# Automatically Assigning Research Methods to Journal Articles in the Domain of Social Sciences

**Judith Eckle-Kohler[1,2], Tri-Duc Nghiem[1], Iryna Gurevych[1,2]**
[1]Ubiquitous Knowledge Processing (UKP) Lab, Department of Computer Science,
Technische Universität Darmstadt, Germany
[2]Information Center for Education, German Institute for International Educational Research,
Germany
http://www.ukp.tu-darmstadt.de

## ABSTRACT

We investigate the automatic assignment of research methods to journal articles from the domain of Social Sciences. We employ Computer Science and Computational Linguistics methodology to perform this automatic assignment of metadata. The multi-label classification system we present uses only abstracts and titles of journal articles as input. Our best system is able to assign the important research methods *empirical* and *quantitative empirical* with F-scores of 0.67 and 0.68. These research methods are in the focus of many recent manual analyses of publications databases. Our classification approach could be applied to automatically analyze large publications databases and databases of bibliographic references according to the use of *empirical* and *quantitative empirical* methods.

## Keywords

Metadata assignment, research methods, multi-label classification

## INTRODUCTION

The question which research methods are used in scientific publications is of major importance for researchers in many different disciplines, e.g., Social Sciences, Educational Research, Psychology or Medicine. Answers to this question are not only needed when searching for publications using a particular method, but also when analyzing the use of methods in a scientific discipline based on a large publications database. Such an analysis is performed, for instance, when particular scientific subfields are to be characterized according to the research methodologies used (Wallace et al. 2010), or if the use of research methods over time is investigated, e.g., (Sondergeld & Botte 2013).

However, research methods are typically not available as metadata, in contrast to other common bibliographic metadata, such as title, author, or abstract. Previously, the use of methods in a scientific discipline has typically been studied by manually analyzing publication samples. So far, techniques for the automatic identification of methods used in publications have been investigated and developed only by few previous works.

The automatic identification of methods used in publications can either take the full-texts as a basis, or the abstracts. Abstracts of publications are not only available in publications databases (e.g., IEEE Xplore digital library[1]), but also in large databases of bibliographic references, e.g., PsycINFO[2], an abstracting and indexing database in the behavioral sciences and mental health.

While using the full-text for the automatic identification of methods clearly has the advantage that rich textual information on the content of a paper can be exploited, this approach can only be pursued, if the full-texts are available. However, an Open Access policy is not common yet in all scientific disciplines (e.g., Medicine, Psychology). Many large publications databases offer only abstracts and abstracts-based keyword search for free, while access to full-texts is restricted (e.g., IEEE Xplore). Similarly, databases of bibliographic references often provide links to the full-texts only at a fee (e.g., PsycINFO).

Therefore, we explore automatic methods for research method assignment that use publication abstracts only. Our main contribution can be summarized as follows: We show that a multi-label classification system that uses only abstracts and titles as input is able to assign particular

[1] http://ieeexplore.ieee.org

[2] http://www.apa.org/pubs/databases/psycinfo/index.aspx

important research methods, such as *empirical* and *quantitative empirical* with a much better performance than that of current systems for automatic keyphrase assignment.

The remainder of this paper is organized as follows: the next section will summarize and discuss related work. Then, we will present the main research questions addressed in this paper. The subsequent two sections give a detailed description of the data and machine learning methods used in our experiments. Then, we describe the experimental setup and present the results of our experiments. We proceed by discussing these results and conclude by giving an outlook to future work.

## RELATED WORK

### Manual analysis of research literature

Wallace et al. (2010) manually analyzed full-texts of knowledge management articles in order to identify the research methodologies used. Their goal is to characterize the professional discipline "knowledge management" based on research methods use. While they propose to use this kind of manual content analysis for other professional disciplines as well (e.g., Library and Information Science, Computer Science), the high effort of such manual analyses can be expected to be a restricting factor.

### Automatic Method Extraction

Automatic Method Extraction aims at extracting method mentions automatically from the scientific texts.

