International Journal on Digital Libraries
https://doi.org/10.1007/s00799-023-00392-z

Sequential sentence classification in research papers using cross-domain multi-task learning

Arthur Brack[1,3] · Elias Entrup[1] · Markos Stamatakis[2] · Pascal Buschermöhle[3] · Anett Hoppe[1,2] · Ralph Ewerth[1,2]

Received: 31 October 2022 / Revised: 21 November 2023 / Accepted: 17 December 2023

Abstract

The automatic semantic structuring of scientific text allows for more efficient reading of research articles and is an important indexing step for academic search engines. Sequential sentence classification is an essential structuring task and targets the categorisation of sentences based on their content and context. However, the potential of transfer learning for sentence classification across different scientific domains and text types, such as full papers and abstracts, has not yet been explored in prior work. In this paper, we present a systematic analysis of transfer learning for scientific sequential sentence classification. For this purpose, we derive seven research questions and present several contributions to address them: (1) We suggest a novel uniform deep learning architecture and multi-task learning for cross-domain sequential sentence classification in scientific text. (2) We tailor two transfer learning methods to deal with the given task, namely sequential transfer learning and multi-task learning. (3) We compare the results of the two best models using qualitative examples in a case study. (4) We provide an approach for the semi-automatic identification of semantically related classes across annotation schemes and analyse the results for four annotation schemes. The clusters and underlying semantic vectors are validated using *k*-means clustering. (5) Our comprehensive experimental results indicate that when using the proposed multi-task learning architecture, models trained on datasets from different scientific domains benefit from one another. Our approach significantly outperforms state of the art on full paper datasets while being on par for datasets consisting of abstracts.

Keywords Sequential sentence classification · Zone identification · Transfer learning · Multi-task learning · Scholarly communication

1 Introduction

Searching 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 [11, 70, 85]. The exploding number of published articles aggravates this situation further [8]. Therefore, automatic approaches to structure research papers are highly desirable.

The task of *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 whether a particular sentence contains information about the research work's objective, methods or results [22]. Figure 1 shows an example of an abstract with classified sentences. Such a semantification of sentences can help algorithms focus on relevant elements of text and thus assist information retrieval systems [60, 70] or knowledge graph population [61]. The task is called *sequential* to distinguish it from the general *sentence classification* task where a sen-

✉ Anett Hoppe
anett.hoppe@tib.eu

Arthur Brack
arthur.brack@set.de

Ralph Ewerth
ralph.ewerth@tib.eu

[1] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany

[2] L3S Research Center, Leibniz University Hannover, Hannover, Germany

[3] Leibniz University Hannover, Hannover, Germany

Published online: 22 January 2024

tence 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 of neighbouring sentences, e.g. sentences that describe the methods usually precede sentences about results.

In previous work, several approaches have been proposed for *sequential sentence classification* (e.g. [3, 41, 75, 86]), and several datasets were annotated for various scientific domains (e.g. [22, 29, 34, 77]). The datasets contain either abstracts or full papers and were annotated with domain-specific sentence classes. However, research infrastructures usually 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 prediction accuracy. For this purpose, we explore several research questions.

First, although some approaches propose transfer learning for the scientific domain [6, 12, 36, 63], the field lacks a comprehensive empirical study on transfer learning across different scientific domains for *sequential sentence classification*. Transfer learning enables us to combine knowledge from multiple datasets to improve classification performance and, thus, can reduce annotation costs. The annotation of scientific text is particularly costly since it demands expertise in the article's domain [4, 9, 32]. However, studies revealed that the success of transferring neural models largely depends on the relatedness of the tasks, and transfer learning with unrelated tasks may even degrade performance [58, 62, 69, 74]. Two tasks are related if there exists some implicit or explicit relationship between the feature spaces [62]. On the other hand, every scientific domain is characterised by its specific terminology and phrasing, which yields different feature spaces. Thus, it is unclear to what extent datasets from different scientific disciplines are related. This raises the following research questions (RQ) for the task of sequential sentence classification:

> *#RQ1* To what extent are datasets from different 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?

Typically, 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 [53] annotated a dataset multiple times with different schemes and analysed the multi-variate frequency distributions of the classes. They found that the investigated schemes are complementary and should be combined. However, annotating datasets multiple times is costly and time-consuming. To support the consolidation of different annotation schemes across domains, we examine the following RQs:

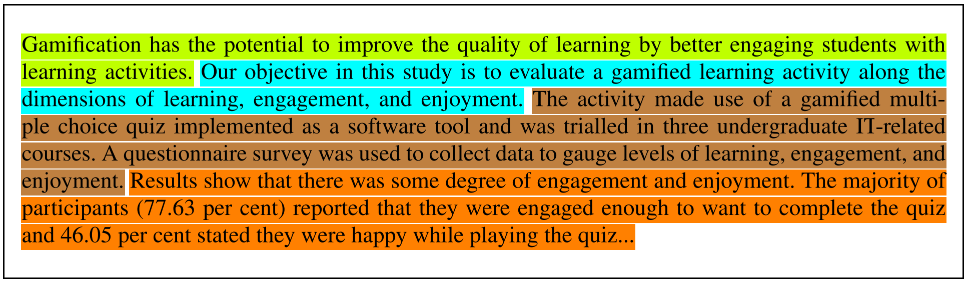
> *#RQ5* Can a model trained with multiple datasets recognise the semantic relatedness of classes from different annotation schemes?
> *#RQ6* Can we derive a consolidated, domain-independent annotation scheme and use that scheme to compile a new dataset to train a domain-independent model?

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 a change of the semantic sentence class. Typically, deep learning is used for abstracts [17, 23, 34, 41, 75, 86] since more training data are presumably available, whereas for full papers, also called *zone identification*, handcrafted features and linear models have been suggested [3, 5, 29, 52]. However, deep learning approaches have also been applied successfully to full papers in related tasks such as argumentation mining [49], scientific document summarisation [1, 18, 25, 33] or n-ary relation extraction [31, 40, 43]. Thus, the potential of deep learning has not been fully exploited yet for sequential sentence classification on full papers, and no unified solution exists for abstracts as well as full papers. This raises the following 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 the classification of 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/TIBHannover/sequential-sentence-classification-extended.



Gamification has the potential to improve the quality of learning by better engaging students with learning activities. Our objective in this study is to evaluate a gamified learning activity along the dimensions of learning, engagement, and enjoyment. The activity made use of a gamified multiple choice quiz implemented as a software tool and was trialled in three undergraduate IT-related courses. A questionnaire survey was used to collect data to gauge levels of learning, engagement, and enjoyment. Results show that there was some degree of engagement and enjoyment. The majority of participants (77.63 per cent) reported that they were engaged enough to want to complete the quiz and 46.05 per cent stated they were happy while playing the quiz...

