10-9-2023

Automated Knowledge Extraction from IS Research Articles Combining Sentence Classification and Ontological Annotation

Sebastian Huettemann
Berlin School of Economics and Law, Germany, sebastian.huettemann@hwr-berlin.de

Automated Knowledge Extraction from IS Research Articles Combining Sentence Classification and Ontological Annotation

Research Paper

Sebastian Huettemann

Berlin School of Economics and Law, Berlin, Germany
sebastian.huettemann@hwr-berlin.de

Abstract. Manually analyzing large collections of research articles is a time- and resource-intensive activity, making it difficult to stay on top of the latest research findings. Limitations of automated solutions lie in limited domain knowledge and not being able to attribute extracted key terms to a focal article, related work, or background information. We aim to address this challenge by (1) developing a framework for classifying sentences in scientific publications, (2) performing several experiments comparing state-of-the-art sentence transformer algorithms with a novel few-shot learning technique and (3) automatically analyzing a corpus of articles and evaluating automated knowledge extraction capabilities. We tested our approach for combining sentence classification with ontological annotations on a manually created dataset of 1,000 sentences from Information Systems (IS) articles. The results indicate a high degree of accuracy underlining the potential for novel approaches in analyzing scientific publications.

Keywords: Knowledge Extraction, Sentence Classification, Ontological Analysis, Few-Shot Learning, Natural Language Processing

1 Introduction

Studying research contributions has a long tradition in IS. In 2000, Claver et al. (2000) set out to analyze research in the IS discipline by examining articles in MISQ and Information & Management within a 17-year time frame. In 2003, Lee et al. (2003) studied the technology acceptance model's past, present, and future by analyzing roughly a hundred articles from leading IS journals. More recently in 2018, Kupfer (2018) analyzed over 1,000 conference articles to craft an overview of research methods employed in IS.

These exemplary projects have something in common: They indicate that there is an ever-present need for extracting knowledge from the existing body of literature in IS to inform researchers about related work, potential knowledge gaps, and future research directions. These approaches unfortunately share another aspect: they highly depend on manual labor as analyzing hundreds of research articles tends to be an extremely time- and resource-intensive activity.

18th International Conference on Wirtschaftsinformatik,
September 2023, Paderborn, Germany

Research builds on cumulative knowledge and the identification of ongoing research conversations in a specific field. In the field of IS, however, the amount of existing research seems to have already surpassed human limits to comprehension. Larsen et al. (2019) recently demonstrated that conventional approaches to literature reviews tend to miss most of the relevant literature. A trend aggravated by the growing number of publications. As human capabilities to analyze the ever-growing body of research are limited, we must ask the question of how such endeavors can be supported by developing and advancing computational methods for knowledge extraction and synthesis.

Today, researchers are supported by various technologies enabling organizing and querying available knowledge. Scopus (Elsevier, 2015), Web of Science (Thomson-Reuters, 2015), Google Scholar (Google Scholar, 2023), and Semantic Scholar (AI2, 2021) are only a few examples of widely used academic search engines. However, albeit providing various search options, none of these solutions is able to sufficiently extract domain knowledge from articles, such as theories, research methods, topics, and technologies.

Tate et al. (2015) highlighted the potential of performing ontological meta-analyses for reviewing, mapping, and visualizing the current body of knowledge in a domain. They referred to Ramaprasad and Syn (2015) who developed an approach to identify under-researched areas by mapping literature to an ontological framework of the IS discipline. This relates to recent calls for novel approaches to improve the discoverability of knowledge in IS. Larsen et al. (2020) and Wagner et al. (2021) underline the importance of ontological indexing to enable automated knowledge mining from scientific articles.

However, extracting domain-specific terminology can lead to ambiguous results as extracted key terms from an article could either refer to a focal article, related work, or background information. To correctly attribute detected terms, we need to be able to analyze and classify contextual information, i.e., a sentence in which key terms are embedded. Consider the following sentences: "We attempt to build a new adaptive behavior theory." vs. "Smith et al. (2009) attempted to build a new adaptive behavior theory." Extracting the term *adaptive behavior theory* out of both sentences would not be sufficient as it cannot be inferred whether the term belongs to a focal article or related work.

In contrast, being able to correctly attribute key terms from a domain ontology to articles could generate additional metadata that describes the contents of articles in greater detail. Such metadata could enable additional filters in search engines or quantitative analyses over large collections of documents to aggregate mentions of specific theories, methods, topics, or technologies.

Therefore, we ask the following research questions: *RQ 1: How can we classify sentences in research articles to correctly attribute scientific key terms to a focal article? RQ 2: Which state-of-the-art machine learning model performs best in classifying sentences from research articles?*

We contribute to the development of automated methods for knowledge extraction from large collections of research articles by (1) developing a framework for sentence classification that supports the correct attribution of extracted key terminology, (2) comparing state-of-the-art open-source transformer models with a novel few-shot learning technique for classifying sentences, and (3) evaluating the potential of our

approach on a corpus of conference articles and providing a use case example. We share the annotated data for training and testing our models via a GitHub repository (Huettemann, 2023).