Gupta & Manning (2011) automatically extract features from abstracts of scientific papers in the domain of Computational Linguistics. Features extracted include the research focus, domain, and the technique used. For evaluation, they use a dataset of 474 abstracts along with the titles, which has been manually annotated for the three categories focus, domain, and technique. The technique terms used for manual annotation all occurred in the abstracts. However, it cannot generally be assumed that values used in instances of metadata elements (e.g. values of the attribute "technique") occur in the abstracts or in the full-texts themselves.

Houngbo & Mercer (2012) describe an approach to automatically build a terminology resource of research method terms. They employ a rule-based approach to extract lexical and terminological information on research methods from sentences that have been identified as containing method mentions in a previous step. These sentences are identified based on method terms ending with *method, analysis, algorithm, approach, model*. To mine further method terms, they apply algorithms for Named Entity Recognition. While their work is useful in the context of building a resource of method terms, it cannot be applied for assigning predefined method terms to abstracts, which do not necessarily contain these method terms at all.

### Automatic Keyphrase Assignment

The problem of automatically assigning keyphrases to scientific articles has been studied as part of the SemEval-2010 initiative of shared tasks (Kim et al. 2010). The performance of the submitted systems was evaluated by matching the keyphrases assigned by the system against manually assigned ones. The dataset used for the competition consisted of 284 scientific articles (i.e., full-texts) with keyphrases assigned by both their authors and readers. By matching this manually assigned answer set with the keyphrases provided by the systems, a number of evaluation metrics were calculated, in particular, the so-called F-score, the harmonic mean of precision and recall.

For each system, the performance over the top 5, 10, and 15 candidate keyphrases returned by the system was determined. The best system achieved an F-score of 27.5% for the top 15 keyphrase candidates. The other top-performing systems achieved F-scores in the upper twenties as well. This low number can partially be explained when looking into the way humans assign keyphrases to papers. Keyphrase assignment is a subjective task, i.e., humans typically do not agree very well on which keyphrases to assign. Matching reader-assigned keyphrases with author-assigned keyphrases yielded the low F-score of 33.6%, which can be considered as an upper bound for automatic systems (Kim et al. 2010).

### Multilabel Classification

Multilabel classification is the problem of classifying documents into multiple categories, i.e., assigning multiple labels to each document.

Spat et al. (2007) consider the problem of classifying German narrative clinical texts into multiple predefined medical field categories. Examples of narrative clinical texts are physician's letters, clinical findings or documentation of disease progression. Their multilabel classification system could assign multiple medical field categories with a high F-score. The system was evaluated on a dataset of 1500 documents which had been labeled by one senior physician only. This domain expert assigned predefined medical field categories to the documents. The main reasons of the good performance are (i) the low number of medical field categories (eight in total) and (ii) the fact that these categories are well reflected by the use of specific medical terminology in the texts.

## RESEARCH QUESTIONS

This work addresses the following research questions:

- Is it possible to automatically assign the scientific method used in Social Science papers based on abstracts and titles of publications and other metadata?
- Are different research methods reflected to varying extent in abstracts of scientific papers (i.e., can an

automatic system assign some methods better than others)?

- Which features are most important when classifying papers based on their abstracts: textual features or metadata features?

## DATA AND METHODS

We consider the problem of automatically assigning multiple method terms to scientific papers in the domain of Social Sciences. As a basis for the automatic assignment, only abstracts and titles of the publications are considered.

### Data

We use data from SOLIS[3] (Social Science Literature Information System), a database of bibliographic references and content descriptive metadata for German social science literature, which has been developed by GESIS, Leibniz Institute for the Social Sciences[4].

SOLIS is a database of metadata which also provides information on the research methods used. Literature from many different subfields of social science is available in SOLIS, ranging from Sociology, Political Science and Educational Research to Communication Sciences and Demography.

These research methods have been assigned by professional indexers at GESIS on the basis of the full-texts. The method terms to be assigned are given as a controlled vocabulary[5] which is used as a guideline document by the indexers. Three facets are used to categorize the controlled vocabulary. Two facets group preferred method terms (called descriptors), and one facet groups non-preferred method terms (i.e., lexical variants, also called non-descriptors) along with references to the preferred terms. Examples of these non-preferred terms are given in table 4.