Fig. 1 An annotated abstract taken from the CSABSTRUCT dataset [17], where the sentences are coloured according to their respective category: *background* (green), *objectives* (blue), *methods* (brown) and *results* (orange) (colour figure online)

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.

This article is an extension of a paper [10] presented during the Joint Conference on Digital Libraries (JCDL'22) [2]. With respect to this previous publication, we provide (a) an extended discussion of related work (Sect. 2); (b) an extended description of the proposed methods, especially the unified deep learning approach (Sect. 3.1); (c) an additional performance comparison between SciBERT, BERT-Base and BERT-Large (Sect. 5.1); (d) a significance test to compare our approach to the previous state-of-the-art results; (e) a qualitative analysis and case study (Sect. 5.4); (f) additional figures and discussion regarding the semi-automatic identification of semantically related classes across several annotation schemes (Sect. 5.5); (g) a comparison of the semi-automatic identification to a fully automatic one (Sect. 5.5); and (h) a discussion of the limitations of our approach and analyses (Sect. 5.6).

The remainder of the paper is organised as follows: Sect. 2 summarises related work on sentence classification in research papers and transfer learning in natural language processing (NLP). Our proposed approaches are presented in Sect. 3. The setup and results of our experimental evaluation are reported in Sects. 4 and 5, while Sect. 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 of information extraction from scientific text, we refer to Brack et al. [11] and Nasar et al. [59].

2.1 Sequential sentence classification in scientific text

Datasets: As depicted in Table 1, there are various annotated benchmark datasets for sentence classification in research papers. These come from several domains, e.g. PubMed-20k [22] consists of biomedical randomised controlled trials, NICTA-PIBOSO [45] comes from evidence-based medicine, Dr. Inventor dataset [29] comes from computer graphics, and the ART/Core Scientific Concepts (CoreSC) dataset [53] comes from chemistry and biochemistry. Most datasets cover only abstracts, while ART/CoreSC and Dr. Inventor cover full papers. Furthermore, each dataset has five to 11 different sentence classes, which are more domain-independent (e.g. Background, Methods, Results, Conclusions) or more domain-specific (e.g. Intervention, Population [45], or Hypothesis, Model, Experiment [53]).

Approaches for Abstracts: Recently, deep learning has been the preferred approach for sentence classification in abstracts [17, 23, 34, 41, 75, 86]. 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 their different implementations of the layers. The approaches use different kinds of word embeddings, e.g. Global Vectors (GloVe) [65], Word2Vec [57] or SciBERT [7] that is BERT [24] pre-trained on scientific text. For sentence encoding, a bidirectional long short-term memory (Bi-LSTM) [38] or a convolutional neural network (CNN) with various pooling strategies is utilised, while Yamada et al. [86] and Shang et al. [75] use the classification token ([CLS]) of BERT or SciBERT. A recurrent neural network (RNN) such as a Bi-LSTM or bidirectional gated recurrent unit (Bi-GRU) [15] is used to enrich sentences with further context. Shang et al. [75] 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) [48] is often used as an output layer

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

Dataset	Domains	# Papers	Text type	Sentence classes
PubMed-20k [22]	Biomedicine	20,000	Abstracts	Background, Objective, Methods, Results, Conclusion
NICTA-PIBOSO [45]	Biomedicine	1000	Abstracts	Background, Intervention,Study, Population, Outcome, Other
CSABSTRUCT [17]	Computer Science	2189	Abstracts	Background, Objective,Method, Result, Other
CS-Abstracts [34]	Computer Science	654	Abstracts	Background, Objective, Methods, Results, Conclusions
Emerald 100k [77]	Management, Engineering, Information Science	103,457	Abstracts	Purpose, Design/methodology and approach, Findings, Originality/value,Social implications,Practical implications, Research limitations and implications
MAZEA [20]	Physics,Engineering, Life and Health Sciences	1335	Abstracts	Background, Gap, Purpose, Method, Result, Conclusion
Dr. Inventor [29]	Computer Graphics	40	Full paper	Background, Challenge, Approach, Outcome, Future Work
ART/CoreSC [53]	Chemistry, Computational Linguistic	225	Full paper	Background, Motivation, Goal, Hypothesis, Object, Model, Method, Experiment, Result, Observation, Conclusion

to capture the interdependence between classes (e.g. results usually follow methods). Yamada et al. [86] 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. [17] 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.

Approaches for Full Papers: For full papers, logistic regression, support vector machines and CRFs with handcrafted features have been proposed [3, 5, 29, 52, 79, 80]. They represent a sentence with various syntactic and linguistic features such as n-grams, part-of-speech tags or citation markers engineered for the respective datasets. Asadi et al. [3] also exploit semantic features obtained from knowledge bases such as Wordnet [28]. 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. [3] suggest fusion techniques within a dynamic window of sentences. However, some approaches [3, 5, 29] exploit the *ground truth label* instead of the predicted label of the preceding sentence ("history feature") during prediction (as confirmed by the authors), which significantly impacts the performance.

Related tasks also classify sentences in full papers with deep learning methods, e.g. for citation intent classification [16, 47], or algorithmic metadata extraction [71] but without exploiting context from surrounding sentences. Comparable to us, Lauscher et al. [49] 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 the task of sequential sentence classification for abstracts as well as full papers has not been proposed and evaluated yet.

2.2 Transfer learning

Transfer learning aims to exploit knowledge from a source task to improve prediction accuracy in a target task. The tasks can have training data from different domains and vary in their objectives. According to Ruder's taxonomy for transfer learning [69], 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 so many applications for transfer learning, we focus on the most relevant cases for sentence classification in scientific texts. For a more comprehensive overview, we refer to [62, 69, 84].

Fine-tuning a pre-trained language model is a popular approach for sequential transfer learning in NLP [13, 24, 37,

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

Approach	Word embedding	Sentence encoding	Context enrichment	Output layer
Dernoncourt and Lee [23]	Character Emb. + GloVe	Bi-LSTM with concatenation	–	CRF
Jin and Szolovits [41]	Bio word2vec	Bi-LSTM with attention pooling	Bi-LSTM	CRF
Cohan et al. [17]	SciBERT	SciBERT-[SEP]	SciBERT-[SEP]	softmax
Gonçalves et al. [34]	GloVe	CNN with max pooling	Bi-GRU	softmax
Yamada et al. [86]	BERT from PubMed	BERT-[CLS]	Bi-LSTM	Semi-Markov CRF
Shang et al. [75]	SciBERT	SciBERT-[CLS]	Bi-LSTM/ attention	CRF

39]. Here, the source task involves learning a language model (or a variant of it) using a sizeable unlabelled text corpus. Then, the model parameters are fine-tuned with labelled data of the target task. Edwards et al. [27] evaluate the importance of domain-specific unlabelled data on pre-training word embeddings for text classification in the general domain (i.e. data such as news, phone conversations, magazines). Pruksachatkun et al. [66] 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 [63] 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.

