RESEARCH ARTICLE

Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches

Joshua Eykens, Raf Guns, and Tim C. E. Engels

Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium

Keywords: disciplinary classification, granularity, multilabel classification, social sciences, supervised machine learning, textual data

ABSTRACT

We compare two supervised machine learning algorithms—Multinomial Naïve Bayes and Gradient Boosting—to classify social science articles using textual data. The high level of granularity of the classification scheme used and the possibility that multiple categories are assigned to a document make this task challenging. To collect the training data, we query three discipline specific thesauri to retrieve articles corresponding to specialties in the classification. The resulting data set consists of 113,909 records and covers 245 specialties, aggregated into 31 subdisciplines from three disciplines. Experts were consulted to validate the thesauri-based classification. The resulting multilabel data set is used to train the machine learning algorithms in different configurations. We deploy a multilabel classifier chaining model, allowing for an arbitrary number of categories to be assigned to each document. The best results are obtained with Gradient Boosting. The approach does not rely on citation data. It can be applied in settings where such information is not available. We conclude that fine-grained text-based classification of social sciences publications at a subdisciplinary level is a hard task, for humans and machines alike. A combination of human expertise and machine learning is suggested as a way forward to improve the classification of social sciences documents.

1. INTRODUCTION

Disciplines have long been considered as the fundamental units of division within the sciences (Stichweh, 2003). These units are knowledge production and communication systems, and can as such serve important classificatory functions (Hammarfelt, 2018; Stichweh, 1992, 2003; Sugimoto & Weingart, 2015; van den Besselaar & Heimeriks, 2006). The subjects of interest for scientometricians (i.e., scientific documents) are classified according to disciplines to facilitate research into knowledge production and dissemination. Over the past few decades, however, we have faced continuous growth of the number of new disciplines and specialties (i.e., internal differentiation), resulting in increasing dynamism and "intensification of the interactions between […] disciplines" (Stichweh, 2003, p. 85).

General classification systems such as the Web of Science (WoS) Subject Categories (SC) or the OECD's Fields of Science are too broad to adequately capture the more complex, fine-grained cognitive reality. Several concerns have been raised in this regard—here we mention

an open access journal

Check for updates

Citation: 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

DOI: https://doi.org/10.1162/qss_a_00106

Received: 12 May 2020
Accepted: 14 December 2020

Corresponding Author: Joshua Eykens
joshua.eykens@uantwerpen.be

Handling Editor: Ludo Waltman

The MIT Press

two central ones. First, Glänzel, Schubert, and Czerwon (1999) point out that the SC approach on the journal level works well for classifying publications in highly specialized journals, but that it is problematic for those appearing in multidisciplinary or general journals. Second, research such as Waltman and van Eck's (2012) large-scale clustering study, grouping publications based on their citation relations, indicated the feasibility of more fine-grained classification schemes. The authors cluster documents on three different levels, the most detailed of which can be conceived of as "small subfields" and consists of 21,412 clusters. While most bibliometric studies still make use of more general classification schemes for publications, these are limited in scope, only indicating broad scientific fields or general disciplines. Empirical studies such as the one conducted by Waltman and van Eck (2012), as well as theoretical arguments raised by sociologists of science, amplify the need for fine-grained classification schemes. More recently, Sjögårde and Ahlgren (2018, 2019) have shown that fine-grained specialized communities can be determined based on citation relations, and these communities in their turn might possibly exhibit specific citation and publication practices.

In Flanders, the Dutch speaking region of Belgium, the Flemish Research Discipline Standard ("Vlaamse Onderzoeksdiscipline Standaard" or VODS) has been introduced to facilitate a detailed classification of research, including research output (Vancauwenbergh & Poelmans, 2019a, 2019b). The VODS builds upon the OECD Fields of Science (2007), adding two more fine-grained levels. While the introduction of the VODS will open up new possibilities for understanding knowledge production and dissemination on a more detailed level, it also poses important challenges, the classification of publications in the social sciences being one of them.

Current bibliometric approaches to classification of publications are not entirely fit for the social sciences. This mainly has to do with lack of coverage in major citation databases (Ossenblok, Engels, & Sivertsen, 2012) and differences in publication and citation practices within the fields (Kulczycki, Engels et al., 2018; Nederhof, 2006). One possible way to address these concerns is including nonsource items in citation-based bibliometric maps (Boyack & Klavans, 2014). An alternative solution is making use of text-based methods.

1.1. Using Textual Data and Machine Learning to Cluster or Classify (Social Science) Publications

Compared to classification approaches making use of reference/citation data (or other metadata), the usage of purely textual information (titles, abstracts, full-texts, etc.) has thus far received less attention. Nevertheless, the theoretical relevance of an article's textual content for this task has already been emphasized since the seminal work by Rip and Courtial (1984). Michel Callon and colleagues have further developed this long tradition of co-word analysis research, which aims to map and describe scientific interaction and the formation of specialist communities (Callon, Courtial et al., 1983; Callon, Courtial, & Laville, 1991). More recently there has been a resurgence of interest in textual data, mainly due to increased computing resources and availability of potential data sources.

Machine learning methods currently spearhead a lot of research that is based on textual data. We can distinguish between supervised and unsupervised approaches. In unsupervised learning, no predefined classes or categories are available to learn from. Supervised learning, on the other hand, starts from a set of predefined categories, each of which has a number of instances or records assigned to it. An algorithm is then trained on these labeled instances, from which it tries to deduce the common characteristics of instances in each category to predict to which category a new, unseen instance might belong. The present article uses such a supervised approach.

In scientometric studies, unsupervised clustering of documents is common. Hybrid approaches to document clustering in which citation information and textual data are used have shown that

adding textual information can ameliorate the outcomes of document clustering (see for example Janssens, Zhang et al., 2009; Yau, Porter et al., 2014). Unsupervised clustering of documents based only on textual similarity (Boyack et al., 2011) has gained traction in the bibliometric community as well. Arguably, supervised ML has been less popular, presumably because in most scientometric clustering studies a granular ground truth classification at the article level is lacking.

An exploration of supervised ML algorithms combined with basic NLP techniques has been described by Read (2010), who used supervised learning to classify documents in, among others, the Ohsumed data set, part of MedLINE. The author reports F1 scores for different multilabel classification techniques, ranging from 0.1 up to 0.43. Classifier Chains (CCs) are proposed by Read (2010) as a possible solution to the task of multilabel, multiclass classification problems. The latter are tasks in which a document can be assigned to multiple categories at the same time. This kind of learning task is considerably more challenging than the single label classification problem.

Recent supervised ML algorithms with neural networks and word embeddings or BERT (Bidirectional Encoder Representations from Transformers) models, respectively, have also been used to vectorize and classify scientific documents. While these recent studies do not deal with multilabel, multiclass classification, they are relevant in that they apply these relatively new NLP techniques to vectorize scientific publications. Kandimalla, Rohatgi et al. (2020) report on a large-scale classification study in which they categorize papers according to WoS Subject Categories by making use of neural networks and word embedding models. The authors show that such classification systems work well, achieving an average F-score of 0.76. For the individual SC, the scores range from 0.5 to 0.95. In this study, however, the subcategories with too few records are merged or omitted from the analysis, as they "decrease the performance of the model." Documents that are labeled with more than one category are also dropped. The authors conclude that their experiment shows that the supervised learning approach scales better than citation clustering-based methods. Dunham, Melot, and Murdick (2020) train SciBERT classifiers on arXiv metadata and subject labels. This model is then used to identify AI-relevant publications in WoS, Digital Science Dimensions and Microsoft Academic. The authors report F1 scores ranging from 0.59 to 0.86 for the four categories within the field of AI.

Annif, an automated subject indexing tool currently being tested and implemented at the National Library of Finland, is also comparable to our approach (Suominen, 2019). Annif annotates terms from different subject vocabularies and thesauri to documents based on textual information, such as abstracts and/or titles. The ML module consists of an ensemble of classification algorithms. Annif annotates documents on a granular level, as the tested module was able to assign up to five indexing terms to documents. The module was evaluated on four corpora, including both academic and nonacademic texts, yielding F1 scores ranging from 0.14 to 0.46.