While the first facet is used to describe the primary designation of the research, the second facet captures the methodological design and the data collection procedures used. The method terms in the first facet are organized by hierarchical relationships. For instance, *empirisch* is a broader term which covers the two narrower terms *empirisch-qualitativ* and *empirisch-quantitativ*. The controlled vocabulary of SOLIS method terms is available both in German and in English[6].

| |
|---|
| empirisch (empirical: data collected or provided by primary survey, secondary analysis or on the basis of already existing material)<br>• empirisch-qualitativ (qualitative empirical: collective term referring to methods used for analyzing empirical data under the interpretative paradigm)<br>• empirisch-quantitativ (quantitative empirical) |
| anwendungsorientiert (applied research: scientific consultation, evaluation or development of measures, programs or projects)<br>• praktisch-informativ (practical information: description of facts relevant for social sciences but without scientific claim)<br>• Theorieanwendung (theory application: explicit application of a specific theory to analyze concrete circumstances)<br>• Evaluation (evaluation) |
| Grundlagenforschung (basic research)<br>• Methodenentwicklung (development of methods, also development of measuring instruments)<br>• Theoriebildung (theory formation, also theory comparison and model construction)<br>• wissenschaftstheoretisch (epistemological) |
| Zustandsanalyse (condition analysis)<br>• deskriptive Studie (descriptive study: systematic compilation of facts with scientific background)<br>• Dokumentation (documentation, e.g. bibliography, biographies, statistics, progress report, textbook) |
| historisch (historical: when the main research object dates back more than 15 years ago) |
| normativ (normative: evaluative in the sense of social ethics, practical philosophy or ontology) |

**Table 1 Method terms and their definitions, which are used in the DIPF project to describe the primary designation of the research.**

[3] http://www.gesis.org/en/services/research/solis-social-science-literature-information-system/

[4] http://www.gesis.org

[5] See ANSI/NISO Z39.19-2005 Standard (Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies).

[6] GESIS – Leibniz-Institut für Sozialwissenschaften (2007): Methodenliste für die Datenbanken SOFIS und SOLIS http://www.gesis.org/fileadmin/upload/dienstleistung/tools_standards/Methodenliste_SOFIS-SOLIS.pdf

In the context of a current research project at the German Leibniz institute DIPF[7], these method terms have been re-organized (Kempf et al. 2013) in order to improve the distinction between the research goal of a publication on the one hand, and the methodological design and data collection procedures used on the other hand.

A part of the re-organized classification scheme is shown in table 1. It lists the re-organized method terms from the first facet of the SOLIS controlled vocabulary, which describe the primary designation of the research (i.e., the research goal) presented in a publication.

From the SOLIS database, we extracted a dataset of 2401 documents of type "journal article" which have been published since 1995. The subset of journal articles written in German consists of 1992 documents. In this subset, fifteen different method terms from table 1 occur in the annotated methods. We divided this dataset into a training set (1798 documents) and a test set (194 documents).

To represent the content of each document, we consider its title and abstract, which are both available as metadata. Table 2 shows statistics of this dataset.

| | **Training set** | **Test set** |
|---|---|---|
| Number of abstracts | 1798 | 194 |
| Vocabulary size | 35563 | 8314 |
| Number of tokens | 260353 | 28911 |
| Average number of tokens per document | 144.80 | 149.03 |
| Size of label (method term) set | 15 | 14 |
| Average number of labels per document | 2.00 | 1.97 |
| Size of index term set | 20549 | 2243 |
| Average number of index terms per document | 11.43 | 11.56 |

**Table 2 Statistics of training set and test set.**

In the following sections, we present our research on automatically assigning the method terms listed in table 1 to German journal articles from SOLIS. Unfortunately, the difficulty of this task for human indexers is unknown, since the SOLIS data provides no evidence of inter-annotator agreement for the task of assigning method terms to documents; methods have typically been assigned by one person only, either the author or an indexer.

## Methods

### *Multi-label Classification*

For datasets consisting of instances that are annotated with more than one label, a multi-label classification problem has to be solved. Over the last few years, many approaches have been developed to handle this problem. They can be grouped into two main categories: *algorithm adaption* and *problem transformation*.