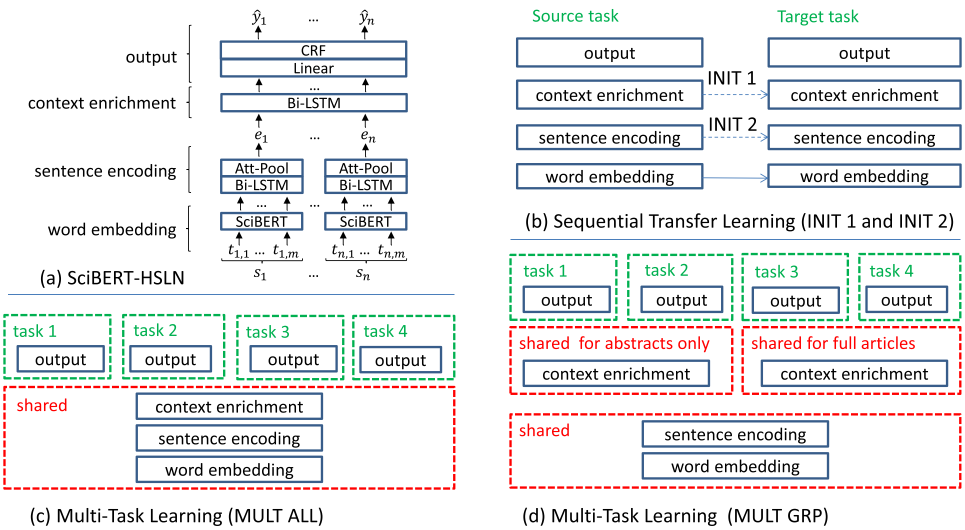
For *sequence tagging*, Yang et al. [87] investigate multi-task learning in the general domain with cross-domain, cross-application and cross-lingual transfer. In particular, target tasks with few labelled data benefit from related tasks. Lee et al. [51] successfully transfer pre-trained parameters from a big dataset to a small dataset in the biological domain. Schulz et al. [73] evaluate multi-task learning for argumentation mining with multiple datasets in the general domain and show that performance improves when training data for the tasks is sparse. For *coreference resolution*, Brack et al. [12] successfully apply sequential transfer learning and use a large dataset from the general domain to improve models for a small dataset in the scientific domain.

For *sentence classification*, Mou et al. [58] 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. They demonstrate that semantically related tasks improve while unrelated tasks degrade the performance of the target tasks. Semwal et al. [74] investigate the extent of task relatedness for product reviews and sentiment classification with sequential transfer learning. Su et al. [78] study multi-task learning for sentiment classification in product reviews from multiple domains. Lauscher et al. [50] evaluate multi-task learning on scientific texts but only on one dataset with different annotation layers. Banerjee et al. [6] apply sequential transfer learning from medicine to computer science for discourse classification, however, only for two domains and on abstracts, whereas Spangher et al. [76] explore this task on news articles with multi-task learning using multiple datasets. Gupta et al. [36] utilise multi-task learning with two scaffold tasks to detect contribution sentences in full papers applied to only one domain and with limited sentence context.

Several approaches have been proposed to *train multiple tasks jointly*: Luan et al. [55] train a model on three tasks (coreference resolution, entity and relation extraction) using one dataset of research papers. Sanh et al. [72] introduce a multi-task model trained on four tasks (mention detection, coreference resolution, entity and relation extraction) with two different datasets. Wei et al. [83] utilise a multi-task model for entity recognition and relation extraction on one dataset in the non-academic domain. Comparable to us, Changpinyo et al. [14] analyse multi-task training with multiple datasets for sequence tagging. *In contrast, we investigate sequential sentence classification across multiple science domains.*

3 Cross-domain multi-task learning for sequential sentence classification

The discussion of related work shows that several approaches and datasets from various scientific domains have been introduced for sequential sentence classification. While transfer learning has been applied to various NLP tasks, it is known that success depends largely on the relatedness of the tasks [58, 62, 69]. However, the field lacks an empirical study on transfer learning between different scientific domains



(a) SciBERT-HSLN

(b) Sequential Transfer Learning (INIT 1 and INIT 2)

(c) Multi-Task Learning (MULT ALL)

(d) Multi-Task Learning (MULT GRP)

Fig. 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 and INIT 2 only the sentence encoding layer; **c** and **d** multi-task learning approaches, i.e. in MULT ALL all possible layers are shared between the tasks, and in MULT GRP, the context enrichment is shared between tasks with the same text type

for sequential sentence classification. Current approaches 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 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 Fig. 2, exploit multiple datasets comprising different text types in the form of abstracts and full papers. The unified approach without transfer learning is described in Sect. 3.1, while Sect. 3.2 introduces the sequential transfer learning and multi-task learning approaches. Finally, in Sect. 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 $(\mathbf{s_1}, \ldots, \mathbf{s_n})$ and the set of dataset specific classes $\mathbb{L}$ (e.g. Background, Methods), the task of *sequential sentence classification* is to predict the corresponding label sequence $(\mathbf{y_1}, \ldots, \mathbf{y_n})$ with $\mathbf{y_i} \in \mathbb{L}$. For this task, we propose a unified deep learning approach as depicted in Fig. 2a, 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 labelling network (HSLN) [41]. In contrast to Jin and Szolovits [41], we utilise SciBERT [7] 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. [17]) and is computationally less expensive than the more recent approaches [75, 86]. Furthermore, their implementation is publicly available. The goal of this study is not to beat state-of-the-art results but rather to provide an empirical study on transfer learning for sequential sentence classification and offer a uniform solution. Our *SciBERT-HSLN* architecture has the following layers:

Word Embedding: Input is a sequence of tokens $(\mathbf{t_{i,1}}, \ldots, \mathbf{t_{i,m}})$ of sentence $\mathbf{s_i}$, while output is a sequence of contextual word embeddings $(\mathbf{w_{i,1}}, \ldots, \mathbf{w_{i,m}})$.