2 Theoretical Background

Our approach for inferring the correct attribution of extracted key terms relates to research in three areas: (1) ontological analysis of research articles where our approach can contribute to automatically extracting necessary data, (2) domain ontologies in IS that are necessary to map the text of research articles to domain-specific concepts, and (3) sentence classification to infer the attribution of extracted key terms by classifying the surrounding context.

2.1 Ontological Analysis of Research Articles

Automatically extracting knowledge from research articles in the IS discipline can serve several purposes, e.g., populating theory repositories, identifying knowledge gaps, supporting the conduct of literature reviews, and increasing transparency regarding the evolution of the discipline. Some examples of knowledge extraction include Li et al. (2020) who proposed TheoryOn, a search engine that allows for directly searching constructs and their relationships in scientific texts, Anisienia et al. (2021) who proposed the application of transfer learning with a deep learning model for extracting research methods, and Köhler et al. (2013) who used natural language processing and machine learning techniques to extract research methods based on a taxonomy.

In contrast to methods that focus on extracting individual aspects, incorporating domain ontologies could lead to a more comprehensive extraction of knowledge from research articles. Ramaprasad et al. (2015) proposed a method for semi-automated ontological meta-analysis and synthesis to automatically derive insights from large collections of documents (Tate et al., 2015). Cameron et al. (2017) applied this approach to the domain of mHealth and demonstrated that such analyses lead to detailed quantitative insights. One example of the application IS is the research from La Paz et al. (2020) who created an ontological overview of twenty-five years of research in the Information Systems Journal.

2.2 Domain Ontologies in IS

Scholars in IS created different taxonomies to capture and hierarchically organize key terms that describe specific aspects within the IS discipline. Barki et al. (1988; 1993) proposed a classification schema containing 1,300 key terms in IS, organized for instance into theories, information technologies, IS management, or IS usage. More recent approaches addressed rather specific areas, such as e-commerce (Gregg & Scott, 2008), mobile applications (Nickerson et al., 2013), and digital platforms (Springer & Petrik, 2021).

Mueller et al. (2022) presented a more comprehensive ontology that captures IS key terminology in a hierarchical and interconnected manner. This IS Ontology comprises

2,700+ key terms including theories, topics, methods, technologies, and other relevant categories. These terms are hierarchically organized, implying "is a"-relationships, e.g., *machine learning* is a *data analysis method* is a *methodological entity*. This structure allows for aggregating findings along upper categories so that for instance all key terms classified as *data analysis method* can be collected. It further includes around 380,000 synonyms for its key terms, aiming to enhance detection capabilities. We, therefore, integrated the IS Ontology into our approach to detect key terms in sentences.

2.3 Sentence Classification

Sentence classification in scientific publications can be used to support information retrieval systems (Neves et al., 2019), knowledge graph population (Oelen et al., 2021), or to predict citation intents (Cohan et al., 2019). Brack et al. (2022) used transformer-based language models for sequential sentence classification in abstracts and full papers from biomedicine, computer graphics, chemistry, and computational linguistics. Goncalves et al. (2020) classified sentences in abstracts by using a deep learning approach based on a convolutional layer and a bi-directional gated recurrent unit. Asadi et al. (2019) classified sentences in existing datasets from biochemistry and computer graphics by using logistic model trees, sequential minimal optimization, and a data fusion technique. Jin and Szolovits (2018) performed sequential sentence classification in medical science abstracts with a hierarchical sequential labeling network. We did not find any approaches to sentence classification using few-shot learning as a technique to reduce the effort in manually annotating sentences.

3 Methodology

In this section, we present our approach to infer the correct attribution of extracted key terms from scientific publications. We developed a sentence classification framework to guide the annotation of training data. We selected a set of state-of-the-art transformer models for classifying sentences and created training data as well as a gold standard for testing the models' performance.

3.1 Development of a Sentence Classification Framework

We developed a framework for sentence classification to infer the correct attribution of extracted key terms from scientific publications. We infer this attribution by classifying the context of the terms – the sentence.

We defined the following criteria for creating a corpus of articles that served as the basis for developing the framework and the data for training and testing the machine learning models: (1) the articles in the corpus should reflect current publication standards and practices in IS, (2) the articles should be randomly selected and the final corpus should contain more than 100 articles to assure a sufficient variety in sampled sentences, and (3) the articles should be open-access so that the annotated dataset can be shared with other researchers. We searched for open-access publications in journals of the Senior Scholars' Basket (AIS, 2021), and randomly selected 117 open-access articles from these journals' web pages covering a timeframe from 2012 until 2020 (AIS

corpus). The complete list of articles can be found in our GitHub repository (Huettemann, 2023).

We used Grobid (2022) for text extraction from PDF documents and spaCy (Honnibal & Montani, 2021) for sentence segmentation and key term detection to detect terms that are included in the IS Ontology (Mueller et al., 2022). Only sentences that contained such terms were extracted. Table 1 illustrates the final framework.