**Problem transformation methods** transform a multi-label classification problem to single-label classification. In our experiments, we test the performance of different methods: Binary Relevance (BR), Classifier Chains (CC) and **Ra**ndom **k** lab**EL**sets.

Binary Relevance learns each label as binary classifier. It divides the original data set into $q$ (number of labels) data sets with all instances in the original data and considers them as positive and negative examples for each label.

Classifier Chains (Read et al. 2011) is also a binary relevance method but it is different from BR in that it forms a classifier chain of all previous classifiers for each binary classifier model. Therefore, CC can model the label correlations.

The Random k-labelsets (RAkEL) algorithm (Tsoumakas et al. 2011) iteratively constructs an ensemble of m Label Powerset (LP) classifiers (Trohidis et al. 2011). In each iteration, it randomly selects a k-labelset, then learns a Label Powerset classifier.

In our experiments, the following methods are associated with single-label classifiers: Random Forest (Breiman 2001) and Support Vector Machines (SVM).

**The algorithm adaption methods** extend classification algorithms to deal with multi-label data. We examine ML-kNN (Zhang & Zhou 2007), a method based on the k Nearest Neighbors learning algorithm. ML-kNN first selects the k nearest examples in the training set for each test example. Then it uses the maximum posterior principle based on statistical information achieved from these neighbor examples to determine the label set for the test example.

### *Features*

We consider four types of features: meta data features, textual content features, linguistic features and controlled vocabulary features, see table 3.

[7] http://www.dipf.de/en/projects/development-and-changing-dynamics-of-educational-research

| | Feature Name | Type |
|---|---|---|
| Meta Data Features | Author | Boolean |
| | Conference Name | Boolean |
| | Index Terms | Boolean |
| Textual Content Features | LDA Topic Model | Numeric |
| | tf-idf scored keyword | Numeric |
| | Number POS tag | Numeric |
| | Named Entity | Boolean |
| | n-gram | Numeric |
| | Bag-Of-Words (BOW) | Numeric |
| Linguistic Features | n-gram over POS tags | Numeric |
| | Verb classes | Numeric |
| | POS tag of verbs | Numeric |
| Controlled Vocabulary Features | Terminology Feature | Numeric |

**Table 3 List of features and their types.**

**Meta data features** are extracted from the metadata associated with each document. We select author names, conference names, and index terms and represent them by boolean features.

**Textual content features** are extracted based on the abstracts. We use Latent Dirichlet Allocation (LDA) (Blei et al. 2003) to build 100 topic models. For each abstract, the top 3 most frequent topics are assigned. We also employ tf-idf scores[8] to select top 2 highest ranked tokens for each abstract as features. Besides, we also use Named Entities (i.e., person names), Bag-Of-Words (BOW) and bi-gram features.

**Linguistic features** are based on linguistic annotation of the abstracts. We use Parts-Of-Speech (POS) tags to determine bi-grams over POS tags. In addition, we use POS tags to identify numbers and to capture information on the verb tense. Moreover, we determine the relative frequency of verb lemmas which belong to particular German semantic verb classes.

**Controlled vocabulary features** are assigned using the list of preferred method terms and their associated non-preferred terms given in the controlled vocabulary of SOLIS.

| Non-preferred term | Preferred term |
|---|---|
| Bedarfsforschung (demand analysis) | anwendungsorientiert (applied research) |
| Implementation (implementation) | anwendungsorientiert (applied research) |
| ethnographisch (ethnographical) | deskriptive Studie (descriptive study) |
| biographisch (biographical) | Dokumentation (documentation) |

**Table 4 Examples of non-preferred method terms and their associated preferred terms in the controlled vocabulary of SOLIS.**

*Evaluation Metrics*

In order to assess the performance of the multi-label classification system, several metrics can be used. They can be categorized into three groups: example based measures, label based measures and ranking based measures. Example based measures score each abstract in the test set and then average them to determine the final score. Metrics such as hamming loss, accuracy, precision, recall, F1, exact match are example based measures. Label based metrics calculate the average score across all labels. Macro and micro measures (precision, recall, F1) are label based metrics. Ranking based metrics evaluate the quality of the ranked labels for each example (i.e., abstract). Further metrics and detailed formulas can be found in (Madjarov et al. 2012)

## EXPERIMENTAL SETUP

The Weka (Hall et al. 2009) Machine Learning system is used with Meka[9] and Mulan extensions (Tsoumakas et al. 2010) in order to perform multi-label classification. To extract features, we use ClearTK (Ogren et al. 2008).