Sentence Encoding: Input $(\mathbf{w_{i,1}}, \ldots, \mathbf{w_{i,m}})$ is transformed via a Bi-LSTM [38] into the hidden token representations $(\mathbf{h_{i,1}}, \ldots, \mathbf{h_{i,m}})$ $(\mathbf{h_{i,t}} \in \mathbb{R}^{d^h})$ which are enriched with contextual information within the sentence. Then, attention pooling [41, 88] with r heads produces a sentence vector $\mathbf{e_i} \in \mathbb{R}^{rd^u}$. An attention head produces a weighted average over the token representations of a sentence. Multiple heads enable to capture several semantics of a sentence. Formally, at first, a token representation $\mathbf{h_{i,t}}$ is transformed via a feed-forward network into a further hidden representation $\mathbf{a_{i,t}}$ with the learned weight matrix $\mathbf{a^{[S]}}$ and bias vector $\mathbf{b^{[S]}}$:

$$\mathbf{a_{i,t}} = FFN(\mathbf{h_{i,t}}) = \tanh(\mathbf{a^{[S]}}\mathbf{h_{i,t}} + \mathbf{b^{[S]}}) \qquad (1)$$

Then, for each attention head k with $1 \leq k \leq r$ the learned token level context vector $\mathbf{u_k} \in \mathbb{R}^{d^u}$ is used to compute importance scores for all token representations which are then normalised by $softmax$:

$$\alpha_{k,i,t} = \frac{\exp(\mathbf{u}_{\mathbf{k}^\intercal}\mathbf{a_{i,t}})}{\sum_{t'} \exp(\mathbf{u}_{\mathbf{k}^\intercal}\mathbf{a_{i,t'}})} \qquad (2)$$

An attention head $\mathbf{e_{k,i}} \in \mathbb{R}^{d^h}$ is computed as a weighted average over the token representations and all heads are concatenated to form the final sentence representation $\mathbf{e_i} \in \mathbb{R}^{rd^h}$:

$$\mathbf{e_{k,i}} = \sum_{t'} \alpha_{k,i,t'} \mathbf{h_{i,t'}} \tag{3}$$

$$\mathbf{e_i} = [\mathbf{e_{1,i}}, \ldots, \mathbf{e_{r,i}}] \tag{4}$$

Context Enrichment: This layer takes as input all sentence representations $(\mathbf{e_1}, \ldots, \mathbf{e_n})$ of the paper and outputs contextualised sentence representations $(\mathbf{c_1}, \ldots, \mathbf{c_n})$ $(\mathbf{c_i} \in \mathbb{R}^{d^h})$ via a Bi-LSTM. Thus, each sentence representation $\mathbf{c_i}$ contains contextual information from surrounding sentences.

Output Layer: This layer transforms sentence representations $(\mathbf{c_1}, \ldots, \mathbf{c_n})$ via a linear transformation to the logits $(\mathbf{l_1}, \ldots, \mathbf{l_n})$ with $\mathbf{l_i} \in \mathbb{R}^{|\mathbb{L}|}$. Each component of vector $\mathbf{l_i}$ contains a score for the corresponding label:

$$\mathbf{l_i} = \mathbf{W^{[O]}} \mathbf{c_i} + \mathbf{b^{[O]}} \tag{5}$$

Finally, the logits serve as input for a CRF [48] that predicts the label sequence $(\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{n})$ $(\hat{\mathbf{y}}_\mathbf{i} \in \mathbb{L})$ with the highest joint probability. A CRF captures linear (one-step) dependencies between the labels (e.g. Methods sentences are usually followed by Methods or Results sentences). Therefore, a CRF learns a transition matrix $\mathbf{T} \in \mathbb{R}^{|\mathbb{L}| \times |\mathbb{L}|}$, where $\mathbf{T}_{l_1, l_2}$ represents the transition score from label l_1 to label l_2, and two vectors $\mathbf{b}, \mathbf{e} \in \mathbb{R}^{|\mathbb{L}|}$, where $\mathbf{b}_l$ and $\mathbf{e}_l$ represent the score of beginning and ending with label l, respectively. The objective is to find the label sequence with the highest conditional joint probability $P(\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{n} | \mathbf{l_1}, \ldots, \mathbf{l_n})$. For this purpose, we define a score function for a label sequence $(\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{n})$, that is, a sum of the scores of the labels and the transition scores:

$$\begin{aligned} &score((\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{n}), (\mathbf{l_1}, \ldots, \mathbf{l_n})) \\ &\quad = \mathbf{b}_{\hat{y}_1} + \sum_{t=1}^{n} \mathbf{l}_{\mathbf{t}, \hat{y}_t} + \sum_{t=1}^{n-1} \mathbf{T}_{\hat{y}_t, \hat{y}_{t+1}} + \mathbf{e}_{\hat{y}_m} \end{aligned} \tag{6}$$

Then, the score is transformed to a probability value with $softmax$:

$$\begin{aligned} &Z(\mathbf{l_1}, \ldots, \mathbf{l_n}) \\ &\quad = \sum_{\mathbf{y_{1'}}, \ldots, \mathbf{y_{n'}}} \exp(score((\mathbf{y_{1'}}, \ldots, \mathbf{y_{n'}}), (\mathbf{l_1}, \ldots, \mathbf{l_n}))) \end{aligned} \tag{7}$$

$$\begin{aligned} &P(\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{n} | \mathbf{l_1}, \ldots, \mathbf{l_n}) \\ &\quad = \frac{\exp(score((\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{n}), (\mathbf{l_1}, \ldots, \mathbf{l_n})))}{Z(\mathbf{l_1}, \ldots, \mathbf{l_n})} \end{aligned} \tag{8}$$

The denominator $Z(.)$ represents a sum of the scores of all possible label sequences for the given logits. The Viterbi algorithm [30] is used to efficiently calculate the sequence with the highest score and the denominator (both with time complexity $O(|\mathbb{L}|^2 \cdot n)$).

During training, the CRF maximises $P(\mathbf{y_1}, \ldots, \mathbf{y_n} | \mathbf{l_1}, \ldots, \mathbf{l_n})$ of the ground truth labels for all m training samples $((\mathbf{x^{(1)}}, \mathbf{y^{(1)}}), \ldots, (\mathbf{x^{(m)}}, \mathbf{y^{(m)}}))$, where $\mathbf{x^{(i)}}$ represents the sentences of paper i and $\mathbf{y^{(i)}}$ the corresponding ground truth label sequence. Thus, the objective is to minimise the following loss function:

$$L = -\frac{1}{m} \sum_{i=1}^{m} \log P(\mathbf{y^{(i)}} | \mathbf{l^{(i)}}) \tag{9}$$

For regularisation, we use dropout after each layer. The SciBERT model is not fine-tuned since it requires training with 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 Fig. 2b, 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 better generalisation by simultaneously training samples in all tasks and sharing parameters of certain layers between the tasks. As depicted in Fig. 2c, 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, sharing the context enrichment layer between both dataset types is not beneficial. Thus, in the *MULT GRP* model, the *context enrichment layers* are only shared between datasets with the same text type. Formally, the objective is to minimise the following loss functions:

$$L_{\text{MULT ALL}} = \sum_{t \in \mathbb{T}^{\mathbb{A}} \cup \mathbb{T}^{\mathbb{F}}} L_t(\Theta^S, \Theta^C, \Theta_t^O) \tag{10}$$

$$\begin{aligned} L_{\text{MULT GRP}} &= \sum_{t \in \mathbb{T}^{\mathbb{A}}} L_t(\Theta^S, \Theta^{C^A}, \Theta_t^O) \\ &\quad + \sum_{t \in \mathbb{T}^{\mathbb{F}}} L_t(\Theta^S, \Theta^{C^F}, \Theta_t^O) \end{aligned} \tag{11}$$

where $\mathbb{T}^{\mathbb{A}}$ and $\mathbb{T}^{\mathbb{F}}$ are the tasks for datasets containing abstracts and full papers, respectively; L_t is the loss function for task t; the parameters Θ^S are for sentence encoding,

