Cross-Domain Multi-Task Learning for Sequential Sentence Classification in Research Papers

Arthur Brack
TIB – Leibniz Information Centre for Science and Technology & Leibniz University
Hannover, Germany
Arthur.Brack@tib.eu

Anett Hoppe
TIB – Leibniz Information Centre for Science and Technology & L3S Research Center, Leibniz University
Hannover, Germany
Anett.Hoppe@tib.eu

Pascal Buschermöhle
Leibniz University Hannover
Germany

Ralph Ewerth
TIB – Leibniz Information Centre for Science and Technology & L3S Research Center, Leibniz University
Hannover, Germany
Ralph.Ewerth@tib.eu

ABSTRACT

Sequential sentence classification deals with the categorisation of sentences based on their content and context. Applied to scientific texts, it enables the automatic structuring of research papers and the improvement of academic search engines. However, previous work has not investigated the potential of transfer learning for sentence classification across different scientific domains and the issue of different text structure of full papers and abstracts. In this paper, we derive seven related research questions and present several contributions to address them: First, we suggest a novel uniform deep learning architecture and multi-task learning for cross-domain sequential sentence classification in scientific texts. Second, we tailor two common transfer learning methods, sequential transfer learning and multi-task learning, to deal with the challenges of the given task. Semantic relatedness of tasks is a prerequisite for successful transfer learning of neural models. Consequently, our third contribution is an approach to semi-automatically identify semantically related classes from different annotation schemes and we present an analysis of four annotation schemes. Comprehensive experimental results indicate that models, which are trained on datasets from different scientific domains, benefit from one another when using the proposed multi-task learning architecture. We also report comparisons with several state-of-the-art approaches. Our approach outperforms the state of the art on full paper datasets significantly while being on par for datasets consisting of abstracts.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Multi-task learning**; *Neural networks*; • **Information systems** → Digital libraries and archives.

JCDL '22, June 20–24, 2022, Cologne, Germany

ACM ISBN 978-1-4503-9345-4/22/06.
https://doi.org/10.1145/3529372.3530922

KEYWORDS

sequential sentence classification, zone identification, transfer learning, multi-task learning, scholarly communication

ACM Reference Format:
Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2022. Cross-Domain Multi-Task Learning for Sequential Sentence Classification in Research Papers. In *The ACM/IEEE Joint Conference on Digital Libraries in 2022 (JCDL '22), June 20–24, 2022, Cologne, Germany*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3529372.3530922

1 INTRODUCTION

To search relevant research papers for a particular field is a core activity of researchers. Scientists usually use academic search engines and skim through the text of the found articles to assess their relevance. However, academic search engines cannot assist researchers adequately in these tasks since most research papers are plain PDF files and not machine-interpretable [9, 60, 74]. The exploding number of published articles aggravates this situation further [7]. Therefore, automatic approaches to structure research papers are highly desired.

Sequential sentence classification targets the categorisation of sentences by their semantic content or function. In research papers, this can be used to classify sentences by their contribution to the article's content, e.g. to determine if a certain sentence contains information about the research work's objective, methods or results [18]. Figure 1 shows an example of an abstract with classified sentences. Such a semantification of sentences can help to focus on relevant elements of text and thus assist information retrieval systems [50, 60] or knowledge graph population [51]. The task is called *sequential* to distinguish it from the general *sentence classification* task where a sentence is classified in isolation, i.e. without using local context. However, in research papers the meaning of a sentence is often informed by the context from neighbouring sentences, e.g. sentences describing the methods usually precede sentences about results.

Several approaches have been proposed for *sequential sentence classification* (e.g. [2, 34, 65]), and several datasets were annotated for various scientific domains (e.g. [18, 23, 28, 67]). The datasets contain either abstracts or full papers and were annotated with domain-specific sentence classes. However, research infrastructures usually

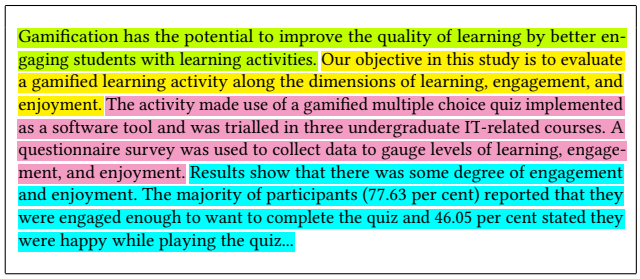

Figure 1: An annotated abstract taken from the CSABSTRUCT dataset [15], in which sentences describing the *background* (green), *objectives* (yellow), *methods* (magenta), and *results* (cyan) of the paper are coloured.

support multiple scientific domains. Therefore, stakeholders of digital libraries are interested in a uniform solution that enables the combination of these datasets to improve the overall accuracy. For this purpose, this paper explores the following research questions.

First, although some approaches propose transfer learning for the scientific domain [5, 10, 29, 53], the field lacks a comprehensive empirical study on transfer learning across different scientific domains for *sequential sentence classification*. Transfer learning enables the combination of knowledge from multiple datasets to improve classification performance and thus to reduce annotation costs. The annotation of scientific text is particularly costly since it demands expertise in the article's domain [3, 8, 26]. However, studies revealed that the success of transferring neural models depends largely on the relatedness of the tasks, and transfer learning with unrelated tasks may even degrade the performance [48, 52, 59, 64]. Two tasks are related if there exists some implicit or explicit relationship between the feature spaces [52]. On the other hand, every scientific domain is characterised by its specific terminology and phrasing, which yields different feature spaces. Thus, it is not clear to which extent datasets from different scientific disciplines are related. This raises the following research questions (RQ) for the task of sequential sentence classification:

- RQ1: To which extent are datasets from different scientific domains semantically related?
- RQ2: Which transfer learning approach works best?
- RQ3: Which neural network layers are transferable under which constraints?
- RQ4: Is it beneficial to train a multi-task model with multiple datasets?

Normally, every dataset has a domain-specific annotation scheme that consists of a set of associated sentence classes. This raises the second set of research questions with regard to the consolidation of these annotation schemes. Prior work [45] annotated a dataset multiple times with different schemes, and analysed the multivariate frequency distributions of the classes. They found that the investigated schemes are complementary and should be combined. However, annotating datasets multiple times is costly. To support the consolidation of different annotation schemes across domains, we examine the following RQs:

- RQ5: Can a model trained with multiple datasets recognise the semantic relatedness of classes from different annotation schemes?
- RQ6: Can we derive a consolidated, domain-independent annotation scheme and use that scheme to compile a new dataset to train a domain-independent model?

Finally, current approaches for sequential sentence classification are designed either for abstracts or full papers. One reason is that these text types follow rather different structures: In abstracts, different sentence classes directly follow one another normally. The general paper text, however, exhibits longer passages without change of the semantic sentence class. Typically, deep learning is used for abstracts [15, 19, 28, 34, 65, 75] since presumably more training data are available, whereas for full papers, also called *zone identification*, hand-crafted features and linear models have been suggested [2, 4, 23, 44]. However, deep learning approaches have also been applied successfully to full papers in related tasks such as argumentation mining [41], document summarisation [1, 16, 21, 27], or n-ary relation extraction [25, 33, 36]. Thus, the potential of deep learning has not been fully exploited yet for sequential sentence classification on full papers, and no unified solution for abstracts as well as full papers exists. This raises the RQ:

- RQ7: Can a unified deep learning approach be applied to text types with very different structures like abstracts or full papers?

In this paper, we investigate these research questions and present the following contributions: (1) We introduce a novel multi-task learning framework for sequential sentence classification. (2) Furthermore, we propose and evaluate an approach to semi-automatically identify semantically related classes from different annotation schemes and present an analysis of four annotation schemes. Based on the analysis, we suggest a domain-independent annotation scheme and compile a new dataset that enables to classify sentences in a domain-independent manner. (3) Our proposed unified deep learning approach can handle both text types, abstracts and full papers, despite their structural differences. (4) To facilitate further research, we make our source code publicly available: https://github.com/arthurbra/sequential-sentence-classification.

Comprehensive experimental results demonstrate that our multi-task learning approach successfully makes use of datasets from different scientific domains, with different annotation schemes, that contain abstracts or full papers. In particular, we outperform state-of-the-art approaches for full paper datasets significantly, while obtaining competitive results for datasets of abstracts.

The remainder of the paper is organised as follows: Section 2 summarises related work on sentence classification in research papers and transfer learning in NLP. Our proposed approaches are presented in Section 3. The setup and results of our experimental evaluation are reported in Section 4 and 5, while Section 6 concludes the paper and outlines areas of future work.

2 RELATED WORK

This section outlines datasets for sentence classification in scientific texts and describes machine learning methods for this task. Furthermore, we briefly review transfer learning methods. For a

more comprehensive overview about information extraction from scientific text, we refer to Brack et al. [9] and Nasar et al. [49].