For linguistic preprocessing, we apply text segmentation, POS tagging and lemmatization using the TreeTagger for java (TT4J[10]). To normalize German compounds, we employ the compound splitter BananaSplit[11]. We postprocess the lemmas annotated by the TreeTagger in order to capture lemmas of particle verbs in sentences, where the particle occurs separately from the verb stem. Consider as an example the sentence *Wir **fangen** morgen **an** – We begin tomorrow*, where the lemma *fangen* annotated by the TreeTagger is replaced by the lemma *anfangen*. This post-processing of verb lemmas is important in order to identify verbs which belong to particular semantic verb classes. We use jgibblda[12] for LDA topic

[8] Tf-idf is a weighting scheme that scores terms based on their document frequency.

[9] http://meka.sourceforge.net

[10] https://code.google.com/p/tt4j/

[11] http://niels.drni.de/s9y/pages/bananasplit.html

[12] http://jgibblda.sourceforge.net/

|  |  | Baseline | RAK-SVM | RAK-RF | CC-SVM | CC-RF | BR-SVM | BR-RF | ML-kNN |
|---|---|---|---|---|---|---|---|---|---|
| Example | Accuracy | 0.403 | **0.427** | 0.295 | 0.299 | 0.192 | 0.298 | 0.27 | 0.279 |
|  | H_Loss | 0.131 | **0.125** | 0.165 | 0.151 | 0.124 | 0.152 | 0.153 | 0.155 |
|  | Exact_match | 0.180 | **0.196** | 0.103 | 0.077 | 0.103 | 0.082 | 0.067 | 0.082 |
|  | Precision | 0.503 | **0.524** | 0.377 | 0.425 | 0.579 | 0.424 | 0.41 | 0.406 |
|  | Recall | 0.516 | **0.539** | 0.398 | 0.416 | 0.191 | 0.416 | 0.382 | 0.39 |
| Label | Micro F1 | 0.509 | **0.532** | 0.387 | 0.421 | 0.287 | 0.421 | 0.396 | 0.398 |
|  | Macro F1 | 0.481 | **0.505** | 0.505 | 0.372 | 0.228 | 0.372 | 0.35 | 0.355 |
| Ranking | One Error | 0.474 | **0.454** | 0.613 | 0.624 | 0.624 | 0.619 | 0.49 | 0.588 |

**Table 5 Overall results with various multi-label (RAkEL, CC, BR) and base (Random Forest, SVM) classifiers and algorithm adaption method ML-kNN using the best configuration.**

modeling, and the Stanford Named Entity Recognizer for annotating Named Entities.

We test the performance of several problem transformation methods: RAkEL, BR, CC using Random Forest and SVM classifiers. For the RAkEL, we set m=10 and k=7 (half the size of the label set).

The baseline system uses RAkEL-SVM configuration, and n-gram (n=1,2) features.

### Parameters and Features Selection

Within a 10-fold cross-validation setup based on the training data, we examined several parameters and features. For the n-gram features, the best configuration uses the top 500 most frequent uni-grams and bi-grams. The best combination of features is uni-grams, bi-grams, POS bi-grams, BOW, and the number feature. We apply chi-square feature selection to reduce the feature space after applying problem transformation to transform the multi-label data into single-label data.

Evaluating the system on our test set, we get the best result based on the 850 highest ranked features of uni-grams, bi-grams, POS bi-grams, BOW, and the number feature.

### Results

Table 5 shows the results of the baseline system and other configurations on the reduced feature space (850 highest ranked features of uni-grams, bi-grams, POS bi-grams, BOW, and the other numeric features). We perform an automated evaluation by matching the method terms assigned by the system with the human assigned method terms given in our test dataset.