Table 3 Characteristics of the benchmark datasets

	PMD	NIC	DRI	ART
Domains	Biomedicine	Biomedicine	Computer graphics	Chemistry, Computational linguistic
Text type	Abstract	Abstract	Full paper	Full paper
# Papers	20,000	1000	40	225
# Sentences	235,892	9771	8777	34,680
∅ # Sentences	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
State of the art	[86] 93.1	[75] 86.8	[5] 72.5	[52] 51.6
Original metric	weighted F1	weighted F1	weighted F1	accuracy

The row "State of the art" depicts the best results for approaches that do not exploit the ground truth label of the preceding sentence during prediction: for PMD [86], for NIC [75], for DRI [5] and for ART [52]

Θ^C, Θ^{C^A} and Θ^{C^F} for context enrichment, and Θ_t^O 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. The loss functions are defined as:

$$L_{\text{MULT ALL SHO}} = \sum_{t \in \mathbb{T}^{\mathbb{A}} \cup \mathbb{T}^{\mathbb{F}}} L_t(\Theta^S, \Theta^C, \Theta^O) \tag{12}$$

$$L_{\text{MULT GRP SHO}} = \sum_{t \in \mathbb{T}^{\mathbb{A}}} L_t(\Theta^S, \Theta^{C^A}, \Theta^{O^A}) + \sum_{t \in \mathbb{T}^{\mathbb{F}}} L_t(\Theta^S, \Theta^{C^F}, \Theta^{O^F}) \tag{13}$$

3.3 Semantic relatedness of classes

Datasets for sentence classification have different domain-specific annotation schemes, that is, different sets of predefined 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 $\mathbb{T}$ be the set of all tasks, $\mathbb{L}$ the set of all classes in all tasks, $m_t(s)$ the label of sentence s predicted by the model for task t and $\mathbb{S}^l$ the set of sentences with the ground truth label l. For each class $l \in \mathbb{L}$, the corresponding semantic vector $\mathbf{v_l} \in \mathbb{R}^{|\mathbb{L}|}$ is defined as:

$$\mathbf{v}_{\mathbf{l},l'} = \frac{\sum_{t \in \mathbb{T}, s \in \mathbb{S}^l} 1(m_t(s) = l')}{|\mathbb{S}^l|} \tag{14}$$

where $\mathbf{v}_{\mathbf{l},l'} \in \mathbb{R}$ is the component of the vector $\mathbf{v_l}$ for class $l' \in \mathbb{L}$ 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 Fig. 4 as an example).

Now, we define the semantic relatedness of two classes $k, l \in \mathbb{L}$ using cosine similarity:

$$\text{semantic\_relatedness}(k, l) = \cos(\mathbf{v_k}, \mathbf{v_l}) = \frac{\mathbf{v_k}^\top \cdot \mathbf{v_l}}{||\mathbf{v_k}|| \cdot ||\mathbf{v_l}||} \tag{15}$$

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) [22], NICTA-PIBOSO (NIC) [45], ART [53] and Dr. Inventor (DRI) [29]. The four datasets are publicly available and provide a good mix to investigate the transferability of sentence classification methods: 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 relatively small datasets, while PMD and ART are about 20 and 3 times larger, respectively; ART has a more fine-granular 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, the 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 [64]. The Adaptive Moment Estimation (ADAM) optimiser [46] is used for training, with 0.01 weight decay and an exponential learning rate decay of 0.9 after each epoch. Since the computational cost for the attention layers in BERT is quadratic in sentence length [81], sentences longer than 128 tokens are truncated to speed up training. To reproduce the results of the original HSLN architecture, we tuned SciBERT-HSLN for PMD and NIC with hyperparameters as proposed in other studies [24, 41]. The following parameters performed best on the validation sets of PMD and NIC: learning rate was set to 3e-5 and dropout rate was 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 our experiments.

For each dataset, we grouped papers into mini-batches without splitting them if the mini-batch did 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 [72]. After a mini-batch, only task-related parameters are updated, i.e. the associated output layer and all the layers below.

4.3 Evaluation

To be consistent with previous results and due to non-determinism in deep neural networks [67], we repeated the experiments and averaged the results. According to Cohan et al. [17], we performed three random restarts for PMD and NIC and used the same train/validation/test sets. For DRI and ART, we performed tenfold and ninefold cross-validation, respectively, as in the original papers [29, 52]. 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.

We compare our results only with approaches that do not exploit the preceding sentence's *ground truth labels* as a feature *during prediction* (see Sect. 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 an SVM classifier already yields an accuracy of 77.7 for DRI and 55.5 for ART (compare also results for the "history" feature in [5]). Best reported results using ground truth labels as input features have an accuracy of 84.15 for DRI and 65.75 for ART [3]. 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 [24] (referred to as SciBERT-[CLS]), (2) original implementation of Jin and Szolovits [41] and (3) the SciBERT-based approach of Cohan et al. [17].

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 several state-of-the-art methods are depicted in Table 4. The results are discussed in the following three subsections regarding the unified approach without transfer learning (Sect. 5.1), with sequential transfer learning (Sect. 5.2) and multi-task learning (Sect. 5.3). Section 5.5 analyses the semantic relatedness of classes for the four annotation schemes.

Table 4 Experimental results for the proposed approaches (in per cent): our SciBERT-HSLN model without transfer learning, parameter initialisation (INIT) and multi-task learning (MULT ALL and MULT GRP)

	PMD [F1]	NIC [F1]	DRI [F1]	ART [Acc]	∅
Previous state of the art	[86] 93.1	[75] 86.8	[5] 72.5	[52] 51.6	76.0
SciBERT-[CLS]	89.6	78.4	69.5	51.5	72.3
Jin and Szolovits [41] (HSLN)	92.6	84.7	75.3	49.3	75.5
Cohan et al. [17]	92.9	84.8	74.3	54.3	76.6
SciBERT-HSLN	92.9	84.9	*78.0*	*58.0*	*78.5*
BERT-Base-HSLN	92.6	82.9	*74.0*	*55.0*	76.1
BERT-Large-HSLN	92.4	83.3	*74.4*	*55.1*	76.3
RoBERTa-Base-HSLN	92.4	83.2	69.4	52.5	74.4
RoBERTa-Large-HSLN	92.6	81.5	67.1	49.5	72.7
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***