Table 1. Framework for sentence classification

Class	**Detailed category**	**Description**	**Examples**
1. Related to article	a. Purpose and section contents	Sentences that describe the focus of an article (e.g. aim, purpose, method); Sentences that describe the content of article sections	"In this study, we performed …", "In the next section, we describe …"
	b. Method Details	Sentences that describe a method (case study, survey, experiment) or a procedure in detail	"We evaluated our system in a series of experiments.", "The data comprised 20 discussion threads."
	c. Research questions and hypotheses	Sentences stating a research question or a hypothesis	"RQ1: How can X influence Y?", "H1: A causes B"
	d. Results	Sentences that describe results or contributions; Sentences that contain performance comparisons	"Our primary contributions are …", "We present empirical evidence for …"
2. Related Work	e. Statement of related work	Sentences must contain a citation and refer to related work	"Smith (2020) developed a method that …", "According to Smith (2010), X is defined as … "
3. Back-ground infor-mation	f. Definitions	Sentences that contain a definition and no cite	"X is defined as …"
	g. Future work	Sentences that imply directions for future work	"Future work could address X"
	h. General information	Sentences that contain more general statements without a citation	"Online communities provide organizations with new opportunities."

To develop the framework, we performed qualitative coding using elements from grounded theory as suggested by Saldaña (2013). In the first round of coding, we sampled 1,525 sentences from the articles in our corpus containing IS-related key terminology and used the sentence categories from the Academic Phrasebank (John Morley, 2018) as a starting point. We performed descriptive coding (Saldaña, 2013) to categorize the sentences into four classes: background, method, related work, and results. Our initial assumption was that all classes, except for related work, would relate to a focal article. However, we found that some sentences described general background knowledge that could also not be attributed to a focal article.

Throughout a second round of coding, we added more detailed categories to better differentiate between sentence classes (see Table 1). In a third round of coding, we performed code mapping (Saldaña, 2013) to refine and group the categories. We found that key terms from the IS Ontology contained in sentences belonging to the detailed categories a to d could directly be attributed to an article. Sentences in subclass e referred to statements associated with a reference. Subclasses f to h indicated general information that did not describe the contents of a focal article. After the third round of coding, further analysis of the literature did not provide additional information.

3.2 Few-Shot Learning and Model Selection

Few-shot learning refers to training machine learning models with only a few training examples. As training state-of-the-art deep learning models typically requires large amounts of data, few-shot learning can be used to fine-tune a model that has already been trained on large datasets to adjust the model to a specific context (Beltagy et al., 2022). In 2022, researchers from Hugging Face, Intel Labs, and the UKP Lab introduced Setfit, a novel framework for few-shot fine-tuning of Sentence Transformers (Tunstall et al., 2022). Compared to other few-shot learning techniques, this framework can be used with less complex models and less training data while still producing highly accurate results.

We defined several criteria for selecting a set of models for our experiments. The models should (1) reflect state-of-art in natural language processing, (2) be publicly available to ensure replicability, (3) have demonstrated very good overall performance, and should preferably have been trained on specific scientific datasets.

We identified sentence-transformers as state-of-the-art in natural language processing (Wolf et al., 2020) and selected three general purpose sentence-transformer models that were extensively tested and evaluated as reported by HuggingFace (Hugging Face, 2023i): *paraphrase-mpnet-base-v2* (Reimers & Gurevych, 2019; Hugging Face, 2023f), *all-MiniLM-L6-v2* (Hugging Face, 2023g), and *distilbert-base-uncased-finetuned-sst-2-english* (Hugging Face, 2023c). These models were the top downloaded models for text classification and sentence similarity with over 2.5m downloads each. We have also included the model *bert-base-uncased* (Hugging Face, 2023b) which was developed by Devlin et al. (2019) and marks a seminal work introducing Bidirectional Encoder Representations from Transformers.

Furthermore, we identified three models that were fine-tuned on scientific datasets: *multicite-multilabel-scibert* was trained on citation context analysis (Lauscher et al., 2022; Hugging Face, 2023e), *longformer-scico* was fine-tuned to detect technical concepts in scientific publications (Cattan & Johnson, 2021; Hugging Face, 2023d), and *specter* implemented a novel method for embedding scientific documents (Cohan et al., 2020; Hugging Face, 2023h).

We did not include GPT models in our approach as the latest version, GPT-4, was released after the submission deadline for the conference. As of to date, GPT-4 is only available via the application ChatGPT and an API. It is not possible to download the model, and the capabilities for fine-tuning are limited. Furthermore, recent research points to GPT models providing different outputs for identical inputs (OpenAI, 2023; Reiss, 2023; Susarla et al., 2023). We therefore conclude that fine-tuning GPT-4 would

require in-depth testing of different prompting strategies which would exceed the scope of this research.