The present paper is an extension of work presented at ISSI 2019, where we applied supervised ML to classify sociology publications into subdisciplinary categories (Eykens, Guns, & Engels, 2019), reaching 81% accuracy. Note, though, that that paper only worked with publications assigned to one specialty. In this article, we study the use of textual data to classify publications from three social science disciplines into one or more subdisciplines. Much like Read (2010) and Kandimalla et al. (2020), we thus primarily aim to exploit textual characteristics of (social science) documents to categorize them into predefined disciplinary categories. As we will describe in more detail further on, we aim to categorize these social science abstracts into granular subcategories. Multiple categories can be assigned to one document at the same time. The novelty of

this study resides in the fact that we have used a procedure to validate the data collected for our ML experiment, and that multiple granular subdisciplinary categories can be assigned to one single document.

1.2. Outline

In Section 2 we describe the classification scheme used in detail. Section 3 describes the data sources used (Sociological Abstracts, ERIC, EconLit), and the collection and processing procedure. We have developed a structured way of collecting and validating textual data based on well-established disciplinary thesauri in tandem with a validation round by experts from the respective fields. This validation procedure is discussed in Section 3.3. Next, Section 4 further details the supervised ML algorithms and feature extraction techniques that we compare. Section 5 describes the results of the comparison, where we evaluate performance on two dimensions: the individual labels and the instances. Finally, we discuss our ML setup and contrast our approach to existing automatic classification techniques. We conclude with some reflections and pathways for future research, and briefly discuss practical applications.

2. THE FLEMISH RESEARCH DISCIPLINE STANDARD (VODS)

We make use of the Flemish Research Discipline Standard ("Vlaamse Onderzoeksdiscipline Standaard", abbreviated VODS, in Dutch), which is available at https://researchportal.be/en/disciplines and has been described in the literature (Vancauwenbergh & Poelmans, 2019a, 2019b). The VODS was introduced in the Flemish Research Information Space (FRIS, see https://researchportal.be/en), an aggregation platform of publicly funded research in Flanders, in 2019. In the future, all scientific output produced by scholars in Flanders may be classified according to the VODS. The VODS is structured as a hierarchical tree with four levels. To allow for international comparison, the first level overlaps with the seven broad fields of science present at the highest level of the OECD Fields of Science (OECD, 2007) coding scheme (hereafter referred to as OECD FOS). For the case of sociology, for example, at the top level of the OECD FOS we find category 5 "social sciences" and subcategory "5.4 sociology and anthropology" (Figure 1). This category is present in the VODS as well.

The VODS adds two more granular layers representing further subdivisions of the second layer of the OECD FOS. The third layer of the VODS might be interpreted as containing subdisciplinary categories, while items on the fourth level can be considered research specialties. To construct and define this scheme, experts from the corresponding fields were consulted by the creators of the VODS. In total, on the most granular level the VODS contains 2,493 codes. For further technical details on this classification scheme, we refer interested readers to Vancauwenbergh and Poelmans (2019a, 2019b).

Hierarchical structure of OECD FOS (ex. 5.4)

5. Social Sciences
   5.4 Sociology

Hierarchical structure of VODS (ex. 050405)

05 Social Sciences
   0504 Sociology and Anthropology
      050401 Anthropology (14 subcategories)
      050402 Applied Sociology (11 subcategories)
      …
      050405 Social Change
         05040501 Social change
         05040502 Social movements and collective action
         05040599 Social change not elsewhere classified
      …

Figure 1. Excerpt of tree structure: OECD FOS (2007) coding scheme and VODS (2019). The VODS classification scheme can be accessed at https://researchportal.be/en/disciplines.

Our objective is to automatically classify articles (based on abstracts and titles) into categories at level 3 of the coding scheme (e.g., 050402 Applied sociology and/or 050405 Social Change and/or …) for three fields within the social sciences, namely (0502) economics & business (10 classes at the third level), (0503) pedagogical & educational sciences (nine classes at the third level), and (0504) sociology & anthropology (12 classes at the third level). At level 3, we have 31 subdisciplinary categories for the three disciplines together. Section 3.3 will further detail the reasons why our approach operates at level 3 rather than level 4. In the following part we introduce the data sources used to collect the titles and abstracts for the three disciplines.

3. DATA SOURCES: SOCIOLOGICAL ABSTRACTS, ERIC AND ECONLIT

The data used for our study were downloaded from ProQuest (https://search.proquest.com). ProQuest provides good journal coverage of the social science literature compared to, for example, Scopus or WoS (Norris & Oppenheim, 2007). ProQuest offers access to a range of existing abstracting services and disciplinary databases. For the purpose of our analyses, we have used Sociological Abstracts to download bibliographic records from sociology & anthropology, EconLit for records from business & economics, and ERIC for records from the pedagogical & educational sciences.

3.1. Combinations of Indexing Terms as Proxies for Subject Specialties

A clear advantage of all three databases is that they make use of controlled vocabularies (or thesauri) for the records which are indexed. The Thesaurus of Sociological Indexing Terms is a well-developed and highly regarded indexing system used by Sociological Abstracts' service. Within EconLit, the Journal of Economic Literature (JEL) classification, also known as the *American Economic Association Classification System*, is used. Within ERIC, the Thesaurus of ERIC Descriptors is used. In addition, ProQuest's search engine allows us to filter on publication types and publication years. We selected all journal articles published between 2000 and 2018. These controlled vocabularies allow us to query ProQuest's command line search page for abstracts on a very fine levels of granularity. Figure 2 shows an example of the query we used for the category "law & economics," within business & economics. The full set of queries for all categories is available online (Eykens & Guns, 2020).

To design the queries, we manually coupled the indexing terms of these different data sets to the fourth level categories in the VODS and downloaded the abstracts found for each query. The first author went through the list of VODS categories and per category collected all relevant indexing terms from the thesaurus at hand.

The VODS provides semantic definitions of each category, which were formulated together with field experts. We used this information to manually retrieve the relevant indexing terms. In many cases, this was straightforward, because there was a perfect overlap with the indexing

```
MAINSUBJECT.EXACT("Personal Bankruptcy Law (K35)")
  OR MAINSUBJECT.EXACT("Regulation and Business Law: General (K20)")
  OR MAINSUBJECT.EXACT("Regulation and Business Law: Other (K29)")
  OR MAINSUBJECT.EXACT("Tax Law (K34)")
  OR MAINSUBJECT.EXACT("Regulated Industries and Administrative Law (K23)")
  OR MAINSUBJECT.EXACT("Regulation and Business Law (K2)")
  OR MAINSUBJECT.EXACT("Business and Securities Law (K22)")
  OR MAINSUBJECT.EXACT("Contract Law (K12)")
  OR MAINSUBJECT.EXACT("Tort Law and Product Liability; Forensic Economics (K13)")
  OR MAINSUBJECT.EXACT("Law and Economics: General (K)")
```

Figure 2. Example of command line query designed for Business & Economics, category 05020109 Law & economics.

terms (for EconLit, this was the case for nearly all categories). In other cases, some additional indexing terms were found to be relevant (see for example Figure 2).

The indexing terms were then used to query ProQuest. The records retrieved for each of the 214 level 4 categories were subsequently downloaded (with an upper limit set to 1,000 records per VODS level four category) and saved in a separate folder, which was labeled with the corresponding VODS category. The data collection was carried out between December 2018 and February 2019. After collecting and processing, the merged sets (all files for the three fields together) resulted in a raw set consisting of 148,341 records (see Table 1).

3.2. Data Cleaning and Processing

To clean the raw data set, we followed a protocol consisting of four steps. At step 1 we removed all records that were missing an abstract or title. Although we limited our search to records published between 2000 and 2018 (step 2), there were still some in our data set that were published before 2000 or after 2018. These were omitted as well. Lower and upper boundaries were set for the word count of the abstracts (step 3): minimum 50 and maximum 1,000, respectively. These limits were found to adequately weed out cases where the abstract field replicated either the title or the entire full text.

Whereas we expected only journal articles resulting from our queries, other publication types were present as well. The reason for this might have to do with the fact that all three data sets have been designed by different organizations, which results in a diverse range of variable names to describe the different publication types used within the data sets. At step 4, for each data set, we compiled a list of unique variable names present in the collected records and filtered out those describing publication types that we did not want to take into consideration (e.g., book reviews, interviews, editorial material, instructional material). A list of the remaining, relevant publication types was used to restrict our data set to research articles published in journals.

Table 1 provides an overview of the number of records in each set before and after cleaning. For Sociological Abstracts and EconLit, our initial collection of records was reduced by a little over 20%. For ERIC, the total number of records was reduced by almost 40%. The large intergroup difference observed is mainly due to a large number of records classified as "instructional material" in ERIC. The large intragroup difference is due to the smaller number of subcategories present in the VODS. For business & economics we queried 84 categories, for pedagogical & educational sciences 53 categories, and for sociology & anthropology 77 categories.