Previous state of the art (see Table 3), SciBERT-[CLS], original HSLN approach of Jin and Szolovits [41] and the approach of Cohan et al. [17] 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 if the transfer method improves SciBERT-HSLN, underline the best overall result

5.1 Unified approach without transfer learning

For the full paper datasets DRI and ART, our SciBERT-HSLN model significantly outperforms the previously reported best results and the deep learning baselines from Jin and Szolovits [41], Cohan et al. [17] and SciBERT-[CLS]. The previous state-of-the-art approaches for DRI and ART [5, 52] 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 [41] uses shallow word embeddings pre-trained on biomedical texts. Thus, incorporating SciBERT's contextual word embeddings into HSLN helps improve performance for the DRI and ART datasets. The approach of

Cohan et al. [17] can process only about ten sentences at once since SciBERT supports sequences of up to 512 tokens. 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 to the current state of the art [86], while they are slightly lower for NIC [75]. In addition, we test the HSLN model with BERT [24] and RoBERTa [54] to evaluate the influence of the embedding model. The base and large versions of BERT and RoBERTa, respectively, show similar performance. In three out of four cases, BERT-Large has a non-significant higher performance of up to 0.4% than BERT-Base. Comparing the RoBERTa models, it is visible that the base model performs better in almost all cases and has up to 3% higher performance. Reasons for that could be (a) fluctuations in the training data causing a decrease in performance or (b) the larger model leading to more generalisation and thus, decreased accuracy on scientific text. Comparing the best values between BERT and RoBERTa, BERT performs better on the NIC, DRI and ART datasets, being up to 5% more accurate. One possible reason could be the different training approaches of the models. RoBERTa is a modification of BERT, which uses different hyperparameters, training data, prediction objectives and masking patterns. These changes could cause the performance difference of RoBERTa. Using the BERT or RoBERTa models consistently performs worse than SciBERT as they are not adapted to scientific text. Overall, our proposed approach SciBERT-HSLN 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).

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 in this setting, 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 at 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 larger 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).

Significance Test: We perform significance tests for our best new models (MULT ALL and MULT GRP) to compare them with the previous state-of-the-art models (PSOTA). We use McNemar's test [56] which is commonly used for model comparison [21, 26]. To apply the test, the predictions of the models are needed at the sentence level. Therefore, we asked the authors of the respective publications to provide the source code in order to generate the predictions. We have received the code for the previous state-of-the-art models for NIC (PSOTA-NIC) and ART (PSOTA-ART) datasets. In our tests, MULT ALL and MULT GRP models are compared with each other, with PSOTA-NIC and PSOTA-ART, and with the different SciBERT-HSLN models. Additionally, the significance between SciBERT-HSLN models and the existing PSOTAs has been checked. The outcomes of the significance test are shown in Table 5. Assuming a significance level of $p = 0.05$, there is no significant difference between the MULT ALL and MULT GRP model. The difference to the various SciBERT-HSLN models is significant in three out of four cases if compared to MULT ALL. The insignifi-

Table 5 Significance test of the different models using McNemar's test

First model	Better Performance	Second model	*p*-value
MULT ALL	<	MULT GRP	0.096
MULT ALL	>	SciBERT-HSLN PubMed	*0.005
MULT ALL	>	SciBERT-HSLN Nicta-piboso	*0.000
MULT ALL	>	SciBERT-HSLN DRI	*0.001
MULT ALL	<	SciBERT-HSLN ART	0.834
MULT ALL	<	PSOTA-NIC	*0.003
MULT ALL	>	PSOTA-ART	*0.000
MULT GRP	>	SciBERT-HSLN PubMed	0.629
MULT GRP	>	SciBERT-HSLN Nicta-piboso	*0.011
MULT GRP	>	SciBERT-HSLN DRI	*0.000
MULT GRP	>	SciBERT-HSLN ART	0.315
MULT GRP	<	PSOTA-NIC	0.386
MULT GRP	>	PSOTA-ART	*0.000
SciBERT-HSLN Nicta-piboso	<	PSOTA-NIC	0.125
SciBERT-HSLN ART	>	PSOTA-ART	*0.000

"PSOTA" is short for "Previous State-of-the-art" and "NIC" for "Nicta-piboso". The column "Better Performance" describes whether the first model (>) or the second model (<) has a higher performance. P-values marked with a star signal that a significant difference exists

Table 6 Experimental results (in per cent) for μPMD, NIC, DRI and μART with our SciBERT-HSLN model and our proposed multi-task learning approaches

	μPMD [F1]	NIC [F1]	DRI [F1]	μART [Acc]	∅
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**

Bold signifies an improvement over SciBERT-HSLN

cant case is between MULT ALL and SciBERT-HSLN ART, where SciBERT-HSLN has a slightly higher performance. For MULT GRP, the difference to SciBERT-HSLN is significant in two out of four cases. Here, it is noticeable that for the ART dataset, the higher performance of MULT GRP is not relevant. In comparison with PSOTA-NIC, MULT ALL performs significantly worse. For both MULT GRP and SciBERT-HSLN, there is no significant difference to PSOTA-NIC. In all cases, the performance difference to PSOTA-ART is significantly better since the p-value is always 0.0. This shows the relevancy of our approaches.

Overall, the results show that the proposed models achieve a significant improvement with regard to the datasets. However, especially for the PMD and the NIC dataset the marginal $F1$ score difference of less than one is not significant.

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 also has 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, referred to as μPMD and μART, with a comparable size with NIC and DRI. Within each fold, we truncated the training data 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 6, all models benefit from the other datasets, whereas the MULT GRP model again performs best.

The results indicate that models for small datasets benefit from multi-task learning independent of differences 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).