3.3 Development of Training Data

We performed two iterations of training and validating the selected models. In the first iteration, we used the 1,525 sentences that we annotated throughout the development of the sentence classification framework for training (see section 3.1). We applied a 70/20/10-split for training, validation, and testing. For testing the impact of changes in hyperparameter settings, we performed several experiments with *multicite-multilabel-scibert* and *paraphrase-mpnet-base-v2*. Adjusting the batch size and number of epochs as well as increasing the number of iterations higher than 20 did not lead to increases in F1 scores. We noticed improvements in F1 scores the more examples we used for each class in the training set.

In the second iteration, we therefore annotated an additional sample of 1,743 sentences from the AIS corpus (see section 3.1). The author annotated these sentences according to the classification framework and we combined this new dataset with the previously annotated set of 1,525 sentences, resulting in a total of 3,268 sentences.

To further improve the performance, we compared the predictions of *multicite-multilabel-scibert* and *paraphrase-mpnet-base-v2* against manual annotations. When both models predicted a class that deviated from manual annotation, we often found classification errors made by the annotator. After performing 10-fold cross-validation, we reviewed predictions for the entire dataset and noticed 230 instances of potentially misclassified sentences, roughly seven percent of the dataset. For each of these sentences, we decided if re-labeling was justified and adjusted the annotation for 198 sentences. Table 2 shows the class distribution in the final dataset.

Table 2. Class distribution in final dataset

Class	**# Sentences**	**% Sentences**
1. Related to article	1,096	33.5%
2. Related Work	656	46.4%
3. Background information	1,516	20.1%

3.4 Development of a Gold Standard

We created a corpus of scientific articles from the International Conference in Wirtschaftsinformatik (WI corpus). We downloaded conference articles from the AIS eLibrary (2022) and removed articles in German as the current approach only works for articles in the English language. This resulted in 506 articles from the years 2015, 2017, 2019, 2021, and 2022. We extracted all sentences from the PDF documents and performed an automated search for IS-related key terms. Only sentences that contained key terms from the IS Ontology were included in the analysis.

We created a gold standard with a sample of this data. To become familiar with the coding process and to discuss deviating results, the author and a master student in IS each manually annotated a sample of 250 sentences by applying the sentence classification framework (see section 3.1). After this, both annotators created a gold

standard by manually annotating a new sample of 1,000 sentences from the WI-corpus. For each sentence, the annotators assigned one of three possible classes: *belongs_to_article, related_work,* or *background_information*. In addition, the annotators assessed whether the identified key terms in a sentence could be attributed to the focal article by assigning *yes* or *no*.

The sentence "Our study provides a useful framework for interdisciplinary research." contains for instance the key term *framework*. For such sentences, we would use the labels *belongs_to_article* and *yes*. A sentence such as "DevOps is practiced by Facebook [29]" would be classified as *related_work* and *no*.

For the task of sentence classification, both annotators independently assigned the same labels in 866 out of 1,000 cases. This resulted in a Cohen's Kappa of 0.79 and fell into the category of substantial agreement (Landis & Koch, 1977; SciKit Learn, 2022). For the task of assessing whether an identified key term could be attributed to a focal article, both annotators independently assigned the same labels in 889 out of 1,000 cases. This resulted in a Cohen's Kappa of 0.77 and fell into the category of substantial agreement (Landis & Koch, 1977; SciKit Learn, 2022). Throughout discussing the results, the annotators assigned final labels for sentences with deviating classes.

4. Experiments and Results

4.1 Evaluation of Sentence Classification Framework

We assumed that only key terms in sentences classified as *belongs_to_article* describe a focal article. Key terms in other sentence classes either describe related work or provide background information. To test this assumption and thereby evaluate the sentence classification framework, we analyzed the annotations in the gold standard for the task of assessing whether an identified key term could be attributed to a focal article. Table 3 shows the sentence classes and the number of extracted key terms from the IS Ontology regarding whether those describe a focal article. The results confirmed our initial assumption and show that the classification can be used to infer key term attribution in articles.

Table 3. Key term attribution in sentence classes

Sentence-class	Describes focal article	Describes focal article %	Does not describe focal article	Does not describe focal article %	Total
belongs to article	401	99.3%	3	0.7%	410
related work	1	0.4%	232	99.6%	233
background information	0	0.0%	357	100.0%	357

4.2 Comparison of Model Performance

To identify the best model for the task of inferring the attribution of ontological key terms, we trained and tested the models in two scenarios: Regular fine-tuning (Hugging

Face, 2023a) and few-shot learning with Setfit (Tunstall et al., 2022). We used the training dataset of 3,268 sentences (see section 3.3): 70% of the data were used for training, 20% for validation, and 10% for final testing. We performed 10-fold cross-validation so that each model was trained and validated 10 times on different portions of the dataset.

For the reporting of our results, we used standard measures in natural language processing. We calculated macro and weighted average F1 scores where F1 = 2 * (precision * recall) / (precision + recall). Macro average F1 scores provide an unweighted average of F1 scores for individual classes whereas weighted average F1 scores take class imbalance into account (SciKit Learn, 2023b).