As discussed above, we have designed queries for each level 4 category in the VODS and collected records from the respective databases. Some records appeared multiple times—that is, some records were retrieved with different queries. After deduplication and relabeling, the

Table 1. Number of records collected from each database: Before and after cleaning

VODS category	Indexing service consulted	Initial number of records	Number of records after cleaning
0502 Business & economics	EconLit	63,407	50,577
0503 Pedagogical & educational sciences	ERIC	23,521	14,527
0504 Sociology & anthropology	Sociological Abstracts	61,413	48,805
Total		148,341	113,909

data set contains 113,909 multilabeled abstracts, with an average of 1.1 labels per abstract (min. = 1, max. = 6, SD = 0.36).

3.3. Expert Validation: Interindexer Consistency and F1 Scores

To validate the reliance on controlled vocabularies described above, a domain expert from each of the three disciplines was contacted. The three experts were given a random sample of 45 abstracts and titles, which they were asked to classify according to the VODS level 4 categories corresponding to their field of expertise (i.e., sociology & anthropology, business & economics, and pedagogical & educational sciences). Each expert was presented a set of abstracts and titles from their own discipline. The expert working in the field of business & economics, for example, was given abstracts and titles originating from EconLit (business & economics) only. No limitations were set on the number of categories the indexers were allowed to assign.

Next, the classification by the experts on the one hand and the classification based on the controlled vocabularies of each database on the other were compared. To do so, the interindexer consistency (IIC) was calculated for each record in the sample. For every sample, we calculated the average IIC using the method described by Rollin (Leininger, 2000). First, a percentage of consistency between two indexers (here: the expert and the controlled vocabulary) is calculated for each document *d*:

$$IIC_d = \frac{2A}{B + C} \tag{1}$$

Here, *A* denotes the number of categories on which both indexers agree, *B* is the number of categories assigned by indexer 1 (expert) and *C* the number of categories assigned by indexer 2 (controlled vocabulary). The IIC at document level is the Dice coefficient of the two sets of categories assigned by the indexers. The average IIC for the whole sample is calculated by dividing the sum of the IICs for all individual documents by the total number of documents *N* (in our case, equal to 45). In addition, we calculated F1 scores for each disciplinary sample. We have calculated these scores for levels 3 and 4 of the VODS.

Table 2 displays the results of the IIC and F1 calculations. The F-scores are also included for the assessment of the performance of the ML models. On level three of the VODS, the IIC varies between 45.2% for the sample from Sociological Abstracts and 62.2% for EconLit. On level four, the IIC scores are considerably lower, with a minimum of 23.7% for EconLit and a maximum of 39.7% for ERIC. Previous research into IIC in the case of the PsycINFO database shows similar results to those obtained for level 3 of the VODS. Leininger (2000) evaluates IIC for a similar classification scheme, based on research areas within psychology. Using Rollins' method, he finds an average IIC of 45% (Rollin, 1981, as cited in Leininger, 2000, p. 6). Sievert and Andrews (1991) study the IIC for a subset from Information Science Abstracts. The authors report average consistency scores of about 50%. Funk and Reid (1983) study the IIC for MEDLINE. They report a consistency score of 61.1% for the MeSH terms assigned to documents. While our scenario is somewhat different (i.e., the first author "reclassified" publications according to the indexing

Table 2. Rollin (1981) interindexer consistency (IIC) and weighted F1 scores for the three data sets at two classification levels

Sample from	IIC level 3	F1 level 3	IIC level 4	F1 level 4
Pedagogical & educational sciences—ERIC	52.9%	0.59	39.7%	0.42
Business & economics—EconLit	62.2%	0.57	23.7%	0.51
Sociology & anthropology—Sociological Abstracts	45.2%	0.67	26.7%	0.48

terms and experts were consulted to validate this reclassification), it seems that these low scores rather indicate the difficulty of the problem at hand. Therefore, we conclude that the level 3 classification is sufficiently robust to be used for our ML learning experiment, and hence we limit ourselves to the classification of journal articles at this level. For matters of interpretation of the differences in scores, at level 3 we have 31 subdisciplinary categories in total, compared to 214 research specialties at level 4.

4. METHODS

For 30 years, the dominant paradigm of text classification (TC) has consisted of ML approaches. ML algorithms are deployed such that "a general inductive process automatically builds an automatic text classifier by learning, from a set of preclassified documents, the characteristics of the categories [or labels] of interest" (Sebastiani, 2002, p. 2). ML approaches have already been applied to classify abstracts or full texts of journal articles. Langlois, Nie et al. (2018) classify papers into two broad domains: empirical and nonempirical. Our approach is different from such studies as the level of granularity of categories into which we classify texts is far greater. Consequently, articles from two different level 3 subdisciplinary categories are overall much more similar than what is encountered in most other classification tasks.

The classification problem discussed in this paper belongs to the domain of multiclass multilabel classification. Multiclass classification refers to assigning one of more than two classes to an instance. Multiclass multilabel classification is an extension of this problem where we assign one or more of multiple classes to an instance (Read, Pfahringer et al., 2009). Some abstracts were thus assigned to multiple classes (up to a maximum of six).

A popular strategy is to transform the multilabel problem into different single-label classification tasks. This can be done making use of binary relevance. As a baseline classifier, we make use of Multinomial Naïve Bayes (MNB). We optimize this classifier to explore the best feature engineering techniques as described below. Next, we compare the results obtained with MNB to those obtained by a Gradient Boosting (GB) model. After discussing the feature engineering steps in the following part, we will present a short description of the algorithms and the metrics that were used to evaluate performance on different aspects.

4.1. Feature Engineering

Feature engineering for multilabel TC is done in the same way as for single-label TC. The "features" or columns of the matrix are representations of words in the abstracts and titles of the publications. The Bag of Words approach (BoW) is a traditional, popular, and simple yet powerful way of vectorizing documents for TC. The BoW approach consists of slicing a text into words or phrases (without taking word order into account). We have built customized tokenizer functions in Python to extract four different textual features: lemma unigrams, lemma bigrams (combined with unigrams), nouns, and noun phrases (Figure 3). Although previous research has shown that for the BoW approach more advanced document representations such as nouns and noun phrases are "not adequate to improve TC accuracy," we wanted to explore this for our specific use-case (Moschitti & Basili, 2004).

We made use of the natural language processing packages NLTK (Loper & Bird, 2002) and SpaCy (Honnibal & Montani, 2018) to parse the texts and to perform part of speech tagging and stemming. For stemming, we made use of NLTK's implementation of the snowball stemming algorithm. Scikit-learn's count vectorizer and TF-IDF (term frequency-inverse document frequency) transformer were used to process the outcomes of the different feature extraction

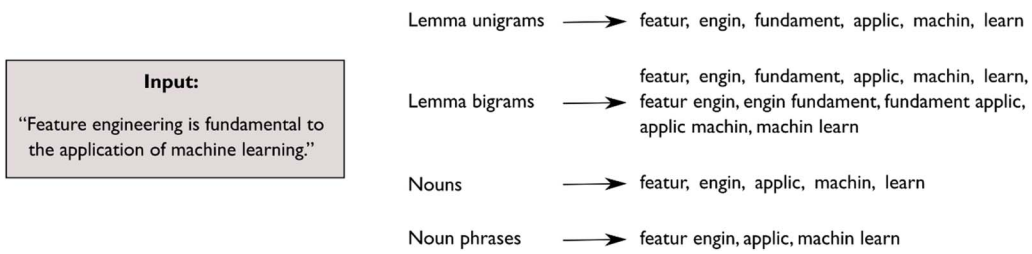

Figure 3. Four feature extraction methods: lemma unigrams, lemma bigrams, nouns, and noun phrases.

methods (Pedregosa, Varoquaux et al., 2011). With each tokenizer, we tested the performance of both (normalized) TF and TF-IDF. This resulted in eight different feature spaces (see Figure 4).

Feature sparseness is a common problem in TC. Transformation methods that make use of bigrams can easily bring about feature matrices with hundreds of thousands or even millions of columns, leading to very high dimensionality. To reduce the dimensionality, we make use of a feature selection method based on randomized decision trees. After extracting textual features, we fit a shallow extra trees classifier (maximum depth of 10) to the data to select the most relevant ones.