2.1 Sequential Sentence Classification in Scientific Text

Datasets: As depicted in Table 1, annotated benchmark datasets for sentence classification in research papers come from various domains, e.g. PubMed-20k [18] consists of biomedical randomised controlled trials, NICTA-PIBOSO [37] comes from evidence-based medicine, Dr. Inventor dataset [23] from computer graphics, and the ART/Core Scientific Concepts (CoreSC) dataset [45] from chemistry and biochemistry. Most datasets cover only abstracts, while ART/CoreSC and Dr. Inventor cover full papers.

Approaches for Abstracts: Deep learning has been the preferred approach for sentence classification in abstracts in recent years [15, 19, 28, 34, 65, 75]. These approaches follow a common *hierarchical sequence labelling architecture*: (1) a word embedding layer encodes tokens of a sentence to word embeddings, (2) a sentence encoder transforms the word embeddings of a sentence to a sentence representation, (3) a context enrichment layer enriches all sentence representations of the abstract with context from surrounding sentences, and (4) an output layer predicts the label sequence.

As depicted in Table 2, the approaches vary in different implementations of the layers. The approaches use different kinds of word embeddings, e.g. Global Vectors (GloVe) [55], Word2Vec [47], or SciBERT [6] that is BERT [20] pre-trained on scientific text. For sentence encoding, a bidirectional long short-term memory (Bi-LSTM) [31] or a convolutional neural network (CNN) with various pooling strategies are utilised, while Yamada et al. [75] and Shang et al. [65] use the classification token ([CLS]) of BERT or SciBERT. To enrich sentences with further context, a recurrent neural network such as a Bi-LSTM or bidirectional gated recurrent unit (Bi-GRU) [13] is used. Shang et al. [65] additionally exploit an attention-mechanism across sentences; however, it introduces quadratic runtime complexity that depends on the number of sentences. A conditional random field (CRF) [40] is mostly used as an output layer to capture the interdependence between classes. Yamada et al. [75] form spans of sentence representations and Semi-Markov CRFs to predict the label sequence by considering all possible span sequences of various lengths. Thus, their approach can better label longer continuous sentences but is computationally more expensive than a CRF. Cohan et al. [15] obtain contextual sentence representations directly by fine-tuning SciBERT and utilising the separation token ([SEP]) of SciBERT. However, their approach can process only about 10 sentences at once since BERT supports sequences of up to 512 tokens only.

Approaches for Full Papers: For full papers, logistic regression, support vector machines and CRFs with hand-crafted features have been proposed [2, 4, 23, 44, 69, 70]. They represent a sentence with various syntactic and linguistic features such as n-grams, part-of-speech tags, or citation markers, which were engineered for the respective datasets. Asadi et al. [2] also exploit semantic features obtained from knowledge bases such as Wordnet [22]. To incorporate contextual information, each sentence representation also contains the label of the previous sentence ("history feature") and the sentence position in the document ("location feature"). To better consider the interdependence between labels, some approaches apply CRFs, while Asadi et al. [2] suggest fusion techniques within a dynamic window of sentences. However, some approaches [2, 4, 23] exploit the *ground-truth label* instead of the predicted label of the preceding sentence ("history feature") during prediction (as confirmed by the authors), which has a significant impact on the performance.

Related tasks also classify sentences in full papers with deep learning methods, e.g. for citation intent classification [14, 39], or algorithmic metadata extraction [61] but without exploiting context from surrounding sentences. Comparable to us, Lauscher et al. [41] utilise a hierarchical deep learning architecture for argumentation mining in full papers but evaluate it only on one corpus.

To the best of our knowledge, a unified approach for sequential sentence classification for abstracts as well as full papers has not been proposed and evaluated yet.

2.2 Transfer Learning

Transfer learning enables a target task to exploit knowledge from another source task to achieve a better prediction accuracy. The tasks can have training data from different domains and vary in their objectives. According to Ruder's taxonomy for transfer learning [59], we investigate inductive transfer learning in this study since the target training datasets are labelled. Inductive transfer learning can be further subdivided into multi-task learning, where tasks are learned simultaneously, and sequential transfer learning (also referred to as parameter initialisation), where tasks are learned sequentially. Since there are many applications for transfer learning, we focus on the most relevant cases to our work here. For a more comprehensive overview, we refer to [52, 59, 73].

Fine-tuning a pre-trained language model is a popular approach for sequential transfer learning in NLP [11, 20, 30, 32]. Here, the source task involves learning a language model (or a variant of it) using a large unlabelled text corpus. Then, the model parameters are fine-tuned with labelled data of the target task. Pruksachatkun et al. [56] improve these language models by *intermediate task transfer learning* where a language model is fine-tuned on a data-rich intermediate task before fine-tuning on the final target task. Park and Caragea [53] provide an empirical study on intermediate transfer learning from the non-academic domain to scientific keyphrase identification. They show that SciBERT in combination with related tasks such as sequence tagging improves performance, while BERT or unrelated tasks degrade the performance.

For *sequence tagging* Yang et al. [76] investigate multi-task learning in the general non-academic domain with a small and a big dataset. Schulz et al. [63] evaluate multi-task learning for argumentation mining with multiple datasets in the general domain. Lee et al. [43] successfully transfer pre-trained parameters from a big dataset to a small dataset in the biological domain. For *coreference resolution*, Brack et al. [10] apply sequential transfer learning and utilise a large dataset from the general domain to improve models for a small dataset in the scientific domain.

For *sentence classification*, Mou et al. [48] compare (1) transferring parameters from a source dataset to a target dataset against (2) training one model with two datasets in the non-academic domain.

Table 1: Characteristics of benchmark datasets for sentence classification in research papers.

Dataset	Domains	# Papers	Text Type	Sentence Classes
PubMed-20k [18]	Biomedicine	20,000	abstracts	Background, Objective, Methods, Results, Conclusion
NICTA-PIBOSO [37]	Biomedicine	1,000	abstracts	Background, Intervention, Study, Population, Outcome, Other
CSABSTRUCT [15]	Computer Science	2,189	abstracts	Background, Objective, Method, Result, Other
CS-Abstracts [28]	Computer Science	654	abstracts	Background, Objective, Methods, Results, Conclusions
Emerald 100k [67]	Management, Engineering, Information Science	103,457	abstracts	Purpose, Design/methodology/approach, Findings, Originality/value, Social implications,Practical implications, Research limitations/implications
MAZEA [17]	Physics, Engineering Life and Health Sciences	1,335	abstracts	Background, Gap, Purpose, Method, Result, Conclusion
Dr. Inventor [23]	Computer Graphics	40	full paper	Background, Challenge, Approach, Outcome, Future Work
ART/CoreSC [45]	Chemistry Computational Linguistic	225	full paper	Background, Motivation, Goal, Hypothesis, Object, Model, Method, Experiment, Result, Observation, Conclusion

Table 2: Comparison of deep learning approaches for sequential sentence classification in abstracts.

Approach	Word embedd.	Sentence encoding	Context enrichm.	Output layer
Dernoncourt and Lee (2016) [19]	Char. Emb. + GloVe	Bi-LSTM/ concat.	-	CRF
Jin and Szolovits (2018) [34]	Bio word2vec	Bi-LSTM/ att. pooling	Bi-LSTM	CRF
Cohan et al. (2019) [15]	SciBERT	SciBERT-[*SEP*]	SciBERT-[*SEP*]	softmax
Gonçalves et al. (2020) [28]	GloVe	CNN / max pooling	Bi-GRU	softmax
Yamada et al. (2020) [75]	BERT from PubMed	BERT-[*CLS*]	Bi-LSTM	Semi-Markov CRF
Shang et al. (2021) [65]	SciBERT	SciBERT-[*CLS*]	Bi-LSTM/ attention	CRF

They demonstrate that semantically related tasks improve while unrelated tasks degrade the performance of the target tasks. Su et al. [68] study multi-task learning for sentiment classification in product reviews from multiple domains. Lauscher et al. [42] evaluate multi-task learning on scientific texts, however, only on one dataset with different annotation layers. Banerjee et al. [5] apply sequential transfer learning from the medical to the computer science domain for discourse classification, however, only for two domains and on abstracts, whereas Spangher et al. [66] explore this task on news articles with multi-task learning using multiple datasets. Gupta et al. [29] utilise a multi-task learning with two scaffold tasks to detect contribution sentences in full papers, however, only in one domain and with limited sentence context.

