ma, sediments, rocks, ocean, observations, climate, zone
Economics and business	firm, market, price, return, find, stock, financial, investor, volatility, investment	firms, markets, prices, returns, market, volatility, investors, we, shocks, evidence

Educational sciences	student, education, learning, teacher, teaching, school, skill, course, educational, classroom	students, education, learning, teachers, skills, educators, teaching, course, participants, training
Electrical engineering, electronic engineering, information engineering	robot, algorithm, controller, proposed, simulation, problem, propose, scheme, robotic, network	problem, robots, robot, systems, algorithm, networks, scheme, this letter, letter, network
Environmental biotechnology	strain, fermentation, enzyme, production microbial, gene, biomass, yeast, bacteria, acid	production, bacteria, strain, genes, enzymes, strains, the production, fermentation, yield, biomass
Environmental engineering	reservoir, oil, rock, pressure, fracture, fluid, permeability, gas, pore, drilling	reservoirs, reservoir, pressure, wells, oil, fractures, permeability, rocks, rock, flow
Health biotechnology	stem cell, stem, mesenchymal, mesenchymal stem, mesenchymal stem cell, cell, differentiation, bone, tissue, msc	cells, mscs, differentiation, expression, proliferation, msc, mesenchymal stem cells, therapy, regeneration, the expression
Health sciences	health, background, infection, parasite, care, disease, prevalence, population, intervention, patient	infection, ci, care, health, transmission, patients, prevalence, disease, interventions, risk
History	century, war, history, article, political, became, historian, historical, empire, british	century, history, article, war, this article, the article, the history, historians, british, empire
Industrial Biotechnology	nanoparticles, biocompatibility, release, vitro, delivery, drug, scaffold, vivo, cell, poly	nanoparticles, biocompatibility, cells, release, cytotoxicity, nm, delivery, properties, scaffolds, therapy
Languages and linguistics	language, english, linguistic, speaker, speech, corpus, word, discourse, verb, syntactic	language, english, languages, speakers, words, speech, corpus, spanish, children, verbs
Law	law, court, legal, criminal, crime, justice, police, right, federal, constitutional	law, courts, court, crime, rights, justice, violence, this article, police, article

Literature	literary, text, essay, narrative, writing, poetry, reading, poem, writer, story	novel, essay, this essay, texts, works, poetry, text, writing, fiction, reading
Materials engineering	ceramic, microstructure, degree c, prepared, diffraction, temperature, thermal, sintering, grain, phase	ceramics, temperature, properties, microstructure, xrd, composites, diffraction, materials, strength, phase
Mathematics	equation, numerical, prove, fractional, solution, existence, nonlinear, aip, differential equation, aip publishing	equations, equation, solutions, aip publishing, publishing, existence, the existence, problem, examples, solution
Mechanical engineering	numerical, simulation, vibration, heat, flow, numerical simulation, proposed, finite, heat transfer, velocity	simulation, simulations, flow, method, equations, parameters, model, the proposed method, system, this paper
Media and communications	user, information, technology, intention, library, service, online, social medium, perceived, survey	users, services, the findings, intention, findings, social media, survey, implications, adoption, libraries
Medical engineering	patient, diagnosis, background, clinical, serum, blood, healthy, laboratory, diagnostic, specificity	patients, diagnosis, background, methods, disease, specificity, conclusions, detection, biomarkers, samples
Other agricultural sciences	crop, agricultural, soil, irrigation, moisture, water, plant, objective study, farmer, speed	the objective, soil, yield, crops, brazil, objective, irrigation, experiment, agriculture, water
Other engineering and technologies	spectroscopy, detection, spectrum, electron, fluorescence, measurement, spectral, microscopy, accuracy, determination	spectra, spectroscopy, detection, nm, method, accuracy, technique, sensor, parameters, determination
Other humanities	art, essay, cultural, history, century, text, museum, author, article, literary	century, article, this article, history, essay, the article, author, this essay, the author, identity
Other social sciences	tourism, tourist, destination, hotel, sport, hospitality, implication, customer, economic, international	tourism, tourists, implications, the findings, findings, industry, the purpose,