4.2. Classification Algorithms

4.2.1. Multinomial Naïve Bayes (MNB)

MNB is one of the most popular TC algorithms used by the ML community. It is a fast, scalable (i.e., iterates very fast over large data sets), and successful approach for many TC problems. Over the years, it has become a popular baseline method, "the punching bag of classifiers" (Lewis, 1998, p. 2). MNB makes use of Bayes' theorem to construct histograms based on the feature vectors—in our case counts or probabilities of the textual features present in a document—for every single instance. The classifier associates these histograms with the labels and estimates likelihoods of a label and a distribution of feature counts occurring together.

If, however, a feature-class combination has zero counts, the probability will be set to zero. This mitigates the necessary information of the other probabilities by multiplying them by zero. For the algorithm to be able to deal with such problems a smoothing parameter is used. Another way of

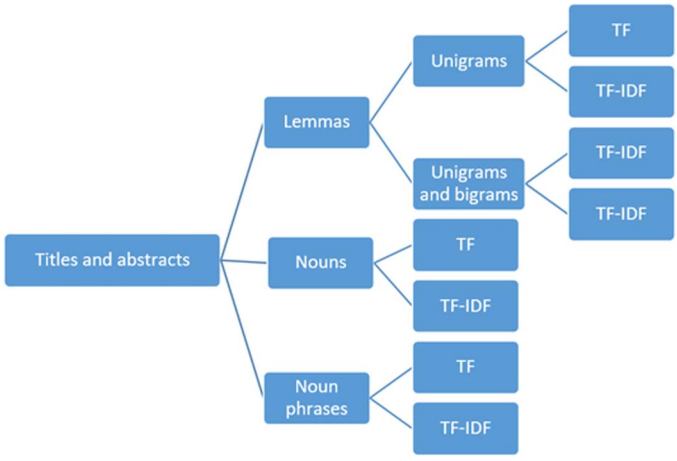

Figure 4. Overview of feature transformation steps.

dealing with this problem is transforming the feature space into a TF-IDF normalized matrix (Rennie, Shih et al., 2003).

4.2.2. Gradient Boosting Decision Trees (LightGBM)

Gradient Boosting Decision Trees (GBDTs) are, as the name indicates, tree-based learning algorithms. These algorithms build ensemble models, or groups of decision trees aimed at reducing residual errors for a split point in a decision tree. Boosting is a specific ensemble technique that sequentially builds the models on random subsets of the features and instances. When an instance is misclassified, its weight is increased and the next model tries to correct for this error.

In practice, this algorithm can be very time-consuming. Ke, Meng et al. (2017) have come up with a solution to this problem by optimizing the randomness of the feature and instance selection step. They combine Gradient-based One-Side Sampling (GOSS) with Exclusive Feature Bundling (EFB) to speed up the training process. The GOSS procedure pays more attention to instances with larger gradients (i.e., having more impact on the classification error of a model) "and randomly drop those instances with small gradients" (Ke et al., 2017, p. 2). This approach is implemented in the LightGBM software package (https://lightgbm.readthedocs.io).

The EFB implementation exploits feature sparseness, which is a very common problem in TC. It bundles sparse features together into a single feature, efficiently reducing the dimensionality. In a previous study, we have found the LightGBM implementation of GB to be the best performing algorithm to classify publications in sociology & anthropology, achieving accuracy scores well over 80% (Eykens et al., 2019). Different from this previous study, in this paper we assess classifier performance for a vastly more complex multilabel setting.

Decision tree-based models, however, come at a cost. They require tuning a wide range of parameter settings. For LightGBM, one can set well over 100 parameters[1]. For our purposes, we have chosen to optimize for 11 core parameters:

- the number of trees that will be built;
- the maximum depth of the trees: to limit tree growth;
- the number of leaves of the decision trees: last splits made in the model when reaching the optimal number of splits for a given loss function, or when reaching the predefined maximum depth;
- the learning rate: sets the weight of the outcomes of each tree for the final output;
- maximum in bin: handles the maximum number of bins in which the feature values will be grouped;
- regularization alpha (L1): limits the impact of the leaves encouraging sparsity (i.e., weights to zero);
- regularization lambda (L2): limits the impact of the leaves by encouraging smaller weights;
- minimum child weight: the minimum sum of instance weight which is needed in a leaf (child);
- bagging fraction: the fraction of the data set used for each iteration;
- bagging frequency: the number of trees training per random subsample of the data set; and
- minimum data in leaf: the minimum number of samples which should be captured in a leaf.

We will describe how we optimized the parameters in Section 4.2.4.

[1] For a complete overview of the parameters used in LightGBM, see https://lightgbm.readthedocs.io/en/latest/Parameters.html.

4.2.3. Multilabel classification: Classifier Chains (CC)

Two main approaches to multilabel classification exist: problem transformation and algorithm adaptation. The most popular and computationally least expensive approach is problem transformation, where a multilabel classification problem is transformed into *N* single-label classification problems. An example of problem transformation is turning the multilabel task into *N*-labels binary classification problems, wherein each binary classification problem is treated by a separate classifier. This is also known as the *Binary Relevance method* and has proven success in the domain of multilabel TC (Zhang, Li et al., 2018).

As each label is treated separately, however, the algorithm effectively ignores label dependence. Read et al. (2009) have suggested an improvement of the Binary Relevance method by "chaining" the results of each classifier to the input space so that the next training round takes the results of previous classifiers into account. As different disciplinary categories might be closer to each other in terms of concepts and topics studied, we do not expect labels to be completely independent of each other. Hence, we opt for the CC approach. It should be noted that other approaches exist, but these come at a cost of computational complexity as well as intuitive understanding of the models.

4.2.4. Cross-validation

After vectorization, dimensionality reduction, and problem transformation with a binary relevance-based CC algorithm, a holdout set (25% of the complete data set) was sliced from the initial data set using an iterative stratification technique as proposed by Sechidis, Tsoumakas, and Vlahavas (2011). This stratification method handles class imbalance for multilabel learning problems in such a way that the distribution of instances over classes in the validation set is kept as close to the actual distribution as possible.

Figure 5 visualizes the cross-validation procedure. The test data (0.25 of the total set) will be used for the final evaluation of our models. For each iteration, a different subset of the remaining 75% of the data (training data) are used to evaluate different parameter settings for the feature engineering options presented above. We make use of randomized parameter grid search and threefold cross-validation to evaluate different parameter settings on parts or "folds" of the training data. This means we run three new random experiments, each of which again divides the training data into two different parts, using 66.66% of the training data to train a model with a random parameter setting, and evaluating that setting on unseen data (the darker grey area represented above). We make use of three different slices of training and test data to make sure that our findings are robust.

4.3. Evaluation Metrics

Evaluating the performance of multilabel classification is not as straightforward as is the case for single-label classification. Single-label classifiers' predictive performance can be evaluated

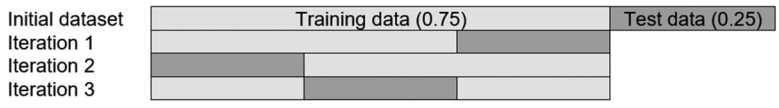

Figure 5. Visualization of training set—validation set folds and test data. Lighter grey represents training samples, and darker grey represents validation samples.

using the accuracy measure (i.e., the fraction of correctly classified instances over the total number of instances). The Accuracy *Acc* is calculated as follows:

$$Acc = \frac{1}{N}\sum_{i=1}^{N} I(Y_i = \hat{Y}_i) \tag{2}$$

I here is the indicator function. Y_i is the set of true labels (subdisciplinary categories, in our case) for document *i*, and $\hat{Y}_i$ is the set of predicted labels for document *i*. For multilabel classification assessing such a score based on the full set of labels per instance would be too harsh, "since even a single false positive or false negative label makes the example incorrect" (Read, 2010). Using multiple metrics to capture different dimensions of the multilabel prediction is advised (Read, 2010; Zhang & Zhou, 2014). Two main dimensions can be assessed: the individual labels and the entire training or testing label sets per instance (Zhang & Zhou, 2014, p. 1822). For full label set evaluation, we calculate accuracy, and for label-based evaluation, we calculate precision and recall.