Several approaches also exist to *train multiple tasks jointly*: Luan et al. [46] train a model on three tasks (coreference resolution, entity and relation extraction) using one dataset of research papers. Wei et al. [72] utilise a multi-task model for entity recognition and relation extraction on one dataset in the non-academic domain. Changpinyo et al. [12] analyse multi-task training with multiple datasets for sequence tagging. *In contrast, we investigate sequential sentence classification across multiple science domains.*

3 SEQUENTIAL SENTENCE CLASSIFICATION

On the one hand, the discussion of related work shows that several approaches and datasets from various scientific domains have been introduced for sequential sentence classification. On the other hand, although transfer learning has been applied to various NLP tasks, it is known that the success depends largely on the relatedness of the tasks [48, 52, 59]. However, the field lacks an empirical study on transfer learning between different scientific domains for sequential sentence classification that cover either only abstracts or entire papers. Furthermore, previous approaches investigated transfer learning for one or two datasets only. To the best of our knowledge, a unified approach for different types of texts that differ noticeably by their structure and semantic context of sentences, as it is the case for abstracts and full papers, has not been proposed yet.

In this section, we suggest a unified cross-domain multi-task learning approach for sequential sentence classification. Our tailored transfer learning approaches, depicted in Figure 2, exploit multiple datasets comprising different text types in form of abstracts and full papers. The unified approach without transfer learning is described in Section 3.1 while Section 3.2 introduces our tailored transfer learning approaches. Finally, in Section 3.3, we present an approach to semi-automatically identify the semantic relatedness of sentence classes between different annotation schemes.

3.1 Unified Deep Learning Approach

Given a paper with the sentences $(s_1, ..., s_n)$ and the set of dataset specific classes L (e.g. *Background*, *Methods*), the task of *sequential sentence classification* is to predict the corresponding label sequence $(y_1, ..., y_n)$ with $y_i \in L$. For this task, we propose a unified deep learning approach as depicted in Figure 2(a), which is applicable to both abstracts *and* full papers. The core idea is to enrich sentence representations with context from surrounding sentences..

Our approach (denoted as *SciBERT-HSLN*) is based on the *Hierarchical Sequential Labeling Network (HSLN)* [34]. In contrast to Jin and Szolovits [34], we utilise SciBERT [6] as word embeddings and evaluate the approach on abstracts *as well as* full papers. We have chosen HSLN as the basis since it is better suited for full papers: It has no limitations on text length (in contrast to the approach of Cohan et al. [15]), and is computationally less expensive than the more recent approaches [65, 75]. Furthermore, their implementation is publicly available. The goal of this paper is not to beat state-of-the-art results but rather to provide an empirical study on transfer learning and offer a uniform solution. Our *SciBERT-HSLN* architecture has the following layers:

(a) *Word embedding:* input is a sequence of tokens $(t_{i,1}, ..., t_{i,m})$ of sentence s_i, and output a sequence of word embeddings $(w_{i,1}, ..., w_{i,m})$.

(b) *Sentence encoding:* input $(w_{i,1}, ..., w_{i,m})$ is transformed via a Bi-LSTM [31] into the representations $(h_{i,1}, ..., h_{i,m})$ $(h_{i,t} \in \mathbb{R}^{d^h})$ which are enriched with contextual information within the sentence. Then, attention pooling [34, 77] with r heads produces a sentence vector $e_i \in \mathbb{R}^{rd^u}$. An attention head produces a weighted average

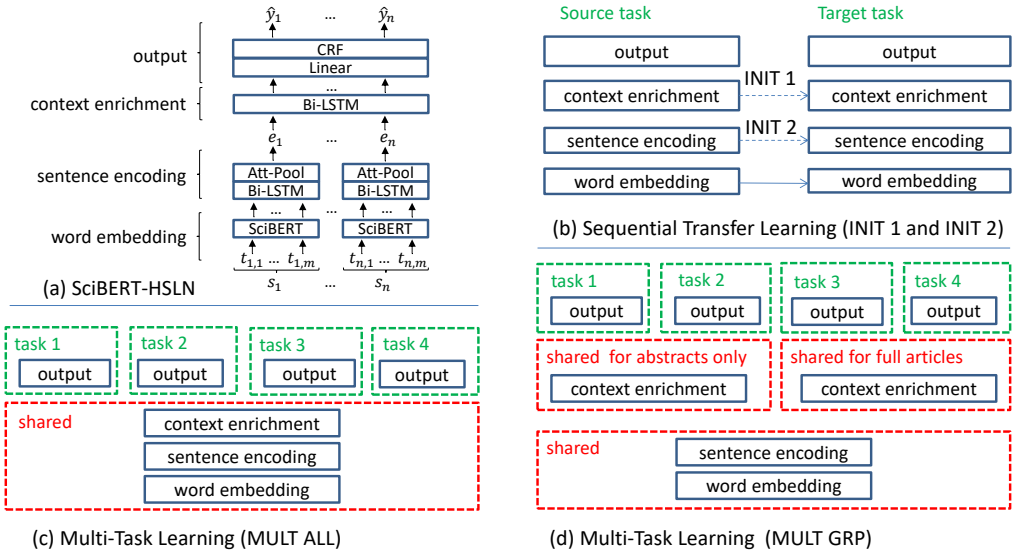

Figure 2: Proposed approaches for sequential sentence classification: (a) unified deep learning architecture *SciBERT-HSLN* for datasets of abstracts and full papers; (b) sequential transfer learning approaches, i.e. INIT 1 transfers all possible layers, INIT 2 only the sentence encoding layer; (c) and (d) are the multi-task learning approaches, i.e. in MULT ALL all possible layers are shared between the tasks, in MULT GRP the context enrichment is shared between tasks with the same text type.

over the token representations of a sentence. Multiple heads enable the model to capture several semantics of a sentence.

(c) *Context enrichment:* sentence vectors $(e_1, ..., e_n)$ are transformed via a Bi-LSTM into $(c_1, ..., c_n)$ with $c_i \in \mathbb{R}^{d^h}$. Thus, each sentence vector c_i is enriched with contextual information from surrounding sentences.

(d) *Output layer:* transforms $(c_1, ..., c_n)$ via a linear transformation to the logits $(l_1, ..., l_n)$ with $l_i \in \mathbb{R}^{|L|}$. Each component in l_i contains a score for the corresponding label. A CRF [40] predicts the labels $(\hat{y}_1, ..., \hat{y}_n)$ with $\hat{y}_i \in L$ with the highest conditional joint probability $P(\hat{y}_1, ..., \hat{y}_n | l_1, ..., l_n)$. In this way, it makes use of patterns that appear in scientific papers (e.g. *Methods* are usually followed by *Results*). During training the CRF maximises $P(y_1, ..., y_n | l_1, ..., l_n)$ of the ground-truth labels for all training samples. The Viterbi algorithm [24] is used for efficient prediction and training.

For regularisation, we use dropout after each layer. We do not fine-tune SciBERT embeddings, since it requires training of 110 Mio. additional parameters.

3.2 Transfer Learning Methods

For sequential sentence classification, we tailor and evaluate the following transfer learning methods.

Sequential Transfer Learning (INIT):. The approach first trains the model for the source task and uses its tuned parameters to initialise the parameters for the target task. Then, the parameters are fine-tuned with the labelled data of the target task. As depicted in Figure 2(b), we propose two types of layer transfers. *INIT 1*: transfer parameters of *context enrichment* and *sentence encoding*; *INIT 2*: transfer parameters of *sentence encoding*. Other layers, except *word embedding*, of the target task are initialised with random values.

Multi-Task Learning (MULT):. Multi-task learning (MULT) aims for a better generalisation by simultaneously training samples in all tasks and sharing parameters of certain layers between the tasks. As depicted in Figure 2(c,d), we propose two multi-task learning architectures. The *MULT ALL* model shares all layers between the tasks except the *output layers* so that the model learns a common feature extractor for all tasks. However, full papers are much longer and have a different rhetorical structure compared to abstracts. Therefore, it is not beneficial to share the context enrichment layer between both dataset types. Thus, in the *MULT GRP* model, the *context enrichment layers* are only shared between datasets with the same text type. The objective functions are defined as:

$$L_{\text{MULT ALL}} = \sum_{t \in T^A \cup T^F} L_t(\Theta^S, \Theta^C, \Theta^O_t) \quad (1)$$

$$L_{\text{MULT GRP}} = \sum_{t \in T^A} L_t(\Theta^S, \Theta^{C^A}, \Theta^O_t) + \sum_{t \in T^F} L_t(\Theta^S, \Theta^{C^F}, \Theta^O_t) \quad (2)$$

where T^A and T^F are the tasks for datasets containing abstracts and full papers; L_t is the loss function for task t; the parameters Θ^S are for sentence encoding, Θ^C, Θ^{C^A}, Θ^{C^F} for context enrichment, and Θ^O_t for the output layer of task t.