		interviews, countries, intentions
Philosophy and ethics	philosophy, argue, philosophical, moral, philosopher, view, claim, argument, ethic, account	philosophy, view, account, argument, philosophers, conception, arguments, thesis, thought, notion
Physical sciences	quantum, dark, dark matter, scalar, physic, matter, decay, mass, energy, particle	physics, matter, energy, universe, spectrum, field, coupling, fields, modes, we
Political Science	policy, political, government, party, public, international, governance, actor, politics, find	governments, politics, actors, government, institutions, this article, article, countries, states, organizations
Psychology	participant, cognitive, disorder, psychological, behavioral, task, stimulus, child, autism, whether	participants, task, these findings, findings, adults, children, disorder, individuals, stimuli, measures
Religion	christian, religious, theology, religion, god, church, theological. tradition, faith, biblical	christian, theology, religion, god, church, article, this article, faith, tradition, the article
Social and economic geography	urban, city, planning, policy, china, economic, rural, government, land, spatial	cities, policies, china, planning, city, impacts, areas, policy, the paper, paper
Sociology	child, family, social, interview, youth, migrant, woman, experience, parent, gender	children, women, experiences, parents, interviews, families, youth, care, findings, research
Veterinary science	dog, animal, veterinary, disease, canine, infection, clinical, horse, virus, cat	dogs, animals, disease, cats, horses, infection, signs, conclusions, diagnosis, prevalence