Per label ℓ from the set of labels *L*, we can determine the set test_ℓ of documents to which this label has been assigned and the set pred_ℓ of documents for which the classifier predicts this label. Weighted average precision *P* is determined as follows:

$$P = \frac{1}{\sum_{\ell \in L}|\text{test}_\ell|}\sum_{\ell \in L}|\text{test}_\ell|\frac{|\text{test}_\ell \cap \text{pred}_\ell|}{|\text{pred}_\ell|} \tag{3}$$

where $|\cdot|$ denotes set cardinality. Similarly, weighted average recall *R* is:

$$R = \frac{1}{\sum_{\ell \in L}|\text{test}_\ell|}\sum_{\ell \in L}|\text{test}_\ell|\frac{|\text{test}_\ell \cap \text{pred}_\ell|}{|\text{test}_\ell|} = \frac{\sum_{\ell \in L}|\text{test}_\ell \cap \text{pred}_\ell|}{\sum_{\ell \in L}|\text{test}_\ell|} \tag{4}$$

The F1 score is the weighted average of precision and recall. Precision and recall are first "macroaveraged" by calculating the weighted mean of precision and recall for each label, and these are used to calculate the final F1 scores. These measures give an indication of the performance of our algorithm across the three different disciplinary data sets. Precision (Eq. 3) in a multilabel setting is "the fraction of predicted relevances which are actually relevant" (Read, 2010, p. 41). In addition, Schapire and Singer (2000, as cited in Tsoumakas & Katakis, 2007) propose Hamming Loss to take into account the fraction of labels that are predicted incorrectly. Hamming Loss is calculated as follows (see Sorower, 2010):

$$Hamming\ \ Loss\ = \frac{1}{N|L|}\sum_{k=1}^{|L|}\sum_{i=1}^{N} y_{i,k} \oplus \hat{y}_{i,k} \tag{5}$$

Here, $\oplus$ is the exclusive-or operator, $y_{i,k}$ is 1 if document *i* has label *k* and 0 otherwise, and similarly, $\hat{y}_{i,k}$ is 1 if document *i* is predicted to have label *k* and 0 otherwise. We average these scores over the total number of classes $|L|$ and predictions *N*. Hamming Loss thus denotes the fraction of incorrectly predicted labels and its optimal value is 0.

5. RESULTS

In the first part, we present the best results obtained for Multinomial Naïve Bayes. As detailed above, we have vectorized the abstracts and titles making use of three slightly different textual characteristics: lemmas, nouns, and noun phrases. Because of the computational requirements, the ML steps were carried out on the High Performance Computing infrastructure of VSC (the Flemish Supercomputer Center) at the University of Antwerp.

Table 3. Results of Multinomial Naïve Bayes classification performance for optimal feature space (lemma bigrams, no IDF normalization). Training and evaluation on hold-out set. Best results (per row) in bold

	TF				TF-IDF			
	Lemma unigrams	Lemma bigrams	Nouns	Noun phrases	Lemma unigrams	Lemma bigrams	Nouns	Noun phrases
Set accuracy	0.20	**0.24**	0.17	0.14	0.19	0.21	0.17	0.15
Hamming Loss	0.04	0.04	0.04	0.03	0.04	0.03	0.04	**0.03**
Precision	0.61	0.61	0.62	**0.66**	0.60	0.64	0.61	0.63
Recall	0.31	**0.37**	0.26	0.19	0.30	0.31	0.25	0.20
F1 score	0.36	**0.42**	0.31	0.27	0.35	0.38	0.31	0.29

5.1. Multinomial Naïve Bayes

For the Multinomial Naïve Bayes classifier, we aim to optimize the smoothing parameter alpha. We randomly sample a value from a loguniform distribution, ranging from very small (i.e., 10^{-10}) up to 1 (i.e., add-one or Laplace smoothing). After finding the optimal value for alpha (0.13883) by fitting the algorithm to the three folds of the training set, we make a prediction for the hold-out test set. The results for the best feature representation method are presented in Table 3. The optimal representation strategy turns out to be lemma bigrams without IDF normalization.

Making use of bigrams for lemmas decreased the Hamming Loss and increased the other scores. We achieved quite similar results with TF-IDF transformed vectors. Interestingly, noun phrases, except for the Hamming Loss evaluation metric, do not yield improved results.

5.2. Gradient Boosting (LightGBM)

For the GB algorithm, we randomly sample values for 11 different parameters. For a more detailed explanation of these parameters, refer to Section 4.2.2. To reduce computing time, we limited the number of random iterations to 100. If we were to perform a full parameter grid search, the number of model fits would be far too high. Keeping in mind that 25 fits take about three hours, this is not desirable.

Compared to the best results achieved with MNB, the GB implementation scores better on almost all evaluation metrics, except for precision (see Table 4). It is interesting to note that

Table 4. Scores for Gradient Boosting classification on the validation set, for each feature space. Best results (per row) in bold

	TF				TF-IDF			
	Lemma unigrams	Lemma bigrams	Nouns	Noun phrases	Lemma unigrams	Lemma bigrams	Nouns	Noun phrases
Set accuracy	0.46	**0.46**	0.43	0.33	0.45	0.45	0.43	0.32
Hamming Loss	**0.03**	0.03	0.03	0.04	0.03	0.03	0.03	0.04
Precision	**0.66**	0.64	0.60	0.49	0.66	0.63	0.60	0.49
Recall	0.48	**0.50**	0.45	0.36	0.48	0.50	0.45	0.36
F1 score	0.54	**0.55**	0.49	0.40	0.54	0.55	0.49	0.39

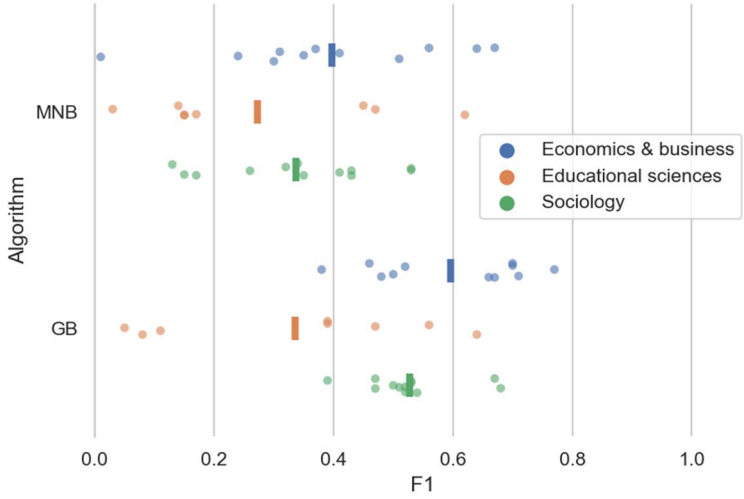

Figure 6. Box plot of F1 scores for all 31 subdisciplines for MNB and Gradient Boosting (GB). Preprocessing: lemma bigrams, no IDF. The subdisciplines are grouped per discipline and the vertical line segments indicate the average F1 scores per discipline.

MNB scores better for the precision metric in some scenarios. Accuracy scores, however, strongly increase for GB, and Hamming Loss also falls back. The same feature transformation strategy seems to work best for GB. For the lemma bigrams feature extraction method without IDF normalization we achieve an F1 score of 0.55. Hamming Loss is considerably lower as well, with a fraction of 0.3% of the labels wrongly assigned. 46% of the label combinations predicted by the algorithm were the same as those in the test set. It is noteworthy that the differences between TF-IDF and TF feature transformations are insignificant.

Figure 6 shows how the F1 scores are distributed across all 31 subdisciplines. We observe that the scores for GB are not only higher on average but also less spread out, with the exception of three poorly scoring subdisciplines. These three are all subdisciplines of educational & pedagogical sciences: Informal learning, General pedagogical & educational sciences, and Parenting &

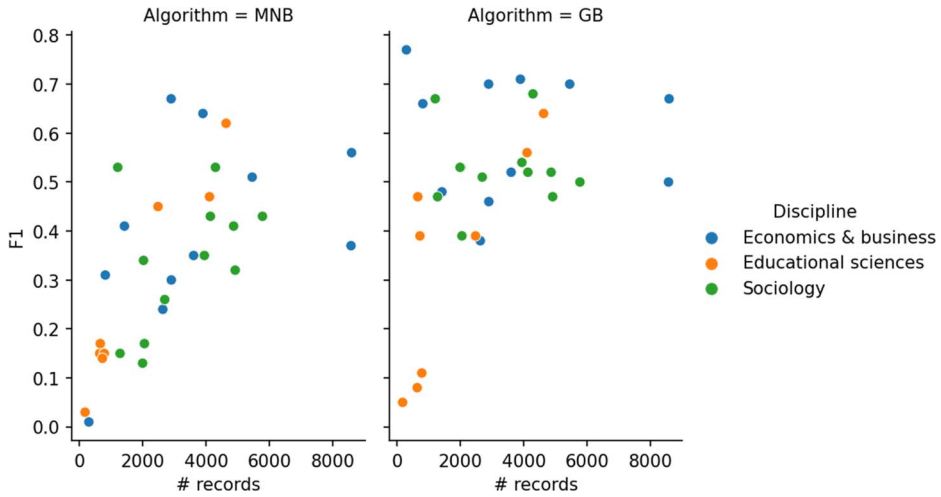

Figure 7. Relation between the number of records and F1 scores for MNB and GB for each of the 31 subdisciplines studied. Preprocessing: lemma bigrams, no IDF.

family education. Except for these subdisciplinary categories, overall no discipline performs clearly better or worse than the others, although the number of training records seems to have some influence: subdisciplines with fewer training records tend to get lower F1 scores (Figure 7). While this relation is somewhat stronger for MNB, the three cases for GB with exceptionally low F1 scores all have few (between 174 and 780) records.