Furthermore, we propose the variants *MULT ALL SHO* and *MULT GRP SHO* that are applicable if all tasks share the same (domain-independent) set of classes. *MULT ALL SHO* shares all layers among all tasks. *MULT GRP SHO* shares the context enrichment and output layer only between tasks with the same text type. Formally, the objective functions are defined as:

$$L_{\text{MULT ALL SHO}} = \sum_{t \in T^A \cup T^F} L_t(\Theta^S, \Theta^C, \Theta^O) \quad (3)$$

$$L_{\text{MULT GRP SHO}} = \sum_{t \in T^A} L_t(\Theta^S, \Theta^{C^A}, \Theta^{O^A}) + \sum_{t \in T^F} L_t(\Theta^S, \Theta^{C^F}, \Theta^{O^F}) \quad (4)$$

3.3 Semantic Relatedness of Classes

Datasets for sentence classification have different domain-specific annotation schemes, that is different sets of pre-defined classes. Intuitively, some classes have a similar meaning across domains, e.g. the classes *Model* and *Experiment* in the ART corpus are semantically related to *Methods* in PubMed-20k (PMD) (see Table 3). An analysis of semantic relatedness can help consolidate different annotation schemes. We propose machine learning models to support the identification of semantically related classes according to the following idea: If a model trained for PMD recognises sentences labelled with *ART:Model* as *PMD:Method*, and vice versa, then the classes *ART:Model* and *PMD:Method* can be assumed to be semantically related.

Let T be the set of all tasks, L the set of all classes in all tasks, $m_t(s)$ the label of sentence s predicted by the model for task t, and S^l the set of sentences with the ground truth label l. For each class $l \in L$ the corresponding semantic vector $v_l \in \mathbb{R}^{|L|}$ is defined as:

$$v_{l,l'} = \frac{\sum_{t \in T, s \in S^l} 1(m_t(s) = l')}{|S^l|} \tag{5}$$

where $v_{l,l'} \in \mathbb{R}$ is the component of the vector v_l for class $l' \in L$ and $1(p)$ is the indicator function that returns 1 if p is true and 0 otherwise. Intuitively, the semantic vectors concatenated vertically to a matrix represent a "confusion matrix" (see Figure 4 as an example). Now, we define the semantic relatedness of two classes $k, l \in L$ using cosine similarity:

$$\text{semantic\_relatedness}(k, l) = \cos(v_k, v_l) = \frac{v_k^\intercal \cdot v_l}{||v_k|| \cdot ||v_l||} \tag{6}$$

4 EXPERIMENTAL SETUP

This section describes the experimental evaluation of the proposed approaches, i.e. used datasets, implementation details, and evaluation methods.

4.1 Investigated Datasets

Table 3 summarises the characteristics of the investigated datasets, namely PubMed-20k (PMD) [18], NICTA-PIBOSO (NIC) [37], ART [45], and Dr. Inventor (DRI) [23]. The four datasets are publicly available and provide a good mix to investigate the transferability: They represent four different scientific domains; PMD and NIC cover abstracts and are from the same domain but have different annotation schemes; DRI and ART cover full papers but are from different domains and have different annotation schemes; NIC and DRI are rather small datasets, while PMD and ART are about 20 and 3 times larger, respectively; ART has a much finer annotation scheme compared to other datasets. As denoted in Table 3, the state-of-the-art results for ART are the lowest ones since ART has more fine-grained classes than the other datasets. In contrast, best results are obtained for PMD: It is a large dataset sampled from PubMed, where authors are encouraged to structure their abstracts. Therefore, abstracts in PMD are more uniformly structured than in other datasets, leading to better classification results.

4.2 Implementation

Our approaches are implemented in PyTorch [54]. The Adaptive Moment Estimation (ADAM) optimiser [38] with 0.01 weight decay and an exponential learning rate decay of 0.9 after each epoch is used for training. To speed up training, sentences longer than 128 tokens are truncated since the computational cost for the attention layers in BERT is quadratic in sentence length [71]. To reproduce the results of the original HSLN architecture, we tuned SciBERT-HSLN for PMD and NIC with hyperparameters as proposed in other studies [20, 34]. The following parameters performed best on the validation sets of PMD and NIC: learning rate 3e-5, dropout rate 0.5, Bi-LSTM hidden size $d^h = 2 \cdot 758$, $r = 15$ attention heads of size $d^u = 200$. We used these hyperparameters in all of our experiments.

For each dataset, we grouped papers to mini-batches without splitting them, if the mini-batch does not exceed 32 sentences. Thus, for full papers a mini-batch may consist of sentences from only one paper. During multi-task training we switched between the mini-batches of the tasks by proportional sampling [62]. After a mini-batch, only task-related parameters are updated, i.e. the associated output layer and all the layers below.

Table 3: Characteristics of the benchmark datasets. The row "SOTA" depicts the best results for approaches that do not exploit the ground-truth label of the preceding sentence during prediction: for PMD [75], for NIC [65], for DRI [4] (cf. Table 7), and for ART [44].

	PMD	NIC	DRI	ART
Domains	Biomedicine	Biomedicine	Computer Graphics	Chemistry, Comp. Linguistic
Text Type	Abstract	Abstract	Full article	Full article
# Articles	20.000	1.000	40	225
# Sentences	235.892	9.771	8.777	34.680
∅ # Sent.	12	10	219	154
# Classes	5	6	5	11
Classes	Background Objective Methods Results Conclusion	Background Intervention Study Population Outcome Other	Background Challenge Approach Outcome FutureWork	Background Motivation Hypothesis Goal Object Experiment Model Method Observation Result Conclusion
SOTA metric	[75] 93.1 weighted F1	[65] 86.8 weighted F1	[4] 72.5 weighted F1	[44] 51.6 accuracy

4.3 Evaluation

To be consistent with previous results and due to non-determinism in deep neural networks [57], we repeated the experiments and averaged the results. According to Cohan et al. [15] we performed three random restarts for PMD and NIC and used the same train/validation/test sets. For DRI and ART, we performed 10-fold and 9-fold cross-validation, respectively, as in the original papers [23, 44]. Within each fold the data is split into train/validation/test sets with the proportions $\frac{k-2}{k}/\frac{1}{k}/\frac{1}{k}$ where k is the number of folds. For multi-task learning, the experiment was repeated with the maximum number of folds of the datasets used, but at least three times. All models were trained for 20 epochs. The test set performance within a fold and restart, respectively, was calculated for the epoch with the best validation performance.

Table 4: Experimental results for the proposed approaches: our SciBERT-HSLN model without transfer learning, parameter initialisation (INIT), and multi-task learning (MULT ALL and MULT GRP). Previous state of the art (see Table 3), SciBERT-[CLS], and the approaches of Jin and Szolovits [34] and Cohan et al. [15] are the baseline results. For PMD (P), NIC (N), and DRI (D) we report weighted F1 score and for ART (A) accuracy. The average of all scores is denoted by ∅. *Italics* depicts whether the result is better than the baseline, bold whether the transfer method improves SciBERT-HSLN, underline the best overall result.

	PMD	NIC	DRI	ART	∅
Prev. SOTA	[75] 93.1	[65] 86.8	[4] 72.5	[44] 51.6	76.0
SciBERT-[CLS] [6]	89.6	78.4	69.5	51.5	72.3
Jin and Szolovits [34]	92.6	84.7	75.3	49.3	75.5
Cohan et al. [15]	92.9	84.8	74.3	54.3	76.6
SciBERT-HSLN	92.9	*84.9*	*78.0*	*58.0*	*78.5*
INIT 1 PMD to T	-	84.8	***81.2***	*57.7*	
INIT 2 PMD to T	-	84.8	***80.1***	*58.0*	
INIT 1 NIC to T	92.9	-	***81.9***	*57.6*	
INIT 2 NIC to T	92.9	-	***79.6***	*57.2*	
INIT 1 DRI to T	92.9	83.5	-	*57.8*	
INIT 2 DRI to T	92.9	83.8	-	*57.6*	
INIT 1 ART to T	**93.0**	84.7	***82.2***	-	
INIT 2 ART to T	92.9	84.7	***81.0***	-	
MULT ALL	**93.0**	**86.0**	***81.8***	*57.7*	***79.6***
PMD, NIC	**93.0**	**86.1**	-	-	
PMD, DRI	92.9	-	***80.6***	-	
PMD, ART	**93.0**	-	-	*58.0*	
NIC, DRI	-	84.2	***80.7***	-	
NIC, ART	-	84.4	-	*57.9*	
DRI, ART	-	-	***82.0***	*57.6*	
PMD, NIC, DRI	**93.0**	**86.2**	***81.0***	-	
PMD, NIC, ART	**93.0**	**86.3**	-	*58.0*	
PMD, DRI, ART	**93.0**	-	***82.7***	*57.8*	
NIC, DRI, ART	-	84.7	***82.0***	*57.7*	
MULT GRP	**93.0**	**86.1**	***83.4***	***58.8***	***80.3***
P,N,D,A	92.9	**85.4**	***84.4***	*58.0*	***80.2***
(P,D),(N,A)	**93.0**	**86.0**	***81.1***	***58.5***	***79.7***
(P,A),(N,D)	92.9	**85.8**	***83.6***	*58.0*	***80.1***
(P,N,D),(A)	92.9	**86.0**	***80.6***	***58.2***	***79.4***
(P,N,A),(D)	**93.0**	**86.0**	***84.1***	***58.1***	***80.3***
(P,D,A),(N)	92.9	**85.5**	***82.2***	*58.0*	***79.6***
(N,D,A),(P)	92.9	**85.9**	***83.3***	***58.5***	***80.1***

We compare our results only with approaches which do not exploit *ground-truth labels* of the preceding sentence as a feature *during prediction* (see Section 2.1). This has a significant impact on the performance: Using the ground truth label of the previous sentences as a sole input feature to a SVM classifier already yields an accuracy of 77.7 for DRI and 55.5 for ART (compare also results for the "history" feature in [4], cf. Table 5). Best reported results using ground truth labels as input features have an accuracy of 84.15 for DRI and 65.75 for ART [2]. In contrast, we pursue a realistic setting by exploiting the *predicted* (not ground truth) label of neighbouring sentences during prediction.