We applied the following hyperparameters for few-shot learning with Setfit: batch_size=64, num_iterations=20, num_epochs=1, loss_class=CosineSimilarityLoss. For regular fine-tuning, we applied: per_device_train_batch_size=32, num_train_epochs=20, learning_rate=2e-5, and weight_decay=0.01. The training was performed on Nvidia v100 GPUs. Table 4 shows the results.

Table 4. Average F1 scores of selected models

	Regular Fine-Tuning		**Few-Shot Learning**	
Model	**Macro F1**	**Weighted F1**	**Macro F1**	**Weighted F1**
multicite-multilabel-scibert	0.951	0.948	0.953	0.950
paraphrase-mpnet-base-v2	0.943	0.938	0.950	0.945
longformer-scico	0.950	0.945	0.941	0.936
bert-base-uncased	0.938	0.932	0.941	0.936
distilbert-base-uncased-finetuned-sst-2-english	0.935	0.932	0.939	0.935
all-MiniLM-L6-v2	0.928	0.932	0.930	0.925
allenai-specter	0.945	0.940	0.945	0.940

We have also trained a multinomial Naive Bayes classifier (SciKit Learn, 2023a) to provide a naive benchmark. For this classifier, we removed stopwords from training data, performed word stemming, and transformed all words into lowercase. We applied a bag-of-words approach and performed 10-fold cross-validation. This resulted in average macro and weighted F1 scores of 0.678 and 0.686, respectively.

4.3 Evaluation of Gold Standard

To find out how well the models generalize to a different dataset, we compared the models by predicting sentence classes for the 1,000 sentences in the gold standard (see section 4.4). We show the detailed results for the best-performing model, *multicite-multilabel-scibert*, in Table 5. Figure 1 provides an additional perspective on the classification performance by illustrating the confusion matrix.

Table 5. Results for sentence classification on a sample of WI-conference articles

Classes and averages	Precision	Recall	F1 score	Number of sentences
belongs_to_article	0.95	0.82	0.88	410
related_work	0.92	0.97	0.94	233
background_information	0.85	0.94	0.89	357
macro average	0.9	0.91	0.9	1,000
weighted average	0.9	0.9	0.9	1,000

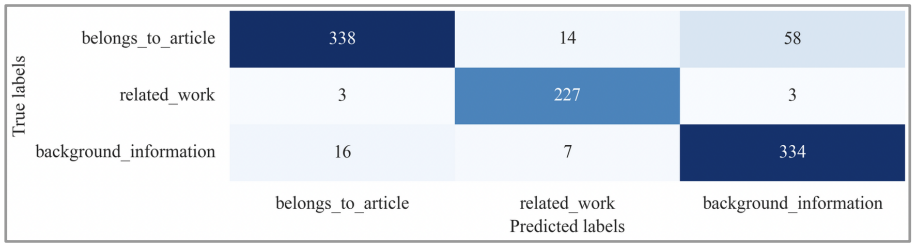

Figure 1. Confusion matrix

We identified two relevant types of errors: (1) the model classifies a sentence that belongs to an article into a different class, and (2) the model classifies a sentence that belongs to a different class as *belongs_to_article*. We regard the first type of error as less critical. As we want to detect key terms that describe an article, this error means that some key terminology is not detected at all. The second type of error is regarded as more severe because key terminology that describes related work or background information is falsely classified as *belongs_to_article*. The confusion matrix shows that the second type of error occurs less frequently than the first one. Only 19 sentences were wrongly classified as *belongs_to_article*.

5 Discussion

5.1 Discussion of Results

Referring to RQ 1, classifying sentences according to the proposed classification framework proved to be a promising approach to attribute keywords to a focal article as shown in section 4.1. In a sample of 1,000 sentences from the WI corpus, we only found four cases where a key term could not be attributed correctly based on the sentence class (see Table 3). In the first case, a specific term was not included in the IS Ontology. Therefore, *Twitter spam* was incorrectly detected as *email spam*. The second case refers to a negated statement where the term *quantitative study* was detected although it was not contained in a focal article: "First, while we derived the conceptual model from theoretical accounts and complementary, exploratory interviews, a rigorous validation (i.e. in terms of a *quantitative study*) is still lacking." In the third case, the term *software support* was detected, although the term *support* was used as a verb: "We reached out to the three online shop vendors to inquire whether their *software supports* personalized price discrimination."

One sentence classified as *related_work* was ambiguous as the first half of the sentence referred to a focal article whereas the second half described related work: "To consider this, these organizational goals were taken into account when developing the *framework* in accordance to Hilty (2008), who further differentiates between rebound effect perspectives of private households, enterprises and states [41]."

For the task of sentence classification (RQ 2), we found only small differences in the F1 scores of the models, ranging between 0.925 and 0.951 (see Table 4). This indicates that training with a limited amount of data – as performed in this research – results in good performance across models based on a transformer architecture. Interestingly, we noticed almost no improvement by applying a few-shot learning framework compared to regular fine-tuning (see section 4.2). We can only assume that the few-shot learning approach performs better than regular fine-tuning when less training data is available.