6. DISCUSSION

Classifying research output into disciplinary categories is of fundamental importance for nearly all bibliometric analyses. In the introduction to this paper, we touched upon the issue of differentiation in the sciences, leading to an ever-increasing number of research communities and disciplines (Stichweh, 2003). This emergence of new disciplines can lead, among other things, to the formation of new research specialties, the organization of new conferences, the formation of new scientific societies and the foundation of new journals (see Shneider, 2009). As the landscape of disciplines grows more diverse, classification schemes are being updated to better fit this dynamic reality.

The development of such an updated classification scheme is exemplified by the implementation of the VODS in Flanders (see Vancauwenbergh & Poelmans, 2019a). Such a diverse and fine-grained classification scheme makes it possible to study interactions between disciplines (i.e., inter- and intradisciplinary knowledge flows) more closely, and map discrepancies between different classification systems with more detail. Yet, it requires new ways of approaching classification tasks as well, in particular in settings such as the classification of expertise, projects, and outputs for which citation data are not available. In this article we take up the specific challenge of a fine-grained classification of social sciences journal articles using the text of their abstracts and titles.

To summarize, our study consists of three elements. First, we constructed a labeled data set. As the VODS classification scheme is relatively new, we lack a data set of classified publications or other documents that can readily be used for ML purposes. This led us to manually construct a training data set consisting of data extracted from EconLit, ERIC, and Sociological Abstracts. Each of the 31 VODS subdisciplines of economics & business, pedagogy & educational sciences, and sociology & anthropology was translated to a thesaurus-based query for the respective databases. Second, the query results were validated by human experts. IIC and F1 scores indicate that categories at level 3 (subdisciplines) and 4 (specialties) of the VODS can sometimes be hard to distinguish between. At the same time, the IIC scores for level 3 categories are comparable to those obtained in earlier IIC studies.

Third, the labeled data set at level 3 was used to train Multinomial Naïve Bayes and GB ML models. If we compare Figure 6 to Table 2, the configuration with the best results yields F1 scores slightly below those for the validation by human experts. This indicates that the models might still be improved somewhat, but very high scores are probably unrealistic or indicative of overfitting. Taken together, the results suggest that level 3 of VODS is so fine grained that some categories are hard to discern in practice and as a result a certain degree of ambiguity becomes unavoidable, at least for the disciplines studied here.

While some of the reported indicators, such as F-scores, are relatively low, we think it is instructive to compare our results to those of the recent studies by Kandimalla et al. (2020) and Dunham et al. (2020). While these authors report better accuracy, it should be highlighted that in this paper we specifically look at the applicability of supervised learning in the context of social sciences. As Kandimalla and colleagues note, this is not an easy task given the large overlap in terminology and the proximity of the categories. Kandimalla et al. (2020) have for that reason

dropped or collapsed 120 out of 235 of the SC from their data set. In addition, they drop documents assigned to multiple disciplines. It should be noted that WoS SC are less granular than the ones used in our study (i.e., at the level of disciplines instead of subdisciplines). Dunham et al. (2020) report good scores for their model, which classifies AI publications into subdisciplinary categories, but their model is restricted to only four categories in AI, hence it is also less prone to errors. Our system works with 31 subcategories, divided over three social science disciplines. Taking these elements into account, it becomes clear that the lower scores are to a large extent a result of the difficulty of the task at hand.

A matter of concern that can be raised in this regard is to what extent classification of documents at a level of granularity that is finer than that of disciplines is feasible. Disciplines, and especially subdisciplines and research specialties, are in constant flux. Whereas most publications might belong to the knowledge base of just one discipline, their contents may be of relevance to two or more subdisciplines and research specialties. Theoretical work such as actor-network theory in the social sciences, for example, has been of relevance for many disciplines, subdisciplines, and research specialties, not only in the social sciences. Interdisciplinary studies, in which an integration of different disciplinary knowledge sources takes place to tackle a research question, may classify under several research specialties, subdisciplines, and disciplines. As these examples illustrate, a multilabel approach, as applied in this paper, is needed in view of the validity of a classification.

This framework requirement needs to be balanced with requirements in terms of the accuracy, feasibility, and reliability of a classification scheme. As the results of our study show, the classification of social sciences publications into subdisciplines (VODS level 3) on the basis of abstracts and titles is a hard task for both humans and machines; classification into research specialisms (VODS level 4) probably is not all that meaningful any more (cf. the IIC and F1 scores in Table 2). We argue that classification at the subdiscipline level should be further explored and fine-tuned, as this level of granularity corresponds to actual policy needs and might be improved by smart combinations of human input and ML. For example, a recommender system might be improved through validation by the authors of papers and machine classifications might gain accuracy through the use of larger sets of texts describing expertise, projects, and publications classified by humans.

6.1. Limitations

Four limitations of this paper should be highlighted. First, we could not compare our results to any benchmark. Although there have been some experiments in which supervised ML techniques are used to classify (or study elements of) scientific articles (see for example Langlois et al., 2018; Matwin & Sazonova, 2012), to date no comparable applications or data sets exist (i.e., medium-sized annotated sets of social science publications classified according to fine-grained disciplinary categories)—at least not to our knowledge. The lack of previous work in this line of research makes it hard to benchmark our results for this specific problem setting.

Second, given that the records in our data set were extracted from EconLit, ERIC, or Sociological Abstracts, each record has been assigned to only one (but possibly multiple subdisciplines of the same) discipline of the VODS level 2 (i.e., to economics & business, to pedagogy & educational sciences, or to sociology & anthropology). Hence, interdisciplinary cases are not present in our initial training data. We cannot compare the performance of the models deployed in this study at different levels of granularity, in particular the discipline and subdiscipline levels. However, our results do show that the subdiscipline level is, at least for articles in social sciences

and using their abstracts and titles only, the most fine-grained level that makes sense for classification exercises.

Third, we have coupled classification systems with two entirely different functions. On the one hand, we have the indexing systems based on the thesauri. These are systems that are designed for information retrieval purposes and have no limit to the number of indexing terms that can be assigned to a document. In such a system, there is no purpose in trying to fit a document into one to six subdisciplinary categories. Thus, we have reduced the complexity and granularity of the thesaurus-based classification to a fixed number of disciplinary groups. This "mismatch" between the two classification systems might lead to relatively low-scoring results when an ML algorithm is tasked with reproducing this classification.

Fourth, as discussed in Section 3.1, the queries have been manually constructed by the first author. The indexing terms in the thesauri were coupled to VODS discipline codes based on the semantic definition of each field in the VODS. It can be argued that this is a highly subjective task, as previous research has shown that disagreement between indexers when annotating records with indexing terms is commonplace. For many categories, however, the indexing terms nicely overlapped with the categories of the VODS. This gave us confidence in the construction. As the expert validation yielded results comparable to previous exercises of this kind, we believe this procedure to be of sufficient quality to allow for an automated (re)classification experiment. On the other hand, one can also interpret the relatively low IIC scores as indicative of the inherent ambiguity at this level of granularity.

6.2. Future Research and Practical Applications

The use of a minimum of textual data makes the approach presented in this study practical to generalize to other data sets (e.g., projects and project applications). Using additional bibliographic metadata would presumably increase the performance of the classification algorithms. Full-text documents would be an interesting path forward, yielding more textual data and a better sensitivity of TF-IDF transformations. In addition, it would be interesting to study ambiguities of the classification resulting from the predictions made by the algorithm and study those in detail.