Moreover, we provide additional results for three strong deep learning baselines: (1) fine-tuning SciBERT using the [CLS] token of individual sentences as in [20] (referred to as SciBERT-[CLS]), (2) original HSLN implementation of Jin and Szolovits [34], and (3) the SciBERT-based approach of Cohan et al. [15]. We cannot provide baseline results for DRI and ART of the approaches [65, 75] since their implementations are not publicly available.

Table 5: Experimental results for μPMD, NIC, DRI and μART with our SciBERT-HSLN model and our proposed multi-task learning approaches.

	μPMD	NIC	DRI	μART	∅
SciBERT-HSLN	90.9	84.9	78.0	52.2	76.5
MULT ALL	**91.1**	**85.7**	**81.0**	**53.8**	**77.9**
MULT GRP	**91.1**	**85.9**	**82.2**	**55.1**	**78.6**

5 RESULTS AND DISCUSSION

In this section, we present and discuss the experimental results for our proposed cross-domain multi-task learning approach for sequential sentence classification. The results for different variations of our approach, the respective baselines, and for several state-of-the-art methods are depicted in Table 4. The results are discussed in the following three subsections with regard to the unified approach without transfer learning (Section 5.1), with sequential transfer learning (Section 5.2), and multi-task learning (Section 5.3). Section 5.4 analyses the semantic relatedness of classes for the four annotation schemes.

5.1 Unified Approach without Transfer Learning (SciBERT-HSLN)

For the full paper datasets DRI and ART, our SciBERT-HSLN model significantly outperforms the previously reported best results, and the deep learning baselines SciBERT-[CLS], Jin and Szolovits [34], and Cohan et al. [15]. The previous state of the art approaches for DRI and ART [4, 44] require feature engineering and a sentence is enriched only with the context of the previous sentence. In SciBERT-[CLS], each sentence is classified in isolation. The original HSLN architecture [34] uses shallow word embeddings pre-trained on biomedical texts. Thus, the incorporation of SciBERT's contextual word embeddings into HSLN helps improve performance for the DRI and ART datasets. The approach of Cohan et al. [15] can process only about 10 sentences at once since SciBERT supports sequences of up to 512 tokens only. Thus, long text has to be split into multiple chunks. Our deep learning approach can process *all* sentences of a paper at once so that all sentences are enriched with context from surrounding sentences.

For the PMD dataset, our SciBERT-HSLN results are equivalent [75] to the current state of the art, while for NIC, they are below [65]. Thus, our proposed approach is competitive with the current approaches for sequential sentence classification in abstracts. *Our unified deep learning approach is applicable to datasets consisting of different text types, i.e. abstracts and full papers, without any feature engineering (RQ7).*

5.2 Sequential Transfer Learning (INIT)

Using the INIT approach, we can only improve the baseline results for the DRI dataset in all settings. The approach INIT 1 performs better than INIT 2 in most cases which indicates that transferring all parameters is more effective. *However, the results suggest that sequential transfer learning is not a very effective transfer method for sequential sentence classification (RQ2).*

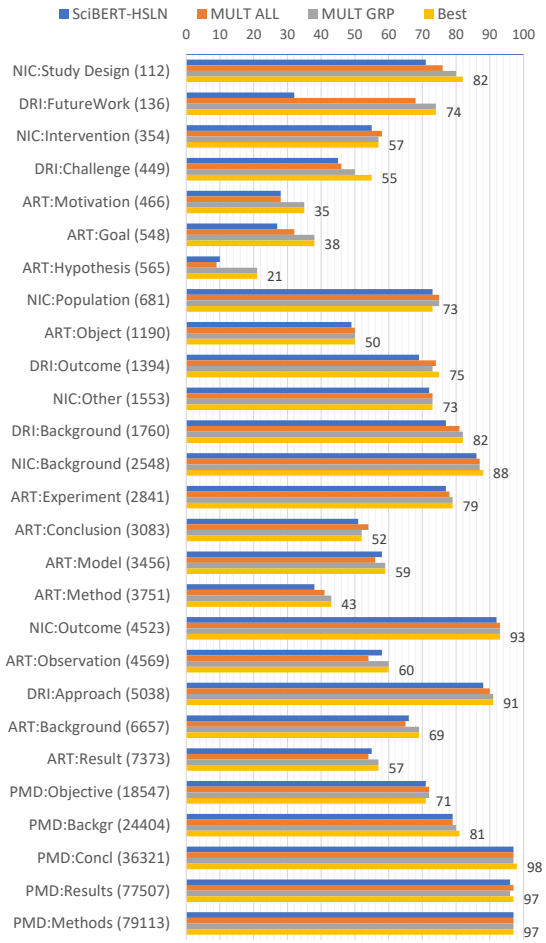

Figure 3: F1 scores per class for the datasets PMD, NIC, DRI, and ART for SciBERT-HSLN, MULT ALL, MULT GRP, and the best combination for the respective dataset. Numbers at the bars depict the F1 scores of the best classifiers and in brackets the number of examples for the given class. The classes are ordered by the number of examples.

5.3 Multi-Task Learning (MULT)

Next, we discuss the results of our multi-task learning approach, and the effects of multi-task learning on smaller datasets and individual sentence classes.

MULT ALL model: All tasks were trained jointly sharing all possible layers. Except for the ART task, all results are improved using the SciBERT-HSLN model. For the PMD task, the improvement is marginal since the baseline results (F1 score) were already on a high level. Pairwise MULT ALL combinations show that the models for PMD and NIC, respectively, benefit from the (respective) other dataset, and the DRI model especially from the ART dataset. The PMD and NIC datasets are from the same domain, and both contain abstracts, so the results are as expected. Furthermore, DRI and ART datasets both contain full papers, and DRI has more coarse-grained classes. However, ART is a related large dataset with fine-grained classes and presumably therefore the model for ART does not benefit from other datasets. In triple-wise MULT ALL combinations the models for PMD and DRI, respectively, benefit from all datasets, and the model for NIC only if the PMD dataset is present. *The results suggest that sharing all possible layers between multiple tasks is effective except for bigger datasets with more fine-grained classes (RQ3, RQ4).*

MULT GRP model: In this setting, the models for all tasks were trained jointly, but only models for the same text type share the *context enrichment layer*, i.e. (PMD, NIC) and (DRI, ART). Here, all models benefit from the other datasets. In our ablation study, we also provide results for sharing only the *sentence encoding layer*, referred to as MULT GRP P,N,D,A, and all pairwise and triple-wise combinations sharing the *context enrichment layer*. Other combinations also yield good results. However, MULT GRP is effective for *all* tasks. *Our results indicate that sharing the sentence encoding layer between multiple models is beneficial. Furthermore, sharing the context enrichment layer only between models for the same text type is an even more effective strategy (RQ3, RQ4).*

Effect of Dataset Size: The NIC and DRI models benefit more from multi-task learning than PMD and ART. However, PMD and ART are bigger datasets than NIC and DRI. The ART dataset has also more fine-grained classes than the other datasets. This raises the following question: *How would the models for PMD and ART benefit from multi-task learning if they were trained on smaller datasets?*

To answer this question, we created smaller variants of PMD and ART (i.e. μPMD and μART) with a comparable size with NIC and DRI. The training data was truncated to $\frac{1}{20}$ for μPMD and $\frac{1}{3}$ for μART while keeping the original size of the validation and test sets. As shown in Table 5, all models benefit from the other datasets, whereas MULT GRP again performs best. *The results indicate that models for small datasets benefit from multi-task learning independent of the difference in the granularity of the classes (RQ1).*