The *multicite-multilabel-scibert* model performed best on the gold standard. Although the model was not able to detect all sentences belonging to a focal article (recall=0.82), it performed very well in predicting the correct classes for the detected sentences (precision=0.95). These results suggest that important keywords are overlooked in some cases, but are rarely incorrectly attributed to an article.

An analysis of misclassified sentences showed that classification errors mainly occurred in cases where sentences contained ambiguous signals, making it hard to infer if mentioned findings relate to a focal article or related work. In those cases, considering additional context information, such as adjacent sentences or the relative position in the article might improve classification results.

5.2 Limitations

One limitation of our approach is the dependence on a domain ontology. As research progresses, new terms are introduced regularly. If such terms are not included in an ontology, they would be missed. We, therefore, need ways to update a domain ontology with current terminology to make our approach sustainable. Semi-automated approaches could for instance identify author-generated keywords from recent articles where a language model could search for similar key terms in an ontology to recommend placing those into the existing hierarchy.

Another limitation relates to the definition of the class *belongs_to_article* in the classification framework. Currently, one of the main indicators for this class is the usage of personal pronouns, e.g., "We performed a case study". However, for articles written in the passive voice, additional indicators must be present to perform a correct classification, e.g., "In this article, a case study is performed." For sentences written in the passive voice without any indicators, adding additional context information from adjacent sentences might be a way to improve classification performance.

5.3 Implications

With the presented approach, we hope to inform the design of novel tools and technologies that support researchers in more quickly gaining insights from analyzing literature. The approach could for instance be implemented in systems aiming to perform semi-automated systematic literature reviews, especially assessing reviews that synthesize the literature to identify trends, research gaps, and under-researched

areas as suggested by Leidner (2018). As a use case example, we searched for the top 5 research methods in the WI corpus (see section 4.4). Based on ontological key terms in sentences that describe the contents of an article, Table 6 shows the number of articles that used a specific method. In contrast to manual approaches, this analysis was not conducted over days or weeks, but in minutes.

Table 6. Top 5 research methods in WI corpus by article count

Research Method	2015	2017	2019	2021	2022	Total
Literature Study	48	47	53	60	71	279
Design Science	35	38	35	39	48	195
Qualitative Interview	39	35	34	41	38	187
Survey	37	27	43	36	30	173
Conceptual Modeling	33	24	22	13	18	110

Implementing our approach in search engines or domain-specific databases such as the AIS Library (2023) could enable semantic filter options making it possible to identify articles that discuss specific theories, methods, or topics. Furthermore, the approach could support research in novel directions where a collection of articles could be dynamically filtered according to ontological key terms to enable automated summarization of findings through the use of large language models such as ChatGPT. Understanding whether fine-tuning large language models on ontologically indexed data could potentially help in reducing the phenomenon of hallucination (Susarla et al., 2023) might be a future research direction.

6 Conclusion

We see the main contribution of this article in the following points:

1. Developing a framework for classifying sentences in scientific publications
2. Creating an annotated dataset of 3,268 sentences from open-access publications that we made publicly available, enabling researchers to build up on our findings
3. Comparing and evaluating state-of-the-art sentence transformer algorithms with a novel few-shot learning technique
4. Performing and evaluating sentence classification on the WI corpus and providing a use case example for automated knowledge extraction capabilities

We showed that a combination of sentence classification and ontological annotation can support researchers in extracting domain knowledge from large corpora of research articles in a fraction of the time usually required.

Future work could further improve prediction accuracy by including additional context information when training language models, such as adjacent sentences, paragraphs, or positional information of sentences within an article. Semi-automated methods for updating domain ontologies will make our approach more sustainable. Furthermore, conducting user studies where researchers use our approach in a practical setting might lead to novel methods to conduct semi-automated analyses of scientific literature.

References

AI2 (2021), Semantic Scholar | AI-Powered Research Tool, https://www.semanticscholar.org/, Accessed: 15.11.2021.

AIS (2021), Senior Scholars' Basket of Journals, https://aisnet.org/page/SeniorScholarBasket, Accessed: 15.11.2021.

AIS eLibrary (2022), AIS Electronic Library - Proceedings of Wirtschaftsinformatik-Conference, https://aisel.aisnet.org/wi/, Accessed: 28.2.2023.

AIS Library (2023), Association for Information Systems (AIS) eLibrary, https://aisel.aisnet.org/, Accessed: 31.5.2023.

Anisienia, A., Mueller, R.M., Kupfer, A. & Staake, T. (2021), Research Method Classification with Deep Transfer Learning for Semi-Automatic Meta-Analysis of Information Systems Papers, *in* 'Hawaii International Conference on System Sciences (HICSS)'.

Asadi, N., Badie, K. & Mahmoudi, M.T. (2019), 'Automatic Zone Identification in Scientific Papers via Fusion Techniques', *Scientometrics* **119**(2), pp. 845–862.