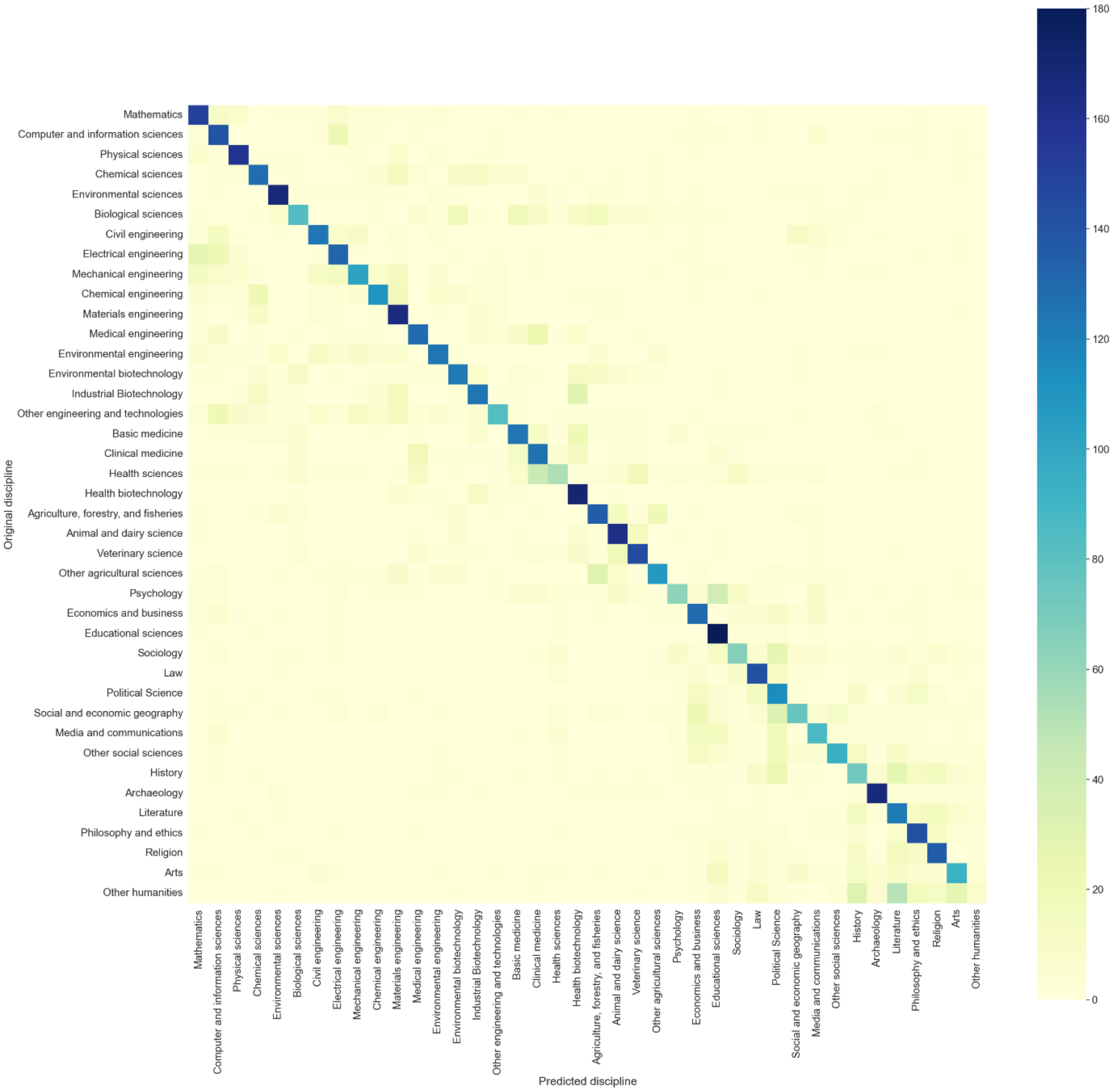

Figure *A1*. Confusion matrix between the "real" disciplines and the classification results using Random Forest with TF-IDF method (top-1 classification)

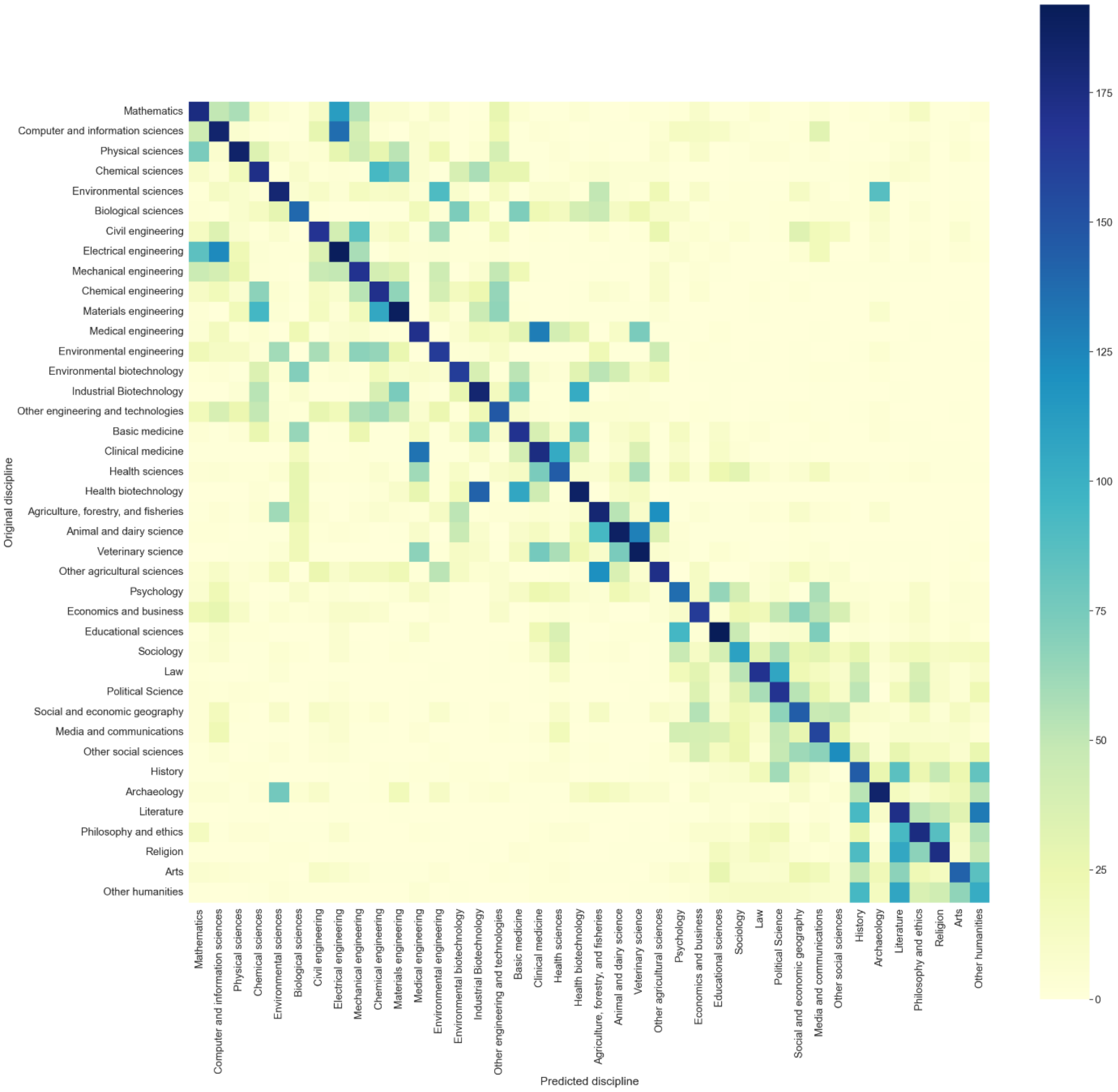

Figure *A2*. Confusion matrix between the "real" disciplines and the classification results using Random Forest with TF-IDF method (top-3 classification)