Effect for each Class: Figure 3 shows the F1 scores per class for the investigated approaches. Classes, which are intuitively highly semantically related (*:Background, *:Results, *:Outcome), and classes with few examples (DRI:FutureWork, DRI:Challenge, ART:Hypothesis, NIC:Study Design) tend to benefit significantly from multi-task learning. The classes ART:Model, ART:Observation, and ART:Result have worse results than SciBERT-HSLN when using MULT ALL, but MULT GRP yields better results. This can be attributed to sharing the *context enrichment layers* only between datasets with the same text type. *The analysis suggests that especially semantically related classes and classes with few examples benefit from multi-task learning (RQ1).*

5.4 Semantic Relatedness of Classes

In this section, we first evaluate our proposed approach for the semi-automatical identification of semantically related classes in the datasets PMD, NIC, DRI, and ART. Based on the analysis, we identify six clusters of semantically related classes. Then, we present a new dataset that is compiled from the investigated datasets and is based on the identified clusters. As a possible down-stream application, this multi-domain dataset with a generic set of classes could help to structure research papers in a domain-independent

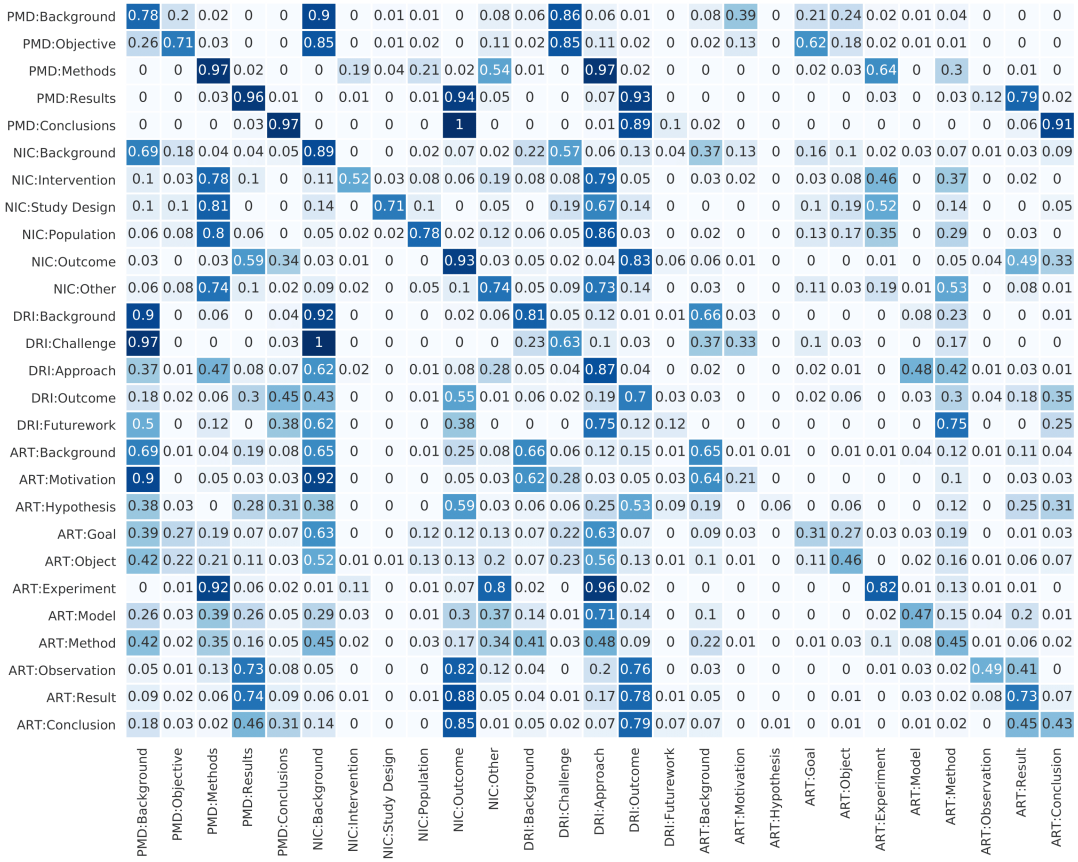

Figure 4: Each row represents a semantic vector as described in Section 3.3 for a class computed with *MULT ALL* classifier.

manner, supporting, for instance, the development of academic search engines.

Analysis of Semantic Relatedness of Classes: Based on the annotation guidelines of the investigated datasets PMD [18], NIC [37], DRI [23], and ART [45], we identified six clusters of semantically related classes, which are depicted in Figure 5. The identification process of the clusters followed the intuition, that most research papers independent of the scientific domain (1) investigate a research problem (*Problem*), (2) provide background information for the problem (*Background*), (3) apply or propose certain methods (*Methods*), (4) yield results (*Results*), (5) conclude the work (*Conclusions*), and (6) outline future work (*Future Work*).

Figure 5 shows the semantic vectors for each label, which were computed with MULT ALL and projected to 2D space using principal component analysis [35]. Each label is assigned to one of the generic clusters for semantically related labels. Except *Problem*, all clusters for semantically related classes are well identifiable in Figure 5. The semantic vector for ART:Hypothesis is an outlier in the *Problem* cluster because ART:Hypothesis is confused mostly with ART:Conclusion and ART:Result (see Figure 4) and has also a very low F1 score (see Figure 3). Table 6 shows the Silhoutte scores [58] for each cluster. A positive score indicates that objects lie well within the cluster, and a negative score that the objects

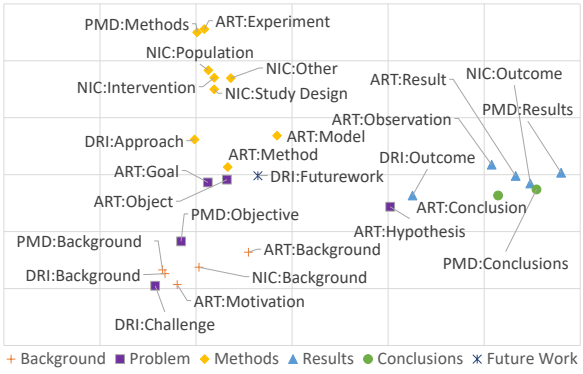

Figure 5: Semantic vectors of labels computed with MULT ALL, and projected to 2D space via PCA. The semantic vectors are assigned to generic clusters of semantically related labels.

are merely somewhere in between clusters. As a distance metric, we use semantic_relatedness as defined in Equation 6. The silhouette coefficients show that MULT ALL forms better clusters than SciBERT-HSLN and MULT GRP, although MULT GRP performs best. We hypothesise that MULT ALL can better capture the semantic

Table 6: Silhouette scores per cluster and overall computed for the semantic vectors of SciBERT-HSLN, MULT GRP and MULT ALL classifiers.

	SciBERT-HSLN	MULT GRP	MULT ALL
Background	0.45	0.18	**0.48**
Problem	-0.27	**-0.04**	-0.29
Methods	0.19	-0.03	**0.31**
Results	-0.38	0.01	**0.32**
Conclusions	**0.92**	-0.49	0.02
Future Work	0.00	0.00	0.00
Overall	0.10	-0.02	**0.20**

Table 7: Characteristics of the domain-independent dataset *G-PNDA* that was compiled from the original datasets *PMD, NIC, DRI,* and *ART*.

	G-PMD	**G-NIC**	**G-DRI**	**G-ART**
Text Type	Abstract	Abstract	Full paper	Full paper
# Papers	1.000	1.000	40	67
# Sentences	11.738	9.771	8.777	9.528
∅ # Sentences	11	10	219	142
Background	1.220	2.548	1.760	1.657
Problem	953	0	449	529
Methods	3.927	2.700	5.038	2.752
Results	3.760	4.523	1.394	3.672
Conclusions	1.878	0	0	918
Future Work	0	0	136	0

relatedness of labels than the other approaches since it is enforced to learn a generic feature extractor across multiple datasets. *The multi-task learning approach sharing all possible layers is able to recognise semantically related classes (RQ5).*

Domain-Independent Sentence Classification: Based on the identified clusters, we compile a new dataset *G-PNDA* from the investigated datasets PMD, NIC, DRI, and ART. The labels of the datasets are collapsed according to the clusters in Figure 5. Table 7 summarises the characteristics of the compiled dataset. To prevent a bias towards bigger datasets, we truncate PMD to $\frac{1}{20}$ and ART to $\frac{1}{3}$ of their original size.

Table 8 depicts our experimental results for the generic dataset *G-PNDA*. We train a model for each dataset part, and the multi-task learning models MULT ALL and MULT GRP. Since we have common sentence classes now, we train also models that share the output layers between the dataset parts, referred to as MULT ALL SHO and MULT GRP SHO (see Section 3.2). For training and evaluation, we split each dataset into train/validation/test sets with the portions 70/10/20, average the results over three random restarts and use the same hyperparameters as before (see Section 4.2).