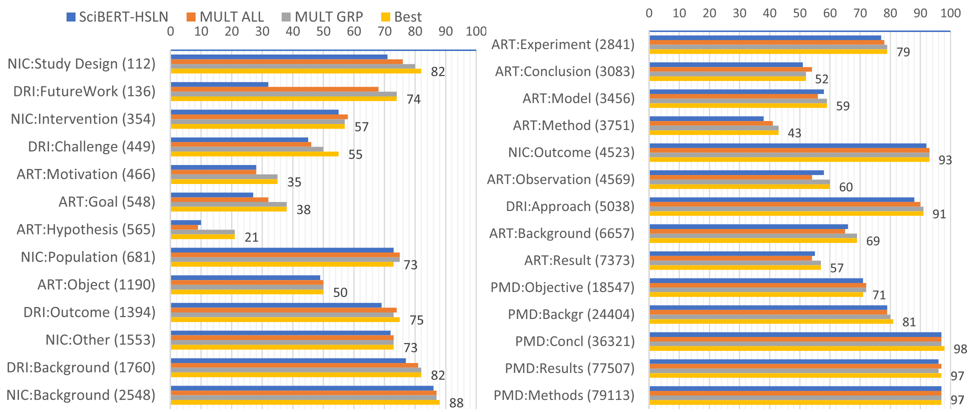
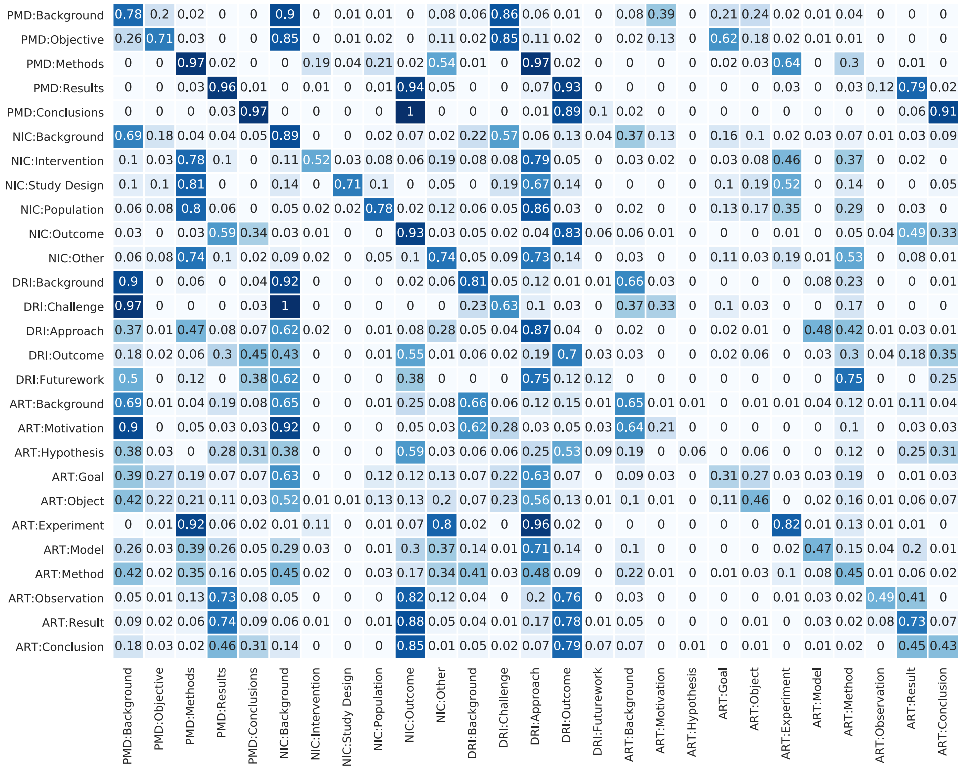


Fig. 3 $F1$ scores (in per cent) per class for the datasets PMD, NIC, DRI and ART for the approaches SciBERT-HSLN, MULT ALL, MULT GRP and the best combination for the respective dataset. Numbers at the bars depict the $F1$ scores of the best classifiers and in brackets the number of examples for the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