Barki, H., Rivard, S. & Talbot, J. (1988), 'An Information Systems Keyword Classification Scheme', *MIS Quarterly* **12**(2), p. 299.

Barki, H., Rivard, S. & Talbot, J. (1993), 'A Keyword Classification Scheme for IS Research Literature: An Update', *MIS Quarterly* **17**(2), p. 209.

Beltagy, I., Cohan, A., Logan Iv, R., Min, S. & Singh, S. (2022), Zero- and Few-Shot NLP with Pretrained Language Models, *in* 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts' Association for Computational Linguistics, pp. 32–37.

Brack, A., Hoppe, A., Buschermöhle, P. & Ewerth, R. (2022), Cross-Domain Multi-Task Learning for Sequential Sentence Classification in Research Papers, *in* 'Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL)' Association for Computing Machinery (ACM).

Cameron, J.D., Ramaprasad, A. & Syn, T. (2017), 'An ontology of and roadmap for mHealth research', *International Journal of Medical Informatics* **100**, pp. 16–25.

Cattan, A. & Johnson, S. (2021), SCICO: Hierarchical Cross-Document Coreference for Scientific Concepts, *in* 'Automated Knowledge Base Construction (AKBC)'.

Claver, E., González, R. & Llopis, J. (2000), 'An analysis of research in information systems (1981–1997)', *Information & Management* **37**(4), pp. 181–195.

Cohan, A., Ammar, W., van Zuylen, M. & Cady, F. (2019), Structural Scaffolds for Citation Intent Classification in Scientific Publications, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)'.

Cohan, A., Feldman, S., Beltagy, I., Downey, D. & Weld, D. (2020), SPECTER: Document-level Representation Learning using Citation-informed Transformers, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics'.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *in* 'Proceedings of NAACL-HLT'

Eckle-Kohler, J., Nghiem, T.-D. & Gurevych, I. (2013), Automatically assigning research methods to journal articles in the domain of social sciences: Automatically Assigning Research Methods to Journal Articles in the Domain of Social Sciences, *in* 'Proceedings of the American Society for Information Science and Technology', pp. 1–8.

Elsevier (2015), Scopus - the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings, http://www.elsevier.com/online-tools/scopus, Accessed: 29.4.2015.

Gonçalves, S., Cortez, P. & Moro, S. (2020), 'A Deep Learning Classifier for Sentence Classification in Biomedical and Computer Science Abstracts', *Neural Computing and Applications* **32**(11), pp. 6793–6807.

Google Scholar (2023), About Google Scholar, https://scholar.google.de/intl/de/scholar/about.html, Accessed: 6.3.2023.
Gregg, D.G. & Scott, J.E. (2008), 'A typology of complaints about eBay sellers', *Communications of the ACM* **51**(4), pp. 69–74.
GROBID (2022), GROBID - A machine learning software for extracting information from scholarly documents, https://github.com/kermitt2/grobid, Accessed: 30.6.2022.
Honnibal, M. & Montani, I. (2021), spaCy - Industrial-strength Natural Language Processing in Python, https://spacy.io/, Accessed: 17.11.2021.
Huettemann, S. (2023), GitHub Repository: Sentence Classification Data, https://github.com/sebastianhuettemann/sentence_classification, Accessed: 13.6.2023.
Hugging Face (2023a), Fine-tune a pretrained model, https://huggingface.co/docs/transformers/training, Accessed: 19.6.2023.
Hugging Face (2023b), Model Overview: bert-base-uncased, https://huggingface.co/bert-base-uncased, Accessed: 15.6.2023.
Hugging Face (2023c), Model Overview: distilbert-base-uncased-finetuned-sst-2-english, https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english, Accessed: 22.2.2023.
Hugging Face (2023d), Model Overview: longformer-scico, https://huggingface.co/allenai/longformer-scico, Accessed: 15.6.2023.
Hugging Face (2023e), Model Overview: multicite-multilabel-scibert, https://huggingface.co/allenai/multicite-multilabel-scibert, Accessed: 15.6.2023.
Hugging Face (2023f), Model Overview: paraphrase-multilingual-mpnet-base-v2, https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2, Accessed: 15.6.2023.
Hugging Face (2023g), Model Overview: sentence-transformers/all-MiniLM-L6-v2, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, Accessed: 22.2.2023.
Hugging Face (2023h), Model Overview: specter, https://huggingface.co/allenai/specter, Accessed: 15.6.2023.
Hugging Face (2023i), Pretrained Models: Sentence-Transformers documentation, https://www.sbert.net/docs/pretrained_models.html, Accessed: 15.6.2023.
Jin, D. & Szolovits, P. (2018), Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts, *in* 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing' Association for Computational Linguistics, pp. 3100–3109.
John Morley (2018), *Academic Phrasebank 2018 Enhanced Edition*. The University of Manchester.
Kupfer, A. (2018), Research Methods in the Information Systems Discipline: A Literature Analysis of Conference Papers, *in* 'Americas Conference on Information Systems (AMCIS) 2018 Proceedings'.
La Paz, A., Merigó, J.M., Powell, P., Ramaprasad, A. & Syn, T. (2020), 'Twenty-five years of the Information Systems Journal: A bibliometric and ontological overview', *Information Systems Journal* **30**(3), pp. 431–457.
Landis, J.R. & Koch, G.G. (1977), 'The Measurement of Observer Agreement for Categorical Data', *Biometrics* **33**(1), p. 159.
Larsen, K.R., Hekler, E.B., Paul, M.J. & Gibson, B.S. (2020), '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*, pp. 1–17.
Larsen, K.R., Hovorka, D.S., Dennis, A.R. & West, J.D. (2019), 'Understanding the Elephant: The Discourse Approach to Boundary Identification and Corpus Construction for Theory Review Articles', *Journal of the Association for Information Systems*, pp. 887–927.