Table 8 shows that the proposed MULT GRP model outperforms all other settings. Surprisingly, sharing the output layer impairs the performance in all settings. We can attribute this to the fact that the output layer learns different transition distributions between the classes. *Thus, in a domain-independent setting a separate output layer per dataset part helps the model to capture the individual rhetorical structure present in the domains (RQ3, RQ6).*

6 CONCLUSIONS

In this paper, we have presented a unified deep learning architecture for sequential sentence classification. The unified approach

Table 8: Experimental results (F1 scores) for our proposed approaches for the dataset *G-PNDA*: baseline model SciBERT-HSLN with one separate model per dataset and the multi-task learning models MULT ALL SHO, MULT ALL, MULT GRP SHO, and MULT GRP. Bold depicts whether the approach improves the baseline, underline the best overall result.

	G-PMD	G-NIC	G-DRI	G-ART	∅
SciBERT-HSLN (one model per dataset)	90.1	89.3	81.7	70.8	83.0
MULT ALL SHO (shared output layer)	89.8	89.1	**83.5**	67.1	82.4
MULT ALL (separate output layer)	**90.5**	**89.8**	**84.9**	70.5	**83.9**
MULT GRP SHO (shared output layer)	90.0	**89.9**	**86.1**	70.4	**84.1**
MULT GRP (separate output layer)	**90.6**	**89.7**	**87.2**	**71.0**	**84.6**

can be applied to datasets that contain abstracts as well as full articles. For datasets of full papers, the unified approach significantly outperforms the state of the art without any feature engineering.

Furthermore, we have tailored two common transfer learning approaches to sequential sentence classification and compared their performance. We found that training a multi-task model with multiple datasets works better than sequential transfer learning. Our comprehensive experimental evaluation with four different datasets offers useful insights under which conditions transferring or sharing of specific layers is beneficial or not. In particular, it is always beneficial to share the sentence encoding layer between datasets from different domains. However, it is most effective to share the context enrichment layer, which encodes the context of neighbouring sentences, only between datasets with the same text type (abstracts vs. full papers). This can be attributed to different rhetorical structures in abstracts and full papers. Our tailored multi-task learning approach makes use of multiple datasets and yields new state-of-the-art results for two full paper datasets. In particular, models for tasks with small datasets and classes with few labelled examples benefit significantly from models of other tasks.

Our study suggests that the classes of the different dataset annotation schemes are semantically related, even though the datasets come from different domains and have different text types (e.g. abstract or full papers). This semantic relatedness is an important prerequisite for transfer learning in NLP tasks [48, 52, 59],

Finally, we proposed an approach to semi-automatically identify semantically related classes from different datasets to support manual comparison and inspection of different annotation schemes across domains. We demonstrated the usefulness of the approach with an analysis of four annotation schemes. This approach can support the investigation of annotation schemes across disciplines without re-annotating datasets. From the analysis, we derived a domain-independent consolidated annotation scheme and compiled a domain-independent dataset. This allows for the classification of sentences in research papers with generic classes across disciplines, which can support, for instance, academic search engines.

In future work, we plan to integrate other tasks (e.g. scientific concept extraction) into the multi-task learning approach to exploit further datasets. Furthermore, we intend to evaluate the domain-independent sentence classifier in an information retrieval scenario.

REFERENCES

[1] Ahmed AbuRa'ed, Horacio Saggion, Alexander Shvets, and Àlex Bravo. 2020. Automatic related work section generation: experiments in scientific document abstracting. *Scientometrics* 125, 3 (2020), 3159–3185. https://doi.org/10.1007/s11192-020-03630-2

[2] Nasrin Asadi, Kambiz Badie, and Maryam Tayefeh Mahmoudi. 2019. Automatic zone identification in scientific papers via fusion techniques. *Scientometrics* 119, 2 (2019), 845–862. https://doi.org/10.1007/s11192-019-03060-9

[3] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M. Cer, and David Jurgens (Eds.). Association for Computational Linguistics, 546–555. https://doi.org/10.18653/v1/S17-2091

[4] Kambiz Badie, Nasrin Asadi, and Maryam Tayefeh Mahmoudi. 2018. Zone identification based on features with high semantic richness and combining results of separate classifiers. *J. Inf. Telecommun.* 2, 4 (2018), 411–427. https://doi.org/10.1080/24751839.2018.1460083

[5] Soumya Banerjee, Debarshi Kumar Sanyal, Samiran Chattopadhyay, Plaban Kumar Bhowmick, and Partha Pratim Das. 2020. Segmenting Scientific Abstracts into Discourse Categories: A Deep Learning-Based Approach for Sparse Labeled Data. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*, Ruhua Huang, Dan Wu, Gary Marchionini, Daqing He, Sally Jo Cunningham, and Preben Hansen (Eds.). ACM, 429–432. https://doi.org/10.1145/3383583.3398598

[6] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3613–3618. https://doi.org/10.18653/v1/D19-1371

[7] Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.* 66, 11 (2015), 2215–2222. https://doi.org/10.1002/asi.23329

[8] Arthur Brack, Jennifer D'Souza, Anett Hoppe, Sören Auer, and Ralph Ewerth. 2020. Domain-Independent Extraction of Scientific Concepts from Research Articles. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12035)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 251–266. https://doi.org/10.1007/978-3-030-45439-5_17

[9] Arthur Brack, Anett Hoppe, Markus Stocker, Sören Auer, and Ralph Ewerth. 2022. Analysing the requirements for an Open Research Knowledge Graph: use cases, quality requirements, and construction strategies. *Int. J. Digit. Libr.* 23, 1 (2022), 33–55. https://doi.org/10.1007/s00799-021-00306-x

[10] Arthur Brack, Daniel Uwe Müller, Anett Hoppe, and Ralph Ewerth. 2021. Coreference Resolution in Research Papers from Multiple Domains. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12656)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 79–97. https://doi.org/10.1007/978-3-030-72113-8_6

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[12] Soravit Changpinyo, Hexiang Hu, and Fei Sha. 2018. Multi-Task Learning for Sequence Tagging: An Empirical Study. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, 2965–2977. https://www.aclweb.org/anthology/C18-1251/

[13] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1724–1734. https://doi.org/10.3115/v1/d14-1179

[14] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 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-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 3586–3596. https://doi.org/10.18653/v1/n19-1361

[15] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S. Weld. 2019. Pretrained Language Models for Sequential Sentence Classification. 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3691–3697. https://doi.org/10.18653/v1/D19-1383

[16] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 615–621. https://doi.org/10.18653/v1/N18-2097

[17] Carmen Dayrell, Arnaldo Candido Jr., Gabriel Lima, Danilo Machado Jr., Ann A. Copestake, Valéria Delisandra Feltrim, Stella E. O. Tagnin, and Sandra M. Aluísio. 2012. Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), 1604–1609. http://www.lrec-conf.org/proceedings/lrec2012/summaries/734.html

[18] Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, Greg Kondrak and Taro Watanabe (Eds.). Asian Federation of Natural Language Processing, 308–313. https://www.aclweb.org/anthology/I17-2052/

[19] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neural Networks for Joint Sentence Classification in Medical Paper Abstracts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, Mirella Lapata, Phil Blunsom, and Alexander Koller (Eds.). Association for Computational Linguistics, 694–700. https://doi.org/10.18653/v1/e17-2110

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[21] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS\^2: Multi-Document Summarization of Medical Studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 7494–7513. https://aclanthology.org/2021.emnlp-main.594

[22] Christiane Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

[23] Beatríz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the Discoursive Structure of Computer Graphics Research Papers. In *Proceedings of The 9th Linguistic Annotation Workshop, LAW@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA*, Adam Meyers, Ines Rehbein, and Heike Zinsmeister (Eds.). The Association for Computer Linguistics, 42–51. https://doi.org/10.3115/v1/w15-1605

[24] G. D. Forney. 1973. The viterbi algorithm. *Proc. IEEE* 61, 3 (1973), 268–278. https://doi.org/10.1109/PROC.1973.9030

[25] Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 1255–1268. https://doi.org/10.18653/v1/2020.acl-main.116

[26] Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*

2018, New Orleans, Louisiana, USA, June 5-6, 2018, Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat (Eds.). Association for Computational Linguistics, 679–688. https://doi.org/10.18653/v1/s18-1111

[27] Sayar Ghosh Roy, Nikhil Pinnaparaju, Risubh Jain, Manish Gupta, and Vasudeva Varma. 2020. Summaformers @ LaySumm 20, LongSumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics, Online, 336–343. https://doi.org/10.18653/v1/2020.sdp-1.39

[28] Sérgio Gonçalves, Paulo Cortez, and Sérgio Moro. 2020. A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Comput. Appl.* 32, 11 (2020), 6793–6807. https://doi.org/10.1007/s00521-019-04334-2