With regard to the ML modules used, we acknowledge that more advanced and complex language-processing techniques have a good track record when it comes to automatically classifying text documents (e.g., BERT and related models). Dunham and colleagues (2020) have shown that SciBERT models outperform other NLP methods when applying them to classify publications in the field of Artificial Intelligence. For our purposes, however, we have opted to keep the setup relatively straightforward. The main motivation behind this study was to investigate and compare the feasibility of using supervised ML algorithms for this particular, challenging fine-grained classification task. We leave comparisons of other methods and feature transformation procedures for future research.

Questions surrounding the properties of interdisciplinarity demand for a clear operationalization of disciplines, which is not straightforward. This is in itself also the main reason why many different classification schemes are used in different contexts, each pointing to insights about different aspects—organizational, cognitive, etc.—of a discipline (Guns, Sīle et al., 2018). Textual approaches might lead to other insights regarding the cognitive structure of disciplines, but these same disciplines are in constant flux (Yan, Ding et al., 2012). A fixed classification scheme will not meet future developments in science; "… human assigned subject categories are akin to using a rearview mirror to predict where a fast-moving car is heading" (Suominen & Toivanen, 2016, p. 2464). To this end, the team working on the VODS has

provided a "not elsewhere classified" category for all the subfields (Vancauwenbergh & Poelmans, 2019a). This particular category has not been studied in this article. Future deployments of the classification system in Flanders will allow researchers to identify themselves and/or their projects with this category and assign documents to it, and, following from this, we could study text residing in these categories to discover emerging research problems and topics.

Once researchers employed by the Flemish universities start to label their expertise, projects, and outputs using the VODS, supervised ML algorithms can be trained on a broader range of disciplinary categories, allowing for a broader evaluation of the method proposed in this paper. This approach will enable us in practice to assist with annotating unlabeled work, or it can serve to underpin an online recommendation system for researchers, embedded in current research information systems. The output of supervised text classifications can also be compared to other existing classification schemes. We can, for example, contrast the publication level classification with journal level classifications of the same publications to study the disciplinary or interdisciplinary diversity of journals.

Finally, we should highlight that measuring IIC consistency is not straightforward. While there exists a long tradition of research that makes use of scoring systems such as IIC or F1 scores to assess the reliability and functionality of classification systems, there have been attempts to include semantic relations between indexing terms or categories to develop more realistic measures of indexing accuracy. Medelyan and Witten (2006) propose calculating the cosine similarity between word vectors of vocabularies or semantic definitions of categories. This is an interesting approach, but, to our knowledge, there are no systematic comparisons with other scoring systems available to date. It would be interesting to use such an approach when assessing classification systems in which a semantic definition of categories is available. The classification error could for example be weighted by the cosine distance between sentence embeddings of semantic definitions of the disciplinary categories. If a classification error is made whereby two distant categories are mistaken for each other, then the error is greater than when these categories are closer to each other in terms of cosine distance.

7. CONCLUSION

In this article we present a supervised ML approach to classify social science journal articles into multiple fine-grained disciplinary categories. By making use of GB with CC we are capable of assigning one or more disciplinary categories to text documents (i.e., abstracts and titles). To do so, we have compiled a new data set consisting of 113,909 records originating from three disciplinary databases in the social sciences (EconLit, ERIC, and Sociological Abstracts).

The novelty of this study lies in two aspects: the construction of the labeled data set, based on discipline-specific thesauri, and the application of supervised ML algorithms to classify social science journal articles into one or more fine-grained disciplinary categories using text. We show in detail how we have collected the data and how we have validated the labeling based on the subject indexing terms from the thesauri. With regard to the ML methods, we compare different feature engineering techniques and two well-established classification algorithms. The Gradient Boosting classifier (LightGBM) in a Classifier Chaining framework is capable of predicting approximately 46% of the exact label combinations correctly, with a fraction of 0.3% of labels assigned incorrectly. The F1 score is 0.55.

In a previous study (Eykens et al., 2019) we assessed the performance of four different ML algorithms for the classification of sociology and anthropology journal articles extracted from Sociological Abstracts into fine-grained disciplinary categories (level 4 of the VODS). Making use of the same LightGBM module (Ke et al., 2017), we were able to correctly classify over

80% of the publications. In this previous study, we made use of simple feature engineering (i.e., lemmas and unigrams) and we did not assess whether multilabel classification was possible. Aside from the work by Read (2010) to date, we are unaware of studies making use of similar methods to achieve fine-grained disciplinary classifications. To our knowledge, no work exists that studies the performance of supervised ML algorithms to classify social science documents on such a granular level.

Because we have significantly scaled up our data set, this study adds more nuance to the previous experimental study (Eykens et al., 2019). We have added textual data from two additional disciplinary databases, namely ERIC and EconLit, and we have assessed more complex feature engineering techniques as well. Importantly, we assess whether multilabel classification is manageable. The results confirm the robustness of our previous work and expand it to additional data sources. We further demonstrate that to a certain extent the approach is indeed generalizable to a multilabel classification task. To achieve this, the quality of the data collection and data validation is crucial. Hence, we encourage others to develop a thought-through data collection and validation procedure to make sure that the complete ML experiment is reproducible, from data collection and processing onwards.

To summarize, this study shows that supervised ML algorithms are capable of classifying social science journal articles into predefined, fine-grained categories based on the limited textual data of abstracts and titles only. However, for both human experts and machines, such classification at the subdisciplinary level proves very hard, to the extent that the question can be raised of whether such an attempt makes sense. Given the need for fine-grained classification in view of assessments, evaluations, and policy, we suggest that the informetric community further explores the possibilities for such fine-grained classification. For example, can the results obtained in this study be improved with different or more advanced NLP techniques, and by combining human expertise with advanced ML techniques? Like others (Boyack & Klavans, 2014; Suominen & Toivanen, 2016), we do not believe it to be fruitful to consider one or other classification system superior. We do instead insist that each approach has its merits, especially when contrasted to others. We hope that our work will spur others to conduct similar studies that explore the limits of the feasibility of classification through algorithms and human experts.

ACKNOWLEDGMENTS

We would like to thank the editor and anonymous reviewers for their comments as well as the three experts, Doctor Pieter Spooren (University of Antwerp), Professor Doctor Raf Vanderstraeten (Ghent University), and Professor Doctor Nick Deschacht (University of Antwerp and KU Leuven) who helped validating the data used for the analysis.

AUTHOR CONTRIBUTIONS

Joshua Eykens: Conceptualization Methodology Investigation Formal analysis Data curation Writing—original draft Writing—review & editing Visualization Project administration. Raf Guns: Conceptualization Methodology Investigation Formal analysis Writing—original draft Writing—review & editing Visualization Supervision. Tim C.E. Engels: Conceptualization Writing—original draft Writing—review & editing Supervision.

COMPETING INTERESTS

The authors have no conflict of interest.

FUNDING INFORMATION

The computing resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation Flanders (FWO) and the Flemish Government.

This investigation has been made possible by the financial support of the Flemish government to the Centre for R&D monitoring (ECOOM). The opinions in the paper are the authors' and not necessarily those of the government.

DATA AVAILABILITY

The Zenodo dataset (Eykens & Guns, 2020) consists of the full set of queries and the number of results for each, the data resulting from the expert evaluation, as well as the source code. Unfortunately, due to copyright restrictions from ERIC, EconLit and Sociological Abstracts, we are not able to make the retrieved records themselves openly available.

REFERENCES

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., … Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE*, *6*(3), e18029. **DOI:** https://doi.org/10.1371/journal.pone.0018029, **PMID:** 21437291, **PMCID:** PMC3060097

Boyack, K. W., & Klavans, R. (2014). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics*, *8*(3), 569–580. **DOI:** https://doi.org/10.1016/j.joi.2014.04.001

Callon, M., Courtial, J.-P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, *22*(1), 155–205. **DOI:** https://doi.org/10.1007/BF02019280

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*(2), 191–235. **DOI:** https://doi.org/10.1177/053901883022002003

Dunham, J., Melot, J., & Murdick, D. (2020). Identifying the development and application of artificial intelligence in scientific text. *arXiv preprint*. arXiv:2002.07143.

Eykens, J., & Guns, R. (2020). Supervised classification of SSH publications. **DOI:** https://doi.org/10.5281/zenodo.3822309

Eykens, J., Guns, R., & Engels, T. C. E. (2019). Article level classification of publications in sociology: An experimental assessment of supervised machine learning approaches. In G. Catalano, C. Daraio, M. Gregori, H. F. Moed, & G. Ruocco (Eds.), *17th International Conference on Scientometrics & Informetrics (ISSI2019)* (vol. 1, pp. 738–743). Sapienza University of Rome, Italy: Edizioni Efesto.