As can be seen from table 5, the RAkEL – SVM system outperforms the other configurations. Our baseline system using RAkEL – SVM and only n-gram (n=1,2) features is a competitive baseline with a performance  slightly lower than that of the best configuration.

Regarding the combination of features, we observed that using some textual content features such as the LDA Topic model, tf-idf scored keywords, Named Entities did not improve the results. Because the texts are rather short (about 150 tokens per document, see table 2), the system could not detect discriminative features: either their frequencies are low or they play no discriminative role.

The metadata features "author", "conference name" and "index terms" introduced much noise and made the results worse. Terminology features, such as the controlled vocabulary features employed in our experiments, could in theory provide good signals to distinguish documents according to the research methods. But since the non-preferred terms listed in the controlled vocabulary typically occur in the full-texts and not in the abstracts, using them did not improve the performance of the system.

Table 6 lists F-scores for each method term label separately, sorted by decreasing F-score. The second and third column give the number of instances in the training and test set. The method term with the highest F-score is *quantitative empirical*, closely followed by *empirical* and *descriptive study*. This is due to the fact that for these three method terms a large number of instances is available in the training data (more than 500 examples each). For method terms with less than 100 instances, the system is not able to provide meaningful results.

We further observe that the method *quantitative empirical* can better be assigned automatically than the methods *empirical* and *descriptive study*, because the system achieves a higher F-score, but with a much lower number of instances.

## DISCUSSION

Our experiments show that even with a small training set of 1798 documents where the maximum number of instances for each method term label is roughly between 500 and 750, our multi-label classification system is able to perform method term assignment with an F-score over 0.6.

| Labels | Size of train | Size of test | F1 |
|---|---|---|---|
| empirisch-quantitativ (quantitative empirical) | 516 | 52 | 0.678 |
| empirisch (empirical) | 754 | 70 | 0.674 |
| Deskriptive Studie (descrip-tive study) | 700 | 74 | 0.611 |
| Grundlagenforschung (basic research) | 350 | 41 | 0.492 |
| Methodenentwicklung (development of methods) | 154 | 23 | 0.368 |
| historisch (historical) | 188 | 22 | 0.359 |
| empirisch-qualitativ (qualitative empirical) | 231 | 24 | 0.326 |
| anwendungsorientiert (applied research) | 305 | 28 | 0.237 |
| Dokumentation (documentation) | 43 | 10 | 0.182 |
| Theorieanwendung (theory application) | 166 | 23 | 0.06 |
| Evaluation (evaluation) | 38 | 2 | 0 |
| normativ (normative) | 24 | 1 | 0 |
| praktisch-informativ (practical information) | 30 | 6 | 0 |
| Theoriebildung (theory formation) | 69 | 6 | 0 |
| wissenschaftstheoretisch (epistemological) | 32 | 0 | 0 |

**Table 6 Average F1 scores for the method term labels from the first facet of the SOLIS controlled vocabulary.**

This performance is significantly better than the performance of state-of-the-art systems for automatic keyphrase assignment (Kim et al. 2010), which achieve F-scores in the upper twenties. However, there is still room for improvement, considering the F-score of more than 0.8 that has been reported for automatic assignment of medical field categories (Spat et al. 2007).

Another way of assessing the performance of our system would be to compare its performance with the inter-annotator-agreement of human annotators on the task of method assignment. The human inter-annotator-agreement can be considered as an upper bound for automatic systems. However, this is not possible, since the SOLIS data provide no evidence of inter-annotator agreement.

Using just abstracts and titles as a basis for feature extraction, makes the problem of learning which method terms to assign very hard for the system, even more so, as the human assigned methods take the full-texts as a basis. This imbalance regarding the information available to the system (only information from abstract and title) and to humans (full-texts) disadvantages the system in the automated evaluation.

Considering the fact that multi-label classification typically requires large amounts of training data, we expect to obtain significant improvements by using a larger training set in the future.

The three method terms which could be assigned with high F-scores include the research methods *empirical* and *quantitative empirical*, which stand in a hierarchical relationship. Both research methods are in the focus of many recent manual analyses of publications databases, e.g., in Sondergeld & Botte (2013). Therefore, the automatic classification approach presented in this paper could be applied to automatically analyze large publications databases and databases of bibliographic references according to the use of *empirical* and *quantitative empirical* methods.

## CONCLUSION AND FUTURE WORK

We investigated the automatic assignment of research methods to journal papers in the domain of Social Sciences based on their abstract and title only. In our experiments, we explored a wide range features and compared many different multi-label classifiers. Our best system was able to assign the important research methods *empirical* and *quantitative empirical* with F-scores of 0.67 and 0.68.

In future work, we plan to explore the possible performance gain for automatic research method assignment when using the full-texts in addition. In this context, we will also increase the number of research methods considered and use a more fine-grained set of method terms, including details of the methodological paradigm used.

We also plan to address the problem of missing evidence of inter-annotator agreement for the task of research method assignment in the Social Science domain by carrying out an annotation study.

## ACKNOWLEDGMENTS

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806.

Our special thanks go to Dr. Andreas Oskar Kempf from GESIS for providing the data for this research and for giving us very helpful comments and background information. We also thank Ute Sondergeld, Alexander Botte and Dr. Daniela Oelke for their contributions to this work, as well as the anonymous reviewers for their valuable feedback.

## REFERENCES

Blei, D.M., Ng, A.Y. & Jordan, M.I., 2003. Latent Dirichlet Allocation J. Lafferty, ed. *Journal of Machine Learning Research*, 3(4-5), pp. 993–1022.

Breiman, L., 2001. Random Forests R. E. Schapire, ed. *Machine Learning*, 45(1), pp. 5–32.

Gupta, S. & Manning, C., 2011. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 1–9.

Hall, M. et al., 2009. The WEKA Data Mining Software: An Update. In *SIGKDD Explorations*. New York, NY, USA, pp. 10–18.

Houngbo, H. & Mercer, R.E., 2012. Method Mention Extraction from Scientific Research Papers. In *Proceedings of COLING 2012*. Mumbai, India, pp. 1211–1222.

Kempf, A.O. et al., 2013. Metadatenfeld „Methoden“. In *Projektdatenbank Bildungsforschung: Bericht zu den Arbeitspaketen 2 und 3 des Projekts „Monitoring Bildungsforschung (MoBi)“. Anhang IIa.*

Kim, S.N. et al., 2010. SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, pp. 21–26.

Madjarov, G. et al., 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9), pp. 3084–3104.

Ogren, P.V., Wetzler, P.G. & Bethard, S.J., 2008. ClearTK: A UIMA Toolkit for Statistical Natural Language Processing. In *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*. Marrakech, Morocco, pp. 32–38.

Read, J. et al., 2011. Classifier chains for multi-label classification W. Buntine et al., eds. *Machine Learning*, 85(3), pp. 333–359.

Sondergeld, U. & Botte, A., 2013. Monitoring Bildungsforschung : Mehrdimensionale szientometrische Untersuchung eines interdisziplinären Forschungsfeldes. In H.-C. Hobohm, ed. *Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten : Proceedings des 13. Internationalen Symposiums für Informationswissenschaft (ISI 2013)*. Potsdam, Germany: Glückstadt: Hülsbusch.

Spat, S. et al., 2007. Multi-label Text Classification of German Language Medical Documents. In *MedInfo*. pp. 1460–1461.

Trohidis, K. et al., 2011. Multi-label classification of music by emotion. *EURASIP Journal on Audio Speech and Music Processing*, 2011(1), p. 4.

Tsoumakas, G., Katakis, I. & Vlahavas, I., 2010. Mining Multi-label Data. In O. Maimon, L Rokach (eds.) *Data Mining and Knowledge Discovery Handbook*. Springer, pp. 667–685.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2011). Random k-Labelsets for Multilabel Classification. IEEE Transactions on Knowledge and Data Engineering, 23(7), pp. 1079–1089.

Wallace, D. P., Fleet, C. V. & Downs, L.J., 2010. The use of research methodologies in the knowledge management literature. *Proceedings of the American Society for Information Science and Technology*, 47(1), pp. 1–7.

Zhang, M.-L. & Zhou, Z.-H., 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7), pp. 2038–2048.