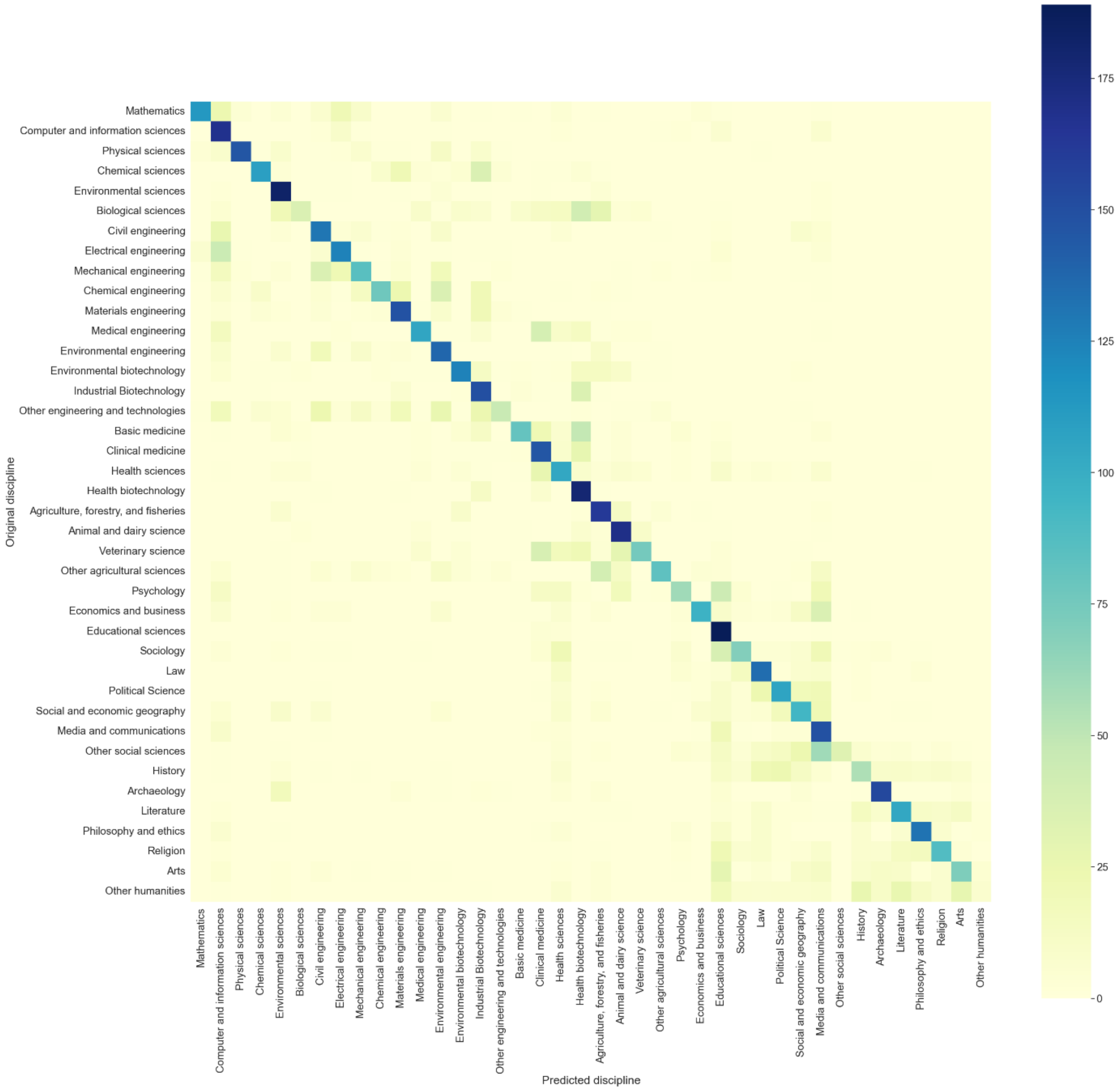

Figure *A3*. Confusion matrix between the “real” disciplines and the classification results using KWBC SKE Abstract method (top-1 classification)