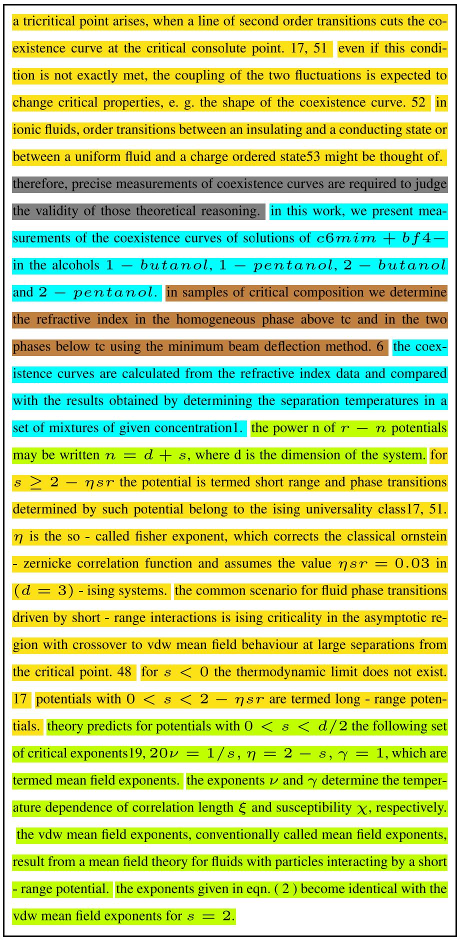


	PMD:Background	PMD:Objective	PMD:Methods	PMD:Results	PMD:Conclusions	NIC:Background	NIC:Intervention	NIC:Study Design	NIC:Population	NIC:Outcome	NIC:Other	DRI:Background	DRI:Challenge	DRI:Approach	DRI:Outcome	DRI:Futurework	ART:Background	ART:Motivation	ART:Hypothesis	ART:Goal	ART:Object	ART:Experiment	ART:Model	ART:Method	ART:Observation	ART:Result	ART:Conclusion
PMD:Background	0.78	0.2	0.02	0	0	0.9	0	0.01	0.01	0	0.08	0.06	0.86	0.06	0.01	0	0.08	0.39	0	0.21	0.24	0.02	0.01	0.04	0	0	0
PMD:Objective	0.26	0.71	0.03	0	0	0.85	0	0.01	0.02	0	0.11	0.02	0.85	0.11	0.02	0	0.02	0.13	0	0.62	0.18	0.02	0.01	0.01	0	0	0
PMD:Methods	0	0	0.97	0.02	0	0	0.19	0.04	0.21	0.02	0.54	0.01	0	0.97	0.02	0	0	0	0	0.02	0.03	0.64	0	0.3	0	0.01	0
PMD:Results	0	0	0.03	0.96	0.01	0	0.01	0	0.01	0.94	0.05	0	0	0.07	0.93	0	0	0	0	0	0	0.03	0	0.03	0.12	0.79	0.02
PMD:Conclusions	0	0	0	0.03	0.97	0	0	0	0	1	0	0	0	0.01	0.89	0.1	0.02	0	0	0	0	0	0	0	0	0.06	0.91
NIC:Background	0.69	0.18	0.04	0.04	0.05	0.89	0	0	0.02	0.07	0.02	0.22	0.57	0.06	0.13	0.04	0.37	0.13	0	0.16	0.1	0.02	0.03	0.07	0.01	0.03	0.09
NIC:Intervention	0.1	0.03	0.78	0.1	0	0.11	0.52	0.03	0.08	0.06	0.19	0.08	0.08	0.79	0.05	0	0.03	0.02	0	0.03	0.08	0.46	0	0.37	0	0.02	0
NIC:Study Design	0.1	0.1	0.81	0	0	0.14	0	0.71	0.1	0	0.05	0	0.19	0.67	0.14	0	0	0	0	0.1	0.19	0.52	0	0.14	0	0	0.05
NIC:Population	0.06	0.08	0.8	0.06	0	0.05	0.02	0.02	0.78	0.02	0.12	0.06	0.05	0.86	0.03	0	0.02	0	0	0.13	0.17	0.35	0	0.29	0	0.03	0
NIC:Outcome	0.03	0	0.03	0.59	0.34	0.03	0.01	0	0	0.93	0.03	0.05	0.02	0.04	0.83	0.06	0.06	0.01	0	0	0	0.01	0	0.05	0.04	0.49	0.33
NIC:Other	0.06	0.08	0.74	0.1	0.02	0.09	0.02	0	0.05	0.1	0.74	0.05	0.09	0.73	0.14	0	0.03	0	0	0.11	0.03	0.19	0.01	0.53	0	0.08	0.01
DRI:Background	0.9	0	0.06	0	0.04	0.92	0	0	0	0.02	0.06	0.81	0.05	0.12	0.01	0.01	0.66	0.03	0	0	0	0	0.08	0.23	0	0	0.01
DRI:Challenge	0.97	0	0	0	0.03	1	0	0	0	0	0	0.23	0.63	0.1	0.03	0	0.37	0.33	0	0.1	0.03	0	0	0.17	0	0	0
DRI:Approach	0.37	0.01	0.47	0.08	0.07	0.62	0.02	0	0.01	0.08	0.28	0.05	0.04	0.87	0.04	0	0.02	0	0	0.02	0.01	0	0.48	0.42	0.01	0.03	0.01
DRI:Outcome	0.18	0.02	0.06	0.3	0.45	0.43	0	0	0.01	0.55	0.01	0.06	0.02	0.19	0.7	0.03	0.03	0	0	0.02	0.06	0	0.03	0.3	0.04	0.18	0.35
DRI:Futurework	0.5	0	0.12	0	0.38	0.62	0	0	0	0.38	0	0	0	0.75	0.12	0.12	0	0	0	0	0	0	0	0.75	0	0	0.25
ART:Background	0.69	0.01	0.04	0.19	0.08	0.65	0	0	0.01	0.25	0.08	0.66	0.06	0.12	0.15	0.01	0.65	0.01	0.01	0	0.01	0.01	0.04	0.12	0.01	0.11	0.04
ART:Motivation	0.9	0	0.05	0.03	0.03	0.92	0	0	0	0.05	0.03	0.62	0.28	0.03	0.05	0.03	0.64	0.21	0	0	0	0	0	0.1	0	0.03	0.03
ART:Hypothesis	0.38	0.03	0	0.28	0.31	0.38	0	0	0	0.59	0.03	0.06	0.06	0.25	0.53	0.09	0.19	0	0.06	0	0.06	0	0	0.12	0	0.25	0.31
ART:Goal	0.39	0.27	0.19	0.07	0.07	0.63	0	0	0.12	0.12	0.13	0.07	0.22	0.63	0.07	0	0.09	0.03	0	0.31	0.27	0.03	0.03	0.19	0	0.01	0.03
ART:Object	0.42	0.22	0.21	0.11	0.03	0.52	0.01	0.01	0.13	0.13	0.2	0.07	0.23	0.56	0.13	0.01	0.1	0.01	0	0.11	0.46	0	0.02	0.16	0.01	0.06	0.07
ART:Experiment	0	0.01	0.92	0.06	0.02	0.01	0.11	0	0.01	0.07	0.8	0.02	0	0.96	0.02	0	0	0	0	0	0	0.82	0.01	0.13	0.01	0.01	0
ART:Model	0.26	0.03	0.39	0.26	0.05	0.29	0.03	0	0.01	0.3	0.37	0.14	0.01	0.71	0.14	0	0.1	0	0	0	0	0.02	0.47	0.15	0.04	0.2	0.01
ART:Method	0.42	0.02	0.35	0.16	0.05	0.45	0.02	0	0.03	0.17	0.34	0.41	0.03	0.48	0.09	0	0.22	0.01	0	0.01	0.03	0.1	0.08	0.45	0.01	0.06	0.02
ART:Observation	0.05	0.01	0.13	0.73	0.08	0.05	0	0	0	0.82	0.12	0.04	0	0.2	0.76	0	0.03	0	0	0	0	0.01	0.03	0.02	0.49	0.41	0
ART:Result	0.09	0.02	0.06	0.74	0.09	0.06	0.01	0	0.01	0.88	0.05	0.04	0.01	0.17	0.78	0.01	0.05	0	0	0	0.01	0	0.03	0.02	0.08	0.73	0.07
ART:Conclusion	0.18	0.03	0.02	0.46	0.31	0.14	0	0	0	0.85	0.01	0.05	0.02	0.07	0.79	0.07	0.07	0	0.01	0	0.01	0	0.01	0.02	0	0.45	0.43

Fig. 4 Each row represents a semantic vector representation as described in Sect. 3.3 for a class computed with the *MULT ALL* classifier

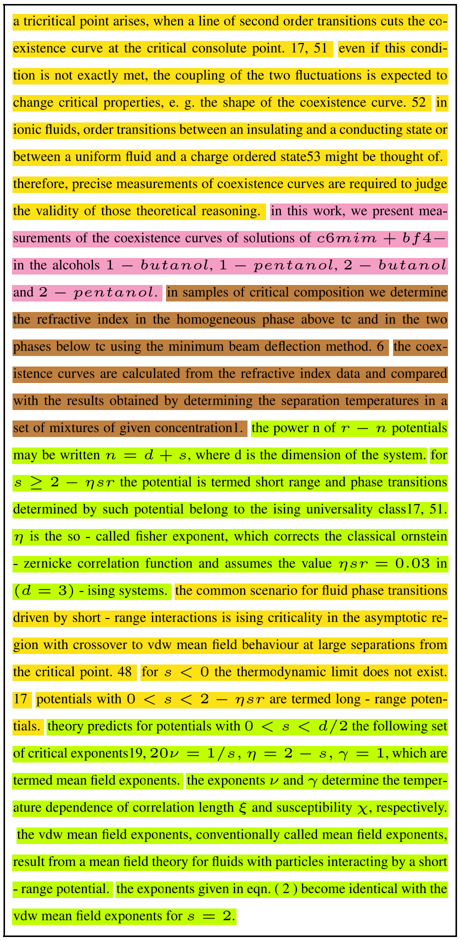
5.4 Qualitative results and case study

To give better suggestions for future work, we complement the above quantitative results with a qualitative evaluation. For this purpose, we use two example texts for each, the *MULT ALL* and the *MULT GRP* model as case studies. We selected texts by their label diversity, choosing one which contains multiple different labels and one with only few assigned labels. Figures 5 and 7 show the annotated ground truth labels for the text examples. If one compares the two texts, