Research Method Classification with Deep Transfer Learning for Semi-Automatic Meta-Analysis of Information Systems Papers

Anna Anisienia
Berlin School of Economics and Law
anna.anisienia@gmail.com

Roland M. Mueller
Berlin School of Economics and Law, University of Twente
roland.mueller@hwr-berlin.de

Anna Kupfer
University of Bamberg
anna.kupfer@uni-bamberg.de

Thorsten Staake
University of Bamberg
thorsten.staake@uni-bamberg.de

Abstract

This paper presents an artifact that uses deep transfer learning methods for the multi-label classification of research methods for an Information Systems corpus. The artifact can support researchers with frequently performed yet time-consuming classification and structure-seeking tasks that are often part of literature analyses. We use a corpus of 5,388 papers from AIS journals and conferences, of which 1,766 have been manually labelled with up to five research methods. The unlabelled papers are used for finetuning the language model, whereas the labelled data are used for training and testing. Our approach outperforms state of the art research method classification that deploy SVM. We show that deep transfer learning models can lead to a better recognition of research methods than shallower word embedding approaches like word2vec or GloVe. The results illustrate the potential of establishing semi-automated methods for meta-analysis.

1. Introduction

Due to the increasing number of scientific articles it is difficult and time-consuming to follow the latest developments and to get an overview of a field of research that is not one's own. The latter, however, is often necessary as transdisciplinary research projects are common in Information Systems (IS) research [18]. As a consequence, attempts to automatically classify scientific publications and extract important concepts have been made. There are some recent advances in automatic meta-analysis of scientific journal and conference contributions [12, 17, 26, 27], leading to an improved quality of the generated analyses. An important subtype of meta-analyses is the creation of an overview of the research methods used [31, 35, 40, 42]. These analyses can also inform about the used philosophical research paradigms (e.g., positivism, interpretivism, critical realism) [29]. However, to the best of our knowledge, no paper has yet been published to automate this task. Using a naïve key word search for research method classification is not sufficient because the description of research methods can be phrased in different terminologies and could also refer to related work. Despite recent progress in automatic meta-analysis of literature, more research is needed to automate the extraction of relevant contents from scientific articles. The results would help to improve search engines and may ultimately lead to tools that generate valuable summaries on their own.

In this context, high hopes rest on deep transfer learning approaches. These approaches refer to multilayer transfer learning approaches for natural language processing (NLP), such as ULMFiT [13], ELMo [33], BERT [5], or OpenAI Transformer [34], which can better capture the semantics of the language, as opposed to shallow word embeddings that have typically been used in the NLP field over the past years. So far, no paper has been published on the use of deep transfer learning and pretrained language models for the automatic research method classification of articles. This paper presents a prototype that applies deep transfer learning to predict the research methods in scientific publications, which facilitates an automatic discovery of crucial research information from large amounts of publications. The current state of the art for classification of research methods uses *Support Vector Models (SVMs),* see Section 2.

This translates into the following research questions *RQ1: Can deep transfer learning be successfully applied to the multilabel classification of research methods compared to the state of the art that use SVMs?*

The goal of RQ1 is to examine the performance of deep transfer learning applied to the multilabel classification of research methods. The application of deep transfer learning is considered successful when it outperforms the baseline model that predicts the most common class for all observations. Additionally, the results of RQ1 are compared against the previous work of other researchers.

URI: https://hdl.handle.net/10125/71357
978-0-9981331-4-0

RQ2: Which form of transfer learning for NLP leads to the best performance of the multilabel classification of research methods?

The goal of RQ2 is to draw a comparison between various ways of approaching deep transfer learning, represented by ULMFiT, ELMo, and OpenAI Transformer. Additionally, RQ2 examines whether those new methods of pretraining can outperform the traditional shallow word representations in form of GloVe vectors [32], as well as embeddings trained from scratch on the target task data. On top of that, we investigate the most effective ways of applying and fine-tuning the latest deep transfer learning models to avoid forgetting of transferred knowledge. As a result, the essential differences in the performance of several pretrained models in various settings are listed and analyzed in the evaluation section. Our approach exceeds the state of the art of research method classification, which rely on Support Vector Machines (SVM) [9]. We show that deep transfer learning models led to better recognition of research methods than shallower word-embedding approaches, such as word2vec or GloVe. From a more general perspective, the results illustrate the possibility of establishing semi-automated methods for knowledge generation in research. In the case presented here, the artifact performs the classification task in seconds, whereas the time span for manual classification was over 400 hours and thus prohibitively long in many contexts. Zooming out further, our contribution provides additional foundations for the discussion on automated knowledge generation in research and touches on aspects such as comprehensibility and impartiality in the creation of knowledge that will serve as a basis for future research.

2. Related Work

First, we present an overview of existing manual research method meta-studies in order to demonstrate the demand for this kind of analysis. Then, we present related literature dealing with theory ontology learning and research method classification. Finally, we discuss the state of the art in transfer learning and language models for NLP.

Research Method Overviews in Information Systems. Several papers have been published that manually analyzed the distribution of different research methods in the Information Systems (IS) discipline. Kupfer [15] developed a research method categorization framework and classified papers from the International Conference on Information Systems (ICIS) and the European Conference on Information Systems (ECIS) from 1995, 2005, and 2015. Vachon et al. [40] studied the evolution of IS research methods from 1984 to 1998 in the journals MIS Quarterly (MISQ) and Information Systems Research (ISR). Palvia et al. [31] looked at the research methods used in the seven major IS journals between 1993 and 2003. Vessey et al. [42] analyzed the research approaches in five top IS journals between 1995 and 1999. Ebeling et al. [8] examined the use of research methods in the main IS conferences between 2006 and 2010. Riedl and Rueckel [35] went a step further by integrating 20 published meta-studies of research method analyses in the IS field. Some papers looked at only one particular journal; for example, Friedrich et al. [10] analyzed 169 papers in the Business & Information Systems Engineering (BISE) journal and analyzed the trend and distribution of research paradigms and methods. Similarly, Dwivedi and Kuljis [6] examined publications in the European Journal of Information Systems (EJIS) from 1997 to 2007. Some papers analyzed the used research methods in an IS subfield, for example Knowledge Management [7, 43]. The existing studies show a demand of the IS community for a regular overview of the trends and distributions of research methods per topic and per journal. However, the previous studies all had to narrow the number of analyzed journals, the years covered, or the foci of interest, because manually analyzing the research methods of a paper is very time consuming. Therefore, an automatic approach of quantitatively analyzing the literature offers new possibilities because larger datasets could be analyzed in shorter time and comparisons become more meaningful.

Theory Ontology Learning for Information Systems Papers. There are several projects in the IS field to better synthesize the ever-increasing number of articles. Nomological networks [12, 19] and theory ontologies [24, 25] allow conceptual search and the automated inference of inter-theory relationships and theory-data maps. Theory ontology learning is the task of using NLP and machine learning methods for extracting these kinds of ontologies. The construct identity detector [17] used NLP algorithms to match constructs that addressed the same real-world phenomenon. CauseMiner [27] is a rule-based NLP system to extract elements of theory ontologies out of IS papers, such as cause, effect, moderator, mediator, context, and relationship direction. DeepCause [26] extends and improves CauseMiner by using different deep learning architectures for this task. A recent call for action in the journal CAIS [18] emphasizes the need for better tools to automatically extract evidences out of IS papers. They also present different knowledge types that could be extracted from papers. Our research is addressing this call to action by focusing on the knowledge type of the used research methods in IS papers.

Automatic Scientific Key-Insights Extraction. Information extraction tries to extract structured information out of unstructured text. For scientific articles, information extraction is used for metadata extraction (author names, affiliations, title, date, journal name, issue, etc.) and key-insight extraction (also called entity recognition, core scientific concepts extraction, or argumentative zoning) [28]. There are only a limited number of papers that tried to identify the research methods of a paper as part of their key-insight extraction [28], see Table 1.

Most papers tried either to classify sentences to different argumentative zones where research methods would be one possibility, or they tried to extract different methods based on the phrases. Only one paper [9] used a predefined taxonomy of research methods and classified the abstracts according to the taxonomy. Most papers that automatically analyzed research methods were in the field of biomedicine or computer science. Only Eckle-Kohler et al. [9] used a corpus of abstracts from the social science field. No paper used any deep learning approaches for the task.

Table 1. Related Work for Research Method Extraction

Pa-per	Scope	Discipline	Methods	Taxo-nomy
[9]	Abstract Classification	Social Science	SVM, RF, kNN	Yes
[11]	Method Phrase Extraction	Biomedicine	Rules, CRF	No
[1]	Sentence Classification	Biomedicine	Naive Bayes	No
[21]	Sentence Classification	Biomedicine	HMM, SVM	No
[39]	Sentence Classification	Computer Science	Naive Bayes	No
[20]	Sentence Classification	Biomedicine	SVM, CRF	No
[36]	Sentence Classification	Computer Science	SVM	No

Our paper provides the following research contributions: (1) developing an artifact that uses deep transfer learning and outperforms the state of the art of research method classification, (2) using full papers (not just abstracts) and classifying them to predefined research methods, and (3) demonstrating the performance based on an extensive IS corpus. Therefore, our contribution might help authors to automate parts of a literature review and therefore mitigate some of the problems associated with the ever-increasing number of papers.

Deep Transfer Learning. In order to apply text mining to the automated knowledge extraction, natural language processing (NLP) researchers used, for a long time, pretrained word vectors such as word2vec [23], fastText [3], or GloVe [32], which enabled the representation of each token by a vector of numbers. Those numbers not only encoded the word itself, as it was the case in one-hot-encoding, but also described the meaning and context of specific tokens [23]. However, in the context of deep learning, those word embeddings were used merely to initialize the first layer of a neural network, of which the remaining layers had to be randomly initialized and trained from scratch based on the data from the target task [37].

Current research has incorporated several ideas for extending the concept of pretrained embeddings, some applied in an analogous way like ImageNet in Computer Vision, others by means of fixed features or attention-based transformer networks. Instead of initializing only the first layer with word embeddings, as most industry-standard neural networks in NLP do, an entire language model (LM), used as a source task, is pretrained and applied to a target task, such as text classification [37]. As it turned out, the transfer of knowledge from the pretrained LM to a target task, such as text classification, significantly improves the model's performance across many different datasets and types of target tasks [13, 33, 34, 41].

The transfer of knowledge from pretrained models allows gaining a richer representation of the natural language and its context beyond just word-level information. Initializing only the first layer of a neural network by using word embeddings can be compared to a pretrained ImageNet model that could only recognize the edges in a convolutional neural network (CNN). Such a shallow transfer learning method would still deliver better performance than a random initialization of weights. However, its use is limited, compared to a fully pretrained LM, which can capture the syntax, semantics, and even complex structures like conjunctions and contradictions [37]. The pretraining process can be viewed as teaching the model to understand English before applying it to a source task, such as the classification of English sentences [14].

A language model has been usually chosen as a source task in deep transfer learning for NLP, as it can be pretrained on any corpus, regardless of the domain. Since this type of model is supposed to predict the next word in a sequence, it eliminates the need for expensive manual annotation. As such, LM is treated in the literature as a self-supervised learning technique [37]. Furthermore, pretrained deep learning models make the target task sample-efficient so that even a small number of labeled observations can achieve a reasonably good performance on various text mining tasks. That is why many researchers and practitioners can benefit from the results of this paper, as state-of-the-art methods in NLP usually require large corpora to obtain useful results.

Language Model Pretraining. All LM involve two steps: (1) LM pretraining and (2) fine-tuning to the target task. Similarly, all of them were pretrained on large corpora like Wikipedia or thousands of books. However, their corpus sizes, settings, and fine-tuning methods differ significantly.

The above-presented methods of pretraining deep contextualized word representations for transfer learning reflect the following three main approaches to the problem: *ULMFiT approach* [13]: pretraining and fine-tuning of an entire LM in the computer vision fashion. *ELMo approach* [33]: pretraining of a language model with a goal of generating task-agnostic fixed feature vectors that can be used as input feature and serve as a replacement for traditionally used shallow word embeddings, such as word2vec. *Transformer approach* [34]: pretraining of attention-based representations that serve as initialization point for parallelizable Transformer-NN.

3. Dataset

This paper used an annotated corpus of journals and conference articles within the domain of IS. Kupfer [15] performed a literature analysis with respect to the utilized research methods and specified a categorization framework, which was employed for annotation of scientific publications from ECIS and ICIS. The data covers the years 1995, 2005, and 2015 and includes 1,023 articles. Building on this corpus, we extended it for all journal papers from the AIS basket of eight (EJIS, ISJ, ISR, JAIS, JIT, JMIS, JSIS, MISQ) for the same years of observation. In total, these were 1,766 papers with up to five manually added research method labels.

Typical deep learning models require more than just over a thousand training examples. This is why transfer learning constitutes an appealing approach, as the knowledge from pretrained models could be transferred to the target task and thus provide additional information that is necessary to learn a mapping of a long textual input (*an entire full-text of a scientific article*) to a multilabel output (*an arbitrary number of research methods*). The use of deep transfer learning could countervail the limitations of a small dataset.

It is worth noting that the precise classification of research methods employed in each article is not a clear-cut issue, even for well-trained researchers. The difficulty of automatically and correctly labeling each document is amplified by the lack of unanimity about the naming standards and a wide range of interpretations.

A scientific publication can simultaneously incorporate multiple research methods. Therefore, our deep transfer learning artifact used multilabel classification to predict up to five research methods for each paper. One particular type of label, *Conceptual*, was assigned to all articles that "develop frameworks, models, and work with theories" [15]. This description, however, applies to many scientific publications, as the innate nature of academic work is to develop new concepts and theories. Thus, a well-developed classification model that generalizes to unseen observations could pick up this pattern and assign the *Conceptual* label to all data points. Generally speaking, the labels *Conceptual*, *Case Study*, *Field Study*, *Survey,* and *Literature Review* account for around 90% of all research methods. This inequality introduces a class imbalance problem.

4. Deep Transfer Classifier for Research Methods

Scientific articles are usually published in the form of PDF documents, which need to be converted to text files before they can be fed to any machine learning model. Since some constituent parts of PDF documents, such as headers and footers, are not useful, they have been removed using the PDF optimizer function from CauseMiner [27]. Four corpora (see Table 2) were created out of two initial datasets:

IS Corpus: a corpus of 5,388 journals and conference full-text articles from AIS journals and conferences.

Annotated Corpus: An Excel spreadsheet containing the information of 1,766 papers that have been labeled using the research method categorization framework of [15] [16]. The manual classification task was undertaken in one round by more than one person. The interrater reliability for a subsample that was categorized by all raters was 0.6 (Mezzich's Kappa), which shows a strong agreement [15]. The labels are research methods with additional metadata, like title, abstract, keyword, and journal.

Table 2. Corpora Overview

Corpus	Details
1	a consolidated version of 1,719 papers that were obtained through matching of the IS Corpus and the Annotated Corpus. The missing papers can be explained by not matching or unavailable PDFs.
2	a modified version of Corpus 1, which replaces all in-text citations, i.e., references such as "(Smith et al., 2018)", with a special token ("xxcite") by means of regular expressions. This dataset was created based on the assumption that a research method mentioned with a citation could have a different semantic meaning than a research method mentioned without citation. In particular, the first could merely reference related work, while the latter could imply the ground-true research method used in the respective paper. The neural network model should learn this pattern through the special token "xxcite".
3	a modified version of Corpus 2, which only uses a concatenated field of Title and Abstract of each paper. The rationale behind this dataset is that transformer-

	based models accept the input of maximum 512 tokens and, thus, are not suitable for processing entire scientific papers. For the sake of comparison of various pretrained word representations, they will be trained with this truncated dataset.
4	a one-hot encoded version of the labeled Corpus 1 created in order to test the one-vs-rest problem transformation approach.

Out of 1,766 papers in the original labeled corpus, 1,719 papers have been used for the artifact, as they have no ambiguous titles and are free of duplicates. Those data form the basis for the Corpus 3, to which an additional *text* field has been generated, based on the *Title* and *Abstract* metadata, concatenated together.

We used 5,388 IS papers, while 1,719 among them have been successfully matched with the labeled database (*1,766 papers*). The amount of available data has a considerable impact on the final results, as deep learning methods require a large number of labeled observations in order to achieve good performance and generalize well to new data, unseen during the training. In the context of this paper, all 5,388 papers were used to create a language model in the self-supervised fashion, while 1,719 papers were applied to train all supervised classification models. For the language model, the dataset has been split, inspired by Merity et al. [22], into 95% training set (*5,118 papers*) and 5% validation set (*270 papers*). For the classifier, in each dataset, 70% of the papers were assigned to the training, 20% to the validation and 10% to the test set, due to the limited number of labeled papers. The training set was required to learn the model's parameters such as weights and biases. By contrast, the validation set was used to provide additional unbiased information that was necessary to adjust the learning rate during the training and to stop early if no further improvement has been observed after a specific number of epochs, declared through a patience parameter. The test set has not been used to train the classifier. Instead, it served as a new, fully independent dataset used to give an unbiased estimate of the model's performance.

Both LM and classifiers were implemented using the encoder–decoder architecture. Both encoders *for LM and for the classifiers* are exactly the same. The decoders, however, differ from each other. While in the LM there is only a linear and a dropout layer that produce probabilities for all words in the vocabulary, the classifier utilizes a more complex architecture that additionally concatenates max- and avg-pooling of its input and passes this through further batch-normalization, dropout, and linear layers. In general, ULMFiT has been implemented in multilabel and one-vs-rest settings.

Besides the deep transfer learning techniques, several simple CNN-based models with task-related embeddings have been constructed. The goal of these experiments was to answer RQ2, i.e., to test whether task-related embeddings, learned through the embedding layer, or shallow transfer learning in form of GloVe, can outperform deep transfer learning techniques.

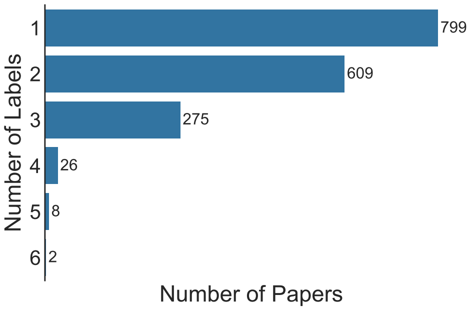

Figure 1. Label Cardinality: Number of research methods assigned per observations

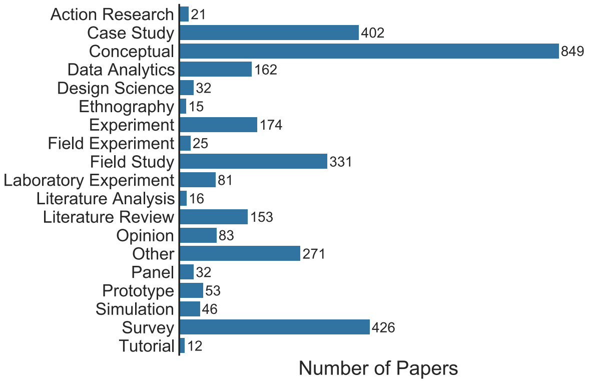

Figure 2. Total Label Counts: Number of documents per label before standardization

One crucial aspect of multilabel classification is the label cardinality, which denotes the number of labels per observation [30]. The distribution shown in Figure 1 demonstrates the number of scientific articles that have a specific number of classes assigned to them, which can be thought of as *label cardinality*. On average, there were two labels assigned per observation. By contrast, the number of scientific articles per label is visualized in Figure 2, which shows the *total label counts*, among which the distribution is dominated by *Conceptual*, *Survey,* and *Case-Study* classes. In Figure 2 the overall class imbalance problem becomes apparent.

Another important aspect of the training process is the sequence length. The distribution of the sequence length is particularly relevant to the CNN model, as its sequences need to be padded to a predefined sequence length. This choice had the effect that longer sequences needed to be truncated and shorter sequences had to be padded to this value. Based on the distribution, the maximum number of tokens (maximum sequence length) in a document has been set to 30,000 to avoid losing any information that might be important to the model.

It is worth mentioning that even though most models used datasets containing only the full-text of papers as input features, much information seems to be encapsulated in the metadata as well. The number of articles associated with each year, journal, or journal type varies. Those features can potentially be used as an auxiliary input to a model. Since the addition of these features is out of the scope of this paper, we focused on the classification of entire documents, and recommend to further elaborate on this as a future research direction.

Based on the review of multilabel classification literature [44] and on metrics used in previous work on research method classification to ensure comparability of results [9], we used as evaluation metrics precision, recall, micro F1, Hamming Loss and exact match. The exact match is the proportion of papers where the classifier predicted all research methods correctly.

Instead of a random search of required hyperparameters, our artifact makes use of a disciplined approach for training neural networks, as proposed by Smith [38]. Inspired by her research paper, each training usually started by performing a learning rate (LR) range test by using a one-cycle policy to find the maximum possible value of the learning rate. The LR search test examined different values, usually from 10^{-5} to 10, and created a plot that helped to decide about the LR value: too large LR can quickly overfit the model, while too small LR would cause slow convergence, i.e., the rate at which the network learns a functional form that generates a mapping of the input features to the desired output.

5. Evaluation of the Deep Transfer Classifier for Research Methods

This section evaluates the artifact's results and explains the model's design choices that were most relevant to the implementation and thus could affect the obtained findings. Table 2 demonstrates the results of conducted experiments, also denoting in brackets, which form of transfer learning (TL) is utilized by each respective method: (1) *shallow TL*, (2) *deep TL*, (3) *no TL*. The dummy classifier assigns all papers to the most common class.

The exact match accuracy cannot be directly computed for the algorithms applying the problem transformation technique. In those cases (*indicated in Table 2 by an asterisk*), this metric has been calculated as the average accuracy of all binary classifiers.

Table 2. Evaluation Results of the Tested Models

		Corpus	Precision	Recall	Micro-F1	Hamming Loss	Exact Match
00	Dummy Classifier	NA	0.55	0.31	0.39	0.11	0.05
01a	GloVe Fixed 100d (shallow TL)	1	0.66	0.15	0.25	0.11	0.05
01b	GloVe Fixed 300d (shallow TL)	1	0.63	0.42	0.51	0.10	0.23
02	GloVe Trainable 300d (shallow TL)	1	0.63	0.47	0.54	0.10	0.25
03	ELMo Small & GloVe 100d Fixed (shallow & deep TL)	1	0.55	0.48	0.51	0.11	0.19
04	ELMo Medium & GloVe 100d Trainable (shallow & deep TL)	2	0.54	0.42	0.47	0.11	0.16
05	ELMo Large & GloVe 100d Trainable (shallow & deep TL)	2	0.42	0.31	0.36	0.13	0.16
06	SVM (no TL)	3	0.46	0.43	0.44	0.14	0.86*
07	OpenAI Transformer (deep TL)	3	0.72	0.35	0.47	**0.09**	0.20
08	ULMFiT Multilabel (deep TL)	1 & 2	0.67	0.51	0.58	**0.09**	0.24
09	ULMFiT One vs Rest (deep TL)	4	**0.74**	**0.64**	**0.66**	**0.09**	**0.91***
10	Target-Task Embeddings (no TL)	1	0.62	0.54	0.58	**0.09**	0.28

6. Discussion

In this section, the results of the presented artifact will be interpreted and compared against the current literature and the research questions.

Performance of Deep Transfer Learning. The primary hypothesis of this paper is centered around the question of whether cutting-edge deep transfer learning techniques can be successfully applied to a problem of multilabel classification of research methods in scientific articles. After running a series of carefully designed experiments, the hypothesis can be confirmed. All deep transfer learning techniques, which were applied with no additional feature engineering, surpassed the performance of a simple baseline model, and some of them outmatched the state of the art in the literature of research method classification.

Comparison Against the Literature. The best test set exact match, micro-F1, and hamming loss, that were achieved on this multilabel problem in existing literature, are 0.196, 0.532 and 0.125, respectively [9]. In contrast, the artifact's best model in the multilabel setting (Model 10) achieved a test set performance of 0.28, 0.58, and 0.09. The one-vs-rest approach (Model 09) obtained an exact match score of 0.91, a micro-F1 of 0.66, and a hamming loss of 0.09. The artifact's deep learning model that was trained on full-texts of entire documents showed considerable improvements (the absolute increase in the exact match ranged from 8% to

71%, and in the micro-F1 score it ranged from 8% to 13%) over the previous state of the art, which was obtained using SVM trained only on abstracts.

The obtained results favor deep learning over simpler ML algorithms and confirm the statement that the research method cannot be identified just by investigating the title and abstract alone. For a better comparability, Model 06 in Table 2 demonstrates the results with the classifier implemented using SVM like Eckle-Kohler et al. [9] were using. Even though this model seems to capture the patterns presented in the dataset surprisingly well—given its simplicity, it is significantly outperformed by deep learning models.

Comparison between different deep transfer learning models. As shown in Section 5, ULMFiT has emerged as the most effective Deep TL method, which outperformed the classification models from the literature, addressing the same problem.

As far as the ULMFiT's configuration is concerned, the default vocabulary size of 60,000 and using the same Corpus 1 for both LM and the classifier have proven to work sufficiently well. The fine-tuning of the last layer for many epochs has led to the most significant improvements in the model's performance. By contrast, when the early layers have been unfrozen too early, the results started deteriorating, which indicates that the network started forgetting the transferred knowledge. The rationale behind this behavior can be explained by the fact that the last layer of the Pooling Linear Classifier is the least general, i.e., the most task-specific [13:5], which is why it had to be trained long enough to achieve the best possible classification results.

Among all tested configurations of ELMo, the *small* version along with fixed 100-dimensional GloVe vectors delivered the best results. However, very similar performance has been observed by applying shallow transfer learning, based on 300-dimensional fixed GloVe representations. However, if the same GloVe embeddings were trainable, the performance improved further, ultimately outperforming ELMo across almost all evaluated metrics.

Even though OpenAI Transformer was trained only on the two metadata fields *Title* and *Abstract*, it was able to outmatch the performance of fixed 100-dimensional GloVe vectors, trained on entire documents. Furthermore, an investigation of different encoder implementations of OpenAI Transformers revealed that CNN, overall, constitutes a more robust architecture than LSTM. The experiments have repeatedly shown that LSTM keeps forgetting the recognized patterns, when encountered with very long input sequences.

Comparison and interpretation of the two best models. In the following subsection, the two best models, indicated in Table 2 as Model 09 and Model 10, will be compared. Model 09 has been obtained using the problem transformation method and LSTM-based ULMFiT as a deep transfer learning technique. In contrast, Model 10 has been created by applying the algorithm adaptation method with a CNN-architecture and embeddings learned from scratch, based on the target task data. This comparison should help to answer RQ2 with respect to which form of transfer learning for NLP leads to the best performance of the multilabel classification of research methods. Additionally, the comparison reveals the differences between deep transfer learning and no transfer learning as well as problem transformation and algorithm adaptation methods. We can compare the models 09 and 10 according to the following criteria:

(1) Test set performance on the evaluation metrics. By investigating the test set performance shown in Table 2, in four of the five examined metrics Model 09 obtained a better score than Model 10**,** and in one metric (hamming loss) both obtained the same score.

(2) Model's complexity and interpretability. If the model's complexity would be considered an additional metric, Model 10 would be preferred, as its structure contains only an embedding layer, a separable Conv1D-decoder followed by a Max Pooling operation, and a dense layer for the final classification, while also applying dropout in multiple places. In contrast, Model 09 is far more complex, as it contains an involved AWD-LSTM language model, which is based on a multilayer bidirectional LSTM with various forms of regularization, and an even more involved classifier utilizing an embedding layer, three LSTM layers, weight dropout, RNN dropout, concatenated average and max pooling layers, and two linear layers with a batch normalization and dropout in between. On top of that, to obtain satisfactory results, Model 09 requires a gradual layer-wise fine-tuning for both, language model and classifier. Therefore, according to Occam's razor principle, the simplicity of Model 10 makes it a better choice in terms of interpretability and maintenance over time.

(3) Performance on the training and validation set. Additionally, Model 10 has performed considerably better than Model 09 on the training and validation sets, which is a good indicator of an adequate model's capacity to extract rich and useful representations.

(4) Extensibility. If additional research methods were added to the algorithm in the future, Model 10 could be quickly retrained after updating a single argument *num_classes*, and applied to new data. By contrast, Model 09 would require training additional binary classifiers from scratch and adding them to the application logic, if applied in a production environment.

(5) Potential inter-label correlations. Since Model 09 trained binary classifiers in isolation, it did not take

dependencies between labels into account. Overall, one of the greatest advantages of deep learning is that it can easily learn dependencies in the data, whereas the problem transformation approach does not take advantage of this.

(6) Generating predictions. An obvious disadvantage of Model 09 compared to Model 10 is the fact that predictions for each label need to be generated separately, as opposed to creating them at once using a single multilabel classifier.

Drawbacks of the one-vs-rest approach. The above comparison revealed several disadvantages of the one-vs-rest problem transformation approach that has been used by Model 09. They can be summarized as follows: (1) increased complexity, which makes it more difficult to interpret and harder to maintain the model, (2) longer training time, as multiple classifiers need to be trained, instead of implementing a single multilabel classification model, (3) inter-label correlations are not taken into account, (4) some metrics, such as exact match score, cannot be computed directly, (5) it is more complicated to make a single prediction of all research methods present in a specific paper at once, and (6), it is difficult to extend the model if new unseen data would contain additional labels, that were not accounted for in the current implementation.

Based on the above-mentioned drawbacks, Model 10 would be recommended for use in a production environment. However, Model 09 constitutes an attractive approach for further research.

Implications of the artifact's results. Overall, the experiments conducted within the scope of this paper deliver promising results for all forms of transfer learning for NLP. However, as Chollet [4:185–186] stated: "What makes a good word-embedding space depends heavily on your task […] because the importance of certain semantic relationships varies from task to task". This might be the reason why the embeddings trained on the target task data ended up in word representations almost as good as those from deep transfer learning. In addition, the obtained results highlight the importance of preprocessing, as considerations like corpus and vocabulary size or inter-label correlations, turned out to have a considerable impact on the obtained results. An overly extensive vocabulary is hard to process and forces the model to learn a lot of noise, while a limited vocabulary may risk failing to recognize some important patterns.

Ultimately, we demonstrated that such tools can already achieve a quality that allow for automating parts of the research process. This saves researchers time and, more importantly from a general perspective, also demonstrates the possibility of establishing semi-automated process for knowledge generation.

We envision multiple use cases that could be facilitated with our tool. First, juxtaposing the publication year and the used research methods to analyze trends and the prevalence of the different methods over time. Second, analyzing the extent to which multi-method approaches are common. Third, comparing the used research methods across the IS journals and conferences and detecting different preferences of methods in different outlets. Fourth, analyzing the used research methods for different topics, like technology acceptance or knowledge management. Fifth, analyzing research methods used by author, institutions, and country. Sixth, in co-citation analyses automatically labelling the nodes with the used research methods and visualizing this, e.g. by color-coding the nodes. Seventh, combining the analyses of multiple dimensions (year, topic, journal, author, institution, country, citation count, research method) in a multi-dimensional data cube, that allows interactive queries and visualizations.

On a larger scale, the artifact represents an improvement of instruments that help scholars to better unlock scientific knowledge over multiple disciplines. The results may contribute to the discussion on meta models and create a common ground for automatically analyzing and summarizing scientific insights. This helps to better promote relevant insights and find open research questions [18].

The consequences of such automation can be far-reaching. As a direct result, summaries can be produced more quickly, resulting in a larger number of papers that can be analyzed. This allows for more profound analyses of paradigms or epistemologies. Since structuring and classifying of research contributions has an influence on their reception and impact (and thus on future research), it will be important to develop "ethical tools", not just "tools that flood us with results": Such ethical tools must provide unbiased, non-discriminatory and comprehensible results. This demands high quality training data. Even though it will probably take several years before such tools will make a significant contribution to publications, IS research in particular should already think about such requirements and their implications.

7. Conclusion

The main contribution of this paper is the development of a deep transfer learning artifact for the multilabel classification of research methods in scientific articles. The presented artifact improves the state of the art in this field across several tested metrics and highlights the best methods to tackle this problem. In particular, this work examined the efficacy of cutting-edge transfer learning techniques, discussed them in detail on a theoretical level, applied them to a multilabel classification of entire text documents, and compared their effectiveness and the best ways of using them. Overall, it has been shown that deep learning models, created by the artifact, led to better recognition of research methods than shallower approaches, such as word2vec and Support Vector Machines, which have been previously applied to this problem in the literature.

All tested deep transfer learning techniques delivered promising results. According to the conducted experiments, ULMFiT has emerged as the best form of pretraining if fine-tuned properly. ELMo has proven to be computationally too expensive to train on this particular dataset for more than a few epochs or to optimize the hyperparameters to the best possible extent. Even though OpenAI Transformers show potential, they are limited to sequences of 512 words. If the problem at hand requires processing of long documents, a sequential network's architecture, such as LSTM, does not work well as an encoder, according to the experiments, which confirmed that it forgets the learned representations from the early stages of the training process. ULMFiT constitutes an exception to this rule thanks to splitting the text into short backpropagation-through-time (bptt) sequences. In contrast to LSTM, more parallelizable architectures such as N-gram-based 1D-CNNs have performed considerably better due to simultaneous extraction of high-level patterns from small parts of the sequence and applying them to other parts of the text.

The investigated deep transfer learning techniques are best applicable to shorter texts, for example abstracts. In comparison, full scientific papers are much longer. Therefore, the created models are prone to forgetting the transferred knowledge. This led to a performance that is only slightly better than training the embeddings from scratch, even with only a limited number of labeled target task data.

In the future work, we also want to examine the effectiveness of Longformer architectures [2] which are transformers that can deal better with long documents.

A simple keyword-based approach for research methods classification has the following shortcomings: research-method keywords appear in the related work and therefore are not always referring to the method used by the paper itself, and papers are not always explicitly mentioning the used research method directly. Our classification-based approach can deal with both cases. Additionally, a keyword-based approach needs a carefully created taxonomy of terms. However, in future work we want to compare our approach with a keyword-based approach that additionally classifies sentences according to whether they are related or original work.

Just as important as the technical improvements is a discussion of the impact of such methods on future research in general. With this work, in which we showed what is possible today and where we have moved the boundaries a bit further towards "automated research", we aim to put the scientific discussion of such methods on more solid grounds. In future work we will use our artifact to analyze and compare the distribution of research methods and philosophical research paradigms among all papers of the last 25 years in (a) the AIS basket of eight, (b) an extended list of additional IS journals and conferences, and (c) non-IS journals in business and social sciences. This kind of large-scale, longitudinal, trans-disciplinary comparison of research methods and philosophical paradigms among thousands of articles is not possible without an automated tool like the one we presented in this paper.

References

[1] Agarwal, S., and H. Yu, "Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion", *Bioinformatics 25*(23), 2009, pp. 3174–3180.

[2] Beltagy, I., M.E. Peters, and A. Cohan, "Longformer: The long-document transformer", *arXiv preprint arXiv:2004.05150*, 2020.

[3] Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information", *arXiv*(1607.04606), 2016.

[4] Chollet, F., *Deep learning with Python*, Manning Publications Co, Shelter Island, New York, 2018.

[5] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv:1810.04805 [cs]*, 2018.

[6] Dwivedi, Y.K., and J. Kuljis, "Profile of IS research published in the European Journal of Information Systems", *European Journal of Information Systems 17*(6), 2008, pp. 678–693.

[7] Dwivedi, Y.K., K. Venkitachalam, A.M. Sharif, W. Al-Karaghouli, and V. Weerakkody, "Research Trends in Knowledge Management: Analyzing the Past and Predicting the Future", *Information Systems Management 28*(1), 2011, pp. 43–56.

[8] Ebeling, B., S. Hoyer, and J. Bührig, "What are your Favorite Methods?-an Examination on the Frequency of Research Methods for is Conferences from 2006 to 2010.", *Proceedings of the 20th European Conference on Information Systems (ECIS)*, (2012).

[9] Eckle-Kohler, J., T.-D. Nghiem, and I. Gurevych, "Automatically Assigning Research Methods to Journal Articles in the Domain of Social Sciences", *Proceedings of the American Society for Information Science and Technology 50*(1), 2013, pp. 1–8.

[10] Friedrich, T., S. Schlauderer, J. Weidinger, and M. Raab, "On the Research Paradigms and Research Methods Employed in the BISE Journal - A Ten-Year Update", *Proceedings of the Wirtschaftsinformatik 2017 Conference*, 2017.

[11] Houngbo, H., and R.E. Mercer, "Method mention extraction from scientific research papers", *Proceedings of COLING 2012*, 2012, pp. 1211–1222.

[12] Hovorka, D.S., K.R. Larsen, J. Birt, and G. Finnie, "A meta-theoretic approach to theory integration in information systems", *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, IEEE (2013), 4656–4665.

[13] Howard, J., and S. Ruder, "Universal Language Model Fine-tuning for Text Classification", *Proceedings of the 56th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), (2018), 328–339.

[14] Howard, J., and R. Thomas, *Practical Deep Learning for Coders, v3 | fast.ai course v3*, 2018.

[15] Kupfer, A., "Research Methods in the Information Systems Discipline: A Literature Analysis of Conference Papers", *Proceedings of the Twenty-fourth Americas Conference on Information Systems (AMCIS)*, 2018.

[16] Kupfer, A., *Private Information Systems and Self-Tracking: Status-Quo and Usage Behavior (PhD Thesis)*, Otto-Friedrich-Universität Bamberg, Bamberg, 2019.

[17] Larsen, K.R., and C.H. Bong, "A tool for addressing construct identity in literature reviews and meta-analyses", *MIS Quarterly 40*(3), 2016, pp. 1–23.

[18] Larsen, K.R., E.B. Hekler, M.J. Paul, and B.S. Gibson, "Improving usability of social and behavioral sciences' evidence: a call to action for a National Infrastructure Project for mining our knowledge", *Communications of the Association for Information Systems 46*(1), 2020, pp. 1.

[19] Li, J., and K. Larsen, "Establishing Nomological Networks for Behavioral Science: a Natural Language Processing Based Approach", *Proceedings of the International Conference on Information Systems (ICIS)*, 2011.

[20] Liakata, M., S. Saha, S. Dobnik, C. Batchelor, and D. Rebholz-Schuhmann, "Automatic recognition of conceptualization zones in scientific articles and two life science applications", *Bioinformatics 28*(7), 2012, pp. 991–1000.

[21] Lin, J., D. Karakos, D. Demner-Fushman, and S. Khudanpur, "Generative content models for structural analysis of medical abstracts", *Proceedings of the hlt-naacl bionlp workshop on linking natural language and biology*, Association for Computational Linguistics (2006), 65–72.

[22] Merity, S., N.S. Keskar, and R. Socher, "An Analysis of Neural Language Modeling at Multiple Scales", *arXiv:1803.08240 [cs]*, 2018.

[23] Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *arXiv preprint arXiv:1301.3781*, 2013.

[24] Mueller, R.M., "A Meta-model for Inferring Inter-theory Relationships of Causal Theories", *Proceedings of the 48th Hawaii International Conference on System Sciences (HICSS)*, (2015), 4908–4917.

[25] Mueller, R.M., "Theory-Data Maps: A Meta-Model and Methods for Inferring and Visualizing Relationships between Causal Theories and Empirical Evidences", *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS)*, (2016), 5288–5297.

[26] Mueller, R.M., and S. Abdullaev, "DeepCause: Hypothesis Extraction from Information Systems Papers with Deep Learning for Theory Ontology Learning", *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS)*, (2019), 6250–6259.

[27] Mueller, R.M., and S. Huettemann, "Extracting Causal Claims from Information Systems Papers with Natural Language Processing for Theory Ontology Learning", *Proceedings of the 51st Hawaii International Conference on System Sciences*, (2018), 5295–5304.

[28] Nasar, Z., S.W. Jaffry, and M.K. Malik, "Information extraction from scientific articles: a survey", *Scientometrics 117*(3), 2018, pp. 1931–1990.

[29] Orlikowski, W.J., and J.J. Baroudi, "Studying Information Technology in Organizations: Research Approaches and Assumptions", *Information Systems Research 2*(1), 1991, pp. 1–28.

[30] Pakrashi, A., D. Greene, and B. MacNamee, "Benchmarking Multi-label Classification Algorithms", *Proceedings of the 24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16)*, (2016).

[31] Palvia, P., D. Leary, E. Mao, V. Midha, P. Pinjani, and A.F. Salam, "Research Methodologies in MIS: An Update", *Communications of the Association for Information Systems 14*(1), 2004.

[32] Pennington, J., R. Socher, and C. Manning, "GloVe: Global vectors for word representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2014), 1532–1543.

[33] Peters, M., M. Neumann, M. Iyyer, et al., "Deep Contextualized Word Representations", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics (2018), 2227–2237.

[34] Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training", *Preprint - OpenAI*, 2018.

[35] Riedl, R., and D. Rueckel, "Historical Development of Research Methods in the Information Systems Discipline.", *Proceedings of the Seventeenth Americas Conference on Information Systems (AMCIS)*, (2011).

[36] Ronzano, F., and H. Saggion, "Dr. inventor framework: Extracting structured information from scientific publications", *International Conference on Discovery Science*, Springer (2015), 209–220.

[37] Ruder, S., "NLP's ImageNet moment has arrived", 2018. http://ruder.io/nlp-imagenet/

[38] Smith, L.N., "A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay", *arXiv:1803.09820 [cs, stat]*, 2018.

[39] Teufel, S., and M. Moens, "Summarizing scientific articles: experiments with relevance and rhetorical status", *Computational linguistics 28*(4), 2002, pp. 409–445.

[40] Vachon, F., M. Pozzebon, and G. Lévesque, "Changes in IS research: a comparative analysis", *International Journal of Business, Humanities and Technology 1*(2), 2011, pp. 184–198.

[41] Vaswani, A., N. Shazeer, N. Parmar, et al., "Attention is All you Need", In I. Guyon, U.V. Luxburg, S. Bengio, et al., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, 5998–6008.

[42] Vessey, I., V. Ramesh, and R.L. Glass, "Research in information systems: An empirical study of diversity in the discipline and its journals", *Journal of Management Information Systems 19*(2), 2002, pp. 129–174.

[43] Wallace, D.P., C. Van Fleet, and L.J. Downs, "The research core of the knowledge management literature", *International Journal of Information Management 31*(1), 2011, pp. 14–20.

[44] Wu, X.-Z., and Z.-H. Zhou, "A unified view of multi-label performance measures", *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, (2017), 3780–3788.