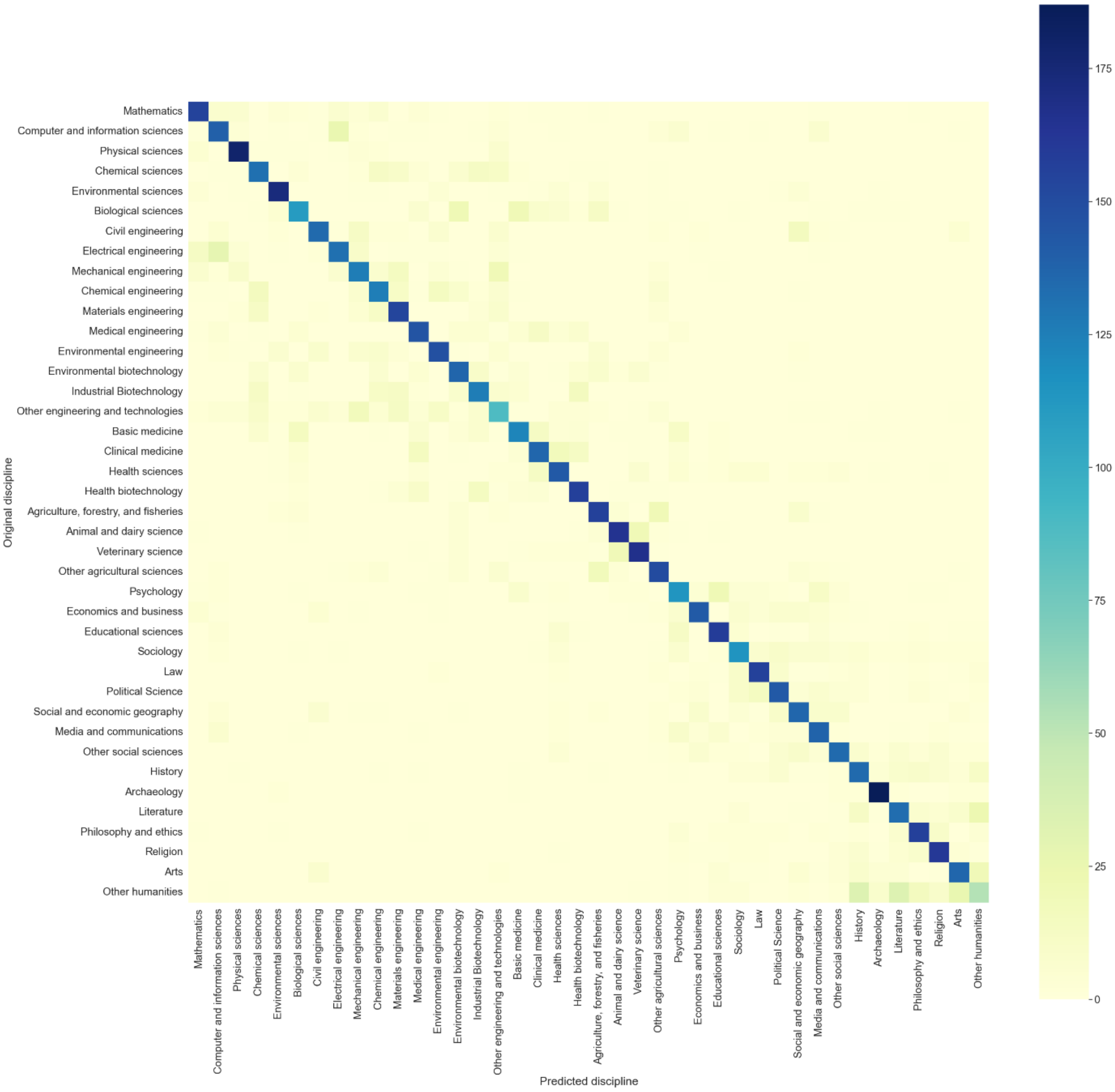

Figure *A4*. Confusion matrix between the "real" disciplines and the classification results using BERT classifier (top-1 classification)