Lauscher, A., Ko, B., Kuehl, B., Johnson, S., Cohan, A., Jurgens, D. & Lo, K. (2022), MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting, *in* 'Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', pp. 1875–1889.

Lee, Y., Kozar, K.A. & Larsen, K.R.T. (2003), 'The Technology Acceptance Model: Past, Present, and Future', *Communications of the Association for Information Systems* **12**

Leidner, D. (2018), 'Review and Theory Symbiosis: An Introspective Retrospective', *Journal of the Association for Information Systems* **19**(06), pp. 552–567.

Li, J., Larsen, K. & Abbasi, A. (2020), 'TheoryOn: A Design Framework and System for Unlocking Behavioral Knowledge Through Ontology Learning', *MIS Quarterly* **44**(4), pp. 1733–1772.

Mueller, R.M., Huettemann, S., Larsen, K.R., Yan, S. & Handler, A. (2022), Toward an Information Systems Ontology, *in* 'Proceedings of the 17th International Conference on Design Science Research in Information Systems and Technology (DESRIST)'.

Neves, M., Butzke, D. & Grune, B. (2019), Evaluation of Scientific Elements for Text Similarity in Biomedical Publications, *in* 'Proceedings of the 6th Workshop on Argument Mining' Association for Computational Linguistics, pp. 124–135.

Nickerson, R.C., Varshney, U. & Muntermann, J. (2013), 'A method for taxonomy development and its application in information systems', *European Journal of Information Systems* **22**(3), pp. 336–359.

Oelen, A., Stocker, M. & Auer, S. (2021), Crowdsourcing Scholarly Discourse Annotations, *in* '26th International Conference on Intelligent User Interfaces' ACM, pp. 464–474.

OpenAI (2023), 'GPT-4 Technical Report', *arXiv*.

Ramaprasad, A. & Syn, T. (2015), 'Ontological Meta-Analysis and Synthesis', *Communications of the Association for Information Systems* **37**, https://aisel.aisnet.org/cais/vol37/iss1/7/, Accessed: 29.9.2021.

Reimers, N. & Gurevych, I. (2019), Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', pp. 3980–3990.

Reiss, M.V. (2023), 'Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark', *arXiv* , http://arxiv.org/abs/2304.11085, Accessed: 15.6.2023.

Saldaña, J. (2013), *The coding manual for qualitative researchers*. SAGE Publications.

SciKit Learn (2022), Cohen's kappa: a statistic that measures inter-annotator agreement, https://scikit-learn/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html, Accessed: 10.3.2023.

SciKit Learn (2023a), Naive Bayes classifier for multinomial models, https://scikit-learn/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html, Accessed: 19.6.2023.

SciKit Learn (2023b), SciKit Learn Metrics - sklearn.metrics.f1_score, https://scikit-learn/stable/modules/generated/sklearn.metrics.f1_score.html, Accessed: 5.3.2023.

Springer, V. & Petrik, D. (2021), Towards a Taxonomy of Impact Factors for Digital Platform Pricing, *in* Lecture Notes in Business Information Processing, Springer, pp. 115–124.

Susarla, A., Gopal, R., Thatcher, J.B. & Sarker, S. (2023), 'The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems', *Information Systems Research.*

Tate, M., Furtmueller, E., Evermann, J. & Bandara, W. (2015), 'Introduction to the Special Issue: The Literature Review in Information Systems', *Communications of the Association for Information Systems* **37**.

Thomson-Reuters (2015), Web of Science - the world's leading source of scholarly research data, http://wokinfo.com, Accessed: 29.4.2015.
Tunstall, L., Reimers, N., Jo, U.E.S., Bates, L., Korat, D., Wasserblat, M. & Pereg, O. (2022), Efficient Few-Shot Learning Without Prompts, *in* '36th Conference on Neural Information Processing Systems (NeurIPS)'.
Wagner, G., Lukyanenko, R. & Paré, G. (2021), 'Artificial intelligence and the conduct of literature reviews', *Journal of Information Technology* **37**(2), pp. 209–226.
Wolf, T. et al. (2020), Transformers: State-of-the-Art Natural Language Processing, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations' Association for Computational Linguistics, pp. 38–45.