Funk, M. E., & Reid, C. A. (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, *71*(2), 176–183.

Glänzel, W., Schubert, A., & Czerwon, H.-J. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics*, *44*(3), 427–439. **DOI:** https://doi.org/10.1007/BF02458488

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. **DOI:** https://doi.org/10.1007/s11192-018-2775-x

Hammarfelt, B. (2018). What is a discipline? The conceptualization of research areas and their operationalization in bibliometric research. In R. Costas, T. Franssen, & A. Yegros-Yegros (Eds.), *Science, Technology and Innovation Indicators in Transition—STI2018* (pp. 197–203). Leiden, The Netherlands: Centre for Science and Technology Studies (CWTS).

Honnibal, M., & Montani, I. (2018). spaCy 2.0.11. **DOI:** https://doi.org/10.5281/zenodo.4291179

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management*, *45*(6), 683–702. **DOI:** https://doi.org/10.1016/j.ipm.2009.06.003

Kandimalla, B., Rohatgi, S., Wu, J., & Lee Giles, C. (2020). Large scale subject category classification of scholarly papers with deep attentive neural networks. *arXiv preprint*. arXiv:2007.13826.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., … Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Neural Information Processing Systems 2017* (pp. 1–9). Long Beach, CA: Curran Associates, Inc.

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., … Zuccala, A. (2018). Publication patterns in the social sciences and humanities: Evidence from eight European countries. *Scientometrics*, *116*, 463–486. **DOI:** https://doi.org/10.1007/s11192-018-2711-0

Langlois, A., Nie, J. Y., Thomas, J., Hong, Q. N., & Pluye, P. (2018). Discriminating between empirical studies and nonempirical works using automated text classification. *Research Synthesis Methods*, *9*(4), 587–601. **DOI:** https://doi.org/10.1002/jrsm.1317, **PMID:** 30103261

Leininger, K. (2000). Interindexer consistency in PsycINFO. *Journal of Librarianship and Information Science*, *32*(1), 4–8. **DOI:** https://doi.org/10.1177/096100060003200102

Lewis, D. D. (1998). Naïve (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.), *10th European Conference on Machine Learning—ECML-98* (vol. 1398, pp. 4–15). Chemnitz: Springer. **DOI:** https://doi.org/10.1007/BFb0026666

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics* (vol. 1, pp. 63–70). Philadelphia, PA: Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/1118108.1118117

Matwin, S., & Sazonova, V. (2012). Direct comparison between support vector machine and multinomial naive Bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, *19*(5), 917. **DOI:** https://doi.org/10.1136/amiajnl-2012-001072, **PMID:** 22683917, **PMCID:** PMC3422847

Medelyan, O., & Witten, I. H. (2006). Measuring inter-indexer consistency using a thesaurus. In *6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 296–297). Chapel Hill, NC: ACM. **DOI:** https://doi.org/10.1145/1141753.1141816

Moschitti, A., & Basili, R. (2004). Complex linguistic features for text classification: A comprehensive study. In S. McDonald & J. Tait (Eds.), *Advances in Information Retrieval. ECIR 2004* (vol. 2997, pp. 181–196). Berlin: Springer. **DOI:** https://doi.org/10.1007/978-3-540-24752-4_14

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and humanities: A review. *Scientometrics*, *66*(1), 81–100. **DOI:** https://doi.org/10.1007/s11192-006-0007-2

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, *1*(2), 161–169. **DOI:** https://doi.org/10.1016/j.joi.2006.12.001

OECD. (2007). *Revised Fields of Science and Technology (FOS) Classification in the Frascati Manual*. Paris: OECD Publishing.

Ossenblok, T., Engels, T. C. E., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science—A comparison of publication patterns and incentive structures in Flanders and Norway (2005–9). *Research Evaluation*, *21*(4), 280–290. **DOI:** https://doi.org/10.1093/reseval/rvs019

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Read, J. (2010). *Scalable Multi-label Classification*. Doctoral thesis, University of Waikato, Hamilton, New Zealand. Retrieved from https://hdl.handle.net/10289/4645

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier chains for multi-label classification. In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.), *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009*. Berlin: Springer. **DOI:** https://doi.org/10.1007/978-3-642-04174-7_17

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naïve Bayes text classifiers. In T. Fawcett & N. Mishra (Eds.), *Twentieth International Conference on Machine Learning (ICML-2003)* (pp. 616–623). Washington, DC: AAAI Press.

Rip, A., & Courtial, J.-P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, *6*(6), 381–400. **DOI:** https://doi.org/10.1007/BF02025827

Rollin, L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management*, *17*(2), 69–76. **DOI:** https://doi.org/10.1016/0306-4573(81)90028-5

Schapire, R. E., & Singer, Y. (2000). Boost-exter: A boosting-based system for text categorization. *Machine Learning*, *39*(2/3), 135–168. **DOI:** https://doi.org/10.1023/A:1007649029923

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computer Surveys*, *34*(1), 1–47. **DOI:** https://doi.org/10.1145/505282.505283

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine Learning and Knowledge Discovery in Databases. Joint European Conference on Machine Learning and Knowledge Discovery in Databases—ECML PKDD 2011* (vol. 6913). Berlin: Springer. **DOI:** https://doi.org/10.1007/978-3-642-23808-6_10

Shneider, A. M. (2009). Four stages of a scientific discipline; four types of scientist. *Trends in Biochemical Sciences*, *34*(5), 217–223. **DOI:** https://doi.org/10.1016/j.tibs.2009.02.002, **PMID:** 19362484

Sievert, M. C., & Andrews, M. J. (1991). Indexing consistency in Information Science Abstracts. *Journal of the American Society for Information Science*, *42*(1), 1–6. **DOI:** https://doi.org/10.1002/(SICI)1097-4571(199101)42:1<1::AID-ASI1>3.0.CO;2-9

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, *12*(1), 133–152. **DOI:** https://doi.org/10.1016/j.joi.2017.12.006

Sjögårde, P., & Ahlgren, P. (2019). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. *Quantitative Studies of Science*, *1*(1), 207–238. **DOI:** https://doi.org/10.1162/qss_a_00004

Sorower, M. S. (2010). A literature survey on algorithms for multi-label learning. Technical Report, Corvallis: Oregon State University.

Stichweh, R. (1992). The sociology of scientific disciplines: On the genesis and stability of the disciplinary structure of modern science. *Science in Context*, *5*(1), 3–15. **DOI:** https://doi.org/10.1017/S0269889700001071

Stichweh, R. (2003). Differentiation in science: Causes and consequences. In G. H. Hardon (Ed.), *Unity of Knowledge in Transdisciplinary Research for Sustainable Development* (vol. 1, pp. 82–90). Oxford: EOLSS Publishers.

Sugimoto, C. R., & Weingart, S. (2015). The kaleidoscope of disciplinarity. *Journal of Documentation*, *77*(4), 775–794. **DOI:** https://doi.org/10.1108/JD-06-2014-0082

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464–2476. **DOI:** https://doi.org/10.1002/asi.23596

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, *29*(1), 1–25. **DOI:** https://doi.org/10.18352/lq.10285

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal for Data Warehousing and Mining*, *3*(3), 1–13. **DOI:** https://doi.org/10.4018/jdwm.2007070101

van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, *68*(3), 377–393. **DOI:** https://doi.org/10.1007/s11192-006-0118-9

Vancauwenbergh, S., & Poelmans, H. (2019a). The creation of the Flemish research discipline list, an important step forward in harmonising research information (systems). *Procedia Computer Science*, *146*, 265–278. **DOI:** https://doi.org/10.1016/j.procs.2019.01.075

Vancauwenbergh, S., & Poelmans, H. (2019b). The Flemish research discipline classification standard: A practical approach. *Knowledge Organisation*, *46*, 354–363. **DOI:** https://doi.org/10.5771/0943-7444-2019-5-354

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. **DOI:** https://doiorg/10.1002/asi.22748

Yan, E., Ding, Y., Milojević, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics*, *6*(1), 140–153. **DOI:** https://doi.org/10.1016/j.joi.2011.10.001

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*, 767–786. **DOI:** https://doi.org/10.1007/s11192-014-1321-8

Zhang, M.-L., Li, Y.-K., Liu, X.-Y., & Geng, X. (2018). Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, *12*, 191–202. **DOI:** https://doi.org/10.1007/s11704-017-7031-7

Zhang, M.-L., & Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1819–1837. **DOI:** https://doi.org/10.1109/TKDE.2013.39