[29] Komal Gupta, Ammaar Ahmad, Tirthankar Ghosal, and Asif Ekbal. 2021. ContriSci: A BERT-Based Multitasking Deep Neural Architecture to Identify Contribution Statements from Research Papers. In *Towards Open and Trustworthy Digital Societies - 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1-3, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 13133)*, Hao-Ren Ke, Chei Sian Lee, and Kazunari Sugiyama (Eds.). Springer, 436–452. https://doi.org/10.1007/978-3-030-91669-5_34

[30] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-Enhanced Bert with Disentangled Attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=XPZIaotutsD

[31] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[32] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 328–339. https://doi.org/10.18653/v1/P18-1031

[33] Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-Level N-ary Relation Extraction with Multiscale Representation Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 3693–3704. https://doi.org/10.18653/v1/n19-1370

[34] Di Jin and Peter Szolovits. 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, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 3100–3109. https://doi.org/10.18653/v1/d18-1349

[35] Ian T. Jolliffe. 2011. Principal Component Analysis. In *International Encyclopedia of Statistical Science*, Miodrag Lovric (Ed.). Springer, 1094–1096. https://doi.org/10.1007/978-3-642-04898-2_455

[36] Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2021. Automated Mining of Leaderboards for Empirical AI Research. In *Towards Open and Trustworthy Digital Societies - 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1-3, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 13133)*, Hao-Ren Ke, Chei Sian Lee, and Kazunari Sugiyama (Eds.). Springer, 453–470. https://doi.org/10.1007/978-3-030-91669-5_35

[37] Su Kim, David Martínez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinform.* 12, S-2 (2011), S5. https://doi.org/10.1186/1471-2105-12-S2-S5

[38] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[39] Suchetha Nambanoor Kunnath, David Pride, Bikash Gyawali, and Petr Knoth. 2020. Overview of the 2020 WOSP 3C Citation Context Classification Task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*. Association for Computational Linguistics, Wuhan, China, 75–83. https://www.aclweb.org/anthology/2020.wosp-1.12

[40] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, Carla E. Brodley and Andrea Pohoreckyj Danyluk (Eds.). Morgan Kaufmann, 282–289.

[41] Anne Lauscher, Goran Glavas, and Kai Eckert. 2018. ArguminSci: A Tool for Analyzing Argumentation and Rhetorical Aspects in Scientific Writing. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, Noam Slonim and Ranit Aharonov (Eds.). Association for Computational Linguistics, 22–28. https://doi.org/10.18653/v1/w18-5203

[42] Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, and Kai Eckert. 2018. Investigating the Role of Argumentation in the Rhetorical Analysis of Scientific Publications with Neural Multi-Task Learning Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 3326–3338. https://doi.org/10.18653/v1/d18-1370

[43] Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer Learning for Named-Entity Recognition with Neural Networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/878.html

[44] Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin R. Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinform.* 28, 7 (2012), 991–1000. https://doi.org/10.1093/bioinformatics/bts071

[45] Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin R. Batchelor. 2010. Corpora for the Conceptualisation and Zoning of Scientific Papers. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (Eds.). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2010/summaries/644.html

[46] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 3219–3232. https://doi.org/10.18653/v1/d18-1360

[47] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 3111–3119. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[48] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How Transferable are Neural Networks in NLP Applications?. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 479–489. https://doi.org/10.18653/v1/d16-1046

[49] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics* 117, 3 (2018), 1931–1990. https://doi.org/10.1007/s11192-018-2921-5

[50] Mariana L. Neves, Daniel Butzke, and Barbara Grune. 2019. Evaluation of Scientific Elements for Text Similarity in Biomedical Publications. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*, Benno Stein and Henning Wachsmuth (Eds.). Association for Computational Linguistics, 124–135. https://doi.org/10.18653/v1/w19-4515

[51] Allard Oelen, Markus Stocker, and Sören Auer. 2021. Crowdsourcing Scholarly Discourse Annotations. In *IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021*, Tracy Hammond, Katrien Verbert, Dennis Parra, Bart P. Knijnenburg, John O'Donovan, and Paul Teale (Eds.). ACM, 464–474. https://doi.org/10.1145/3397481.3450685

[52] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[53] Seoyeon Park and Cornelia Caragea. 2020. Scientific Keyphrase Identification and Classification by Pre-Trained Language Models Intermediate Task Transfer Learning. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 5409–5419. https://doi.org/10.18653/v1/2020.coling-main.472

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[55] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on*

Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. https://doi.org/10.3115/v1/d14-1162

[56] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5231–5247. https://doi.org/10.18653/v1/2020.acl-main.467

[57] Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 338–348. https://doi.org/10.18653/v1/d17-1035

[58] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53 – 65. https://doi.org/10.1016/0377-0427(87)90125-7

[59] Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph. D. Dissertation. National University of Ireland, Galway.

[60] Iqra Safder and Saeed-Ul Hassan. 2019. Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. *Scientometrics* 119, 1 (2019), 257–277. https://doi.org/10.1007/s11192-019-03025-y

[61] Iqra Safder, Saeed-Ul Hassan, Anna Visvizi, Thanapon Noraset, Raheel Nawaz, and Suppawong Tuarob. 2020. Deep Learning-based Extraction of Algorithmic Metadata in Full-Text Scholarly Documents. *Inf. Process. Manag.* 57, 6 (2020), 102269. https://doi.org/10.1016/j.ipm.2020.102269

[62] Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A Hierarchical Multi-Task Approach for Learning Embeddings from Semantic Tasks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 6949–6956. https://doi.org/10.1609/aaai.v33i01.33016949

[63] Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-Task Learning for Argumentation Mining in Low-Resource Settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 35–41. https://doi.org/10.18653/v1/n18-2006

[64] Tushar Semwal, Promod Yenigalla, Gaurav Mathur, and Shivashankar B. Nair. 2018. A Practitioners' Guide to Transfer Learning for Text Classification using Convolutional Neural Networks. In *Proceedings of the 2018 SIAM International Conference on Data Mining, SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA*, Martin Ester and Dino Pedreschi (Eds.). SIAM, 513–521. https://doi.org/10.1137/1.9781611975321.58

[65] Xichen Shang, Qianli Ma, Zhenxi Lin, Jiangyue Yan, and Zipeng Chen. 2021. A Span-based Dynamic Local Attention Model for Sequential Sentence Classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 198–203. https://doi.org/10.18653/v1/2021.acl-short.26

[66] Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask Semi-Supervised Learning for Class-Imbalanced Discourse Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 498–517. https://aclanthology.org/2021.emnlp-main.40

[67] Connor Stead, Stephen Smith, Peter A. Busch, and Savanid Vatanasakdakul. 2019. Emerald 110k: A Multidisciplinary Dataset for Abstract Sentence Classification. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, ALTA 2019, Sydney, Australia, December 4-6, 2019*, Meladel Mistica, Massimo Piccardi, and Andrew MacKinlay (Eds.). Australasian Language Technology Association, 120–125. https://aclweb.org/anthology/papers/U/U19/U19-1016/

[68] Xuefeng Su, Ru Li, and Xiaoli Li. 2020. Multi-domain Transfer Learning for Text Classification. In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12430)*, Xiaodan Zhu, Min Zhang, Yu Hong, and Ruifang He (Eds.). Springer, 457–469. https://doi.org/10.1007/978-3-030-60450-9_36

[69] Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph. D. Dissertation. University of Edinburgh.

[70] Simone Teufel, Advaith Siddharthan, and Colin R. Batchelor. 2009. Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 1493–1502. https://www.aclweb.org/anthology/D09-1155/

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[72] Zhepei Wei, Yantao Jia, Yuan Tian, Mohammad Javad Hosseini, Mark Steedman, and Yi Chang. 2019. Joint Extraction of Entities and Relations with a Hierarchical Multi-task Tagging Model. *CoRR* abs/1908.08672 (2019). arXiv:1908.08672 http://arxiv.org/abs/1908.08672

[73] Karl R. Weiss, Taghi M. Khoshgoftaar, and Dingding Wang. 2016. A survey of transfer learning. *J. Big Data* 3 (2016), 9. https://doi.org/10.1186/s40537-016-0043-6

[74] Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 1271–1279. https://doi.org/10.1145/3038912.3052558

[75] Kosuke Yamada, Tsutomu Hirao, Ryohei Sasano, Koichi Takeda, and Masaaki Nagata. 2020. Sequential Span Classification with Neural Semi-Markov CRFs for Biomedical Abstracts. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 871–877. https://doi.org/10.18653/v1/2020.findings-emnlp.77

[76] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=ByxpMd9lx

[77] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). The Association for Computational Linguistics, 1480–1489. https://doi.org/10.18653/v1/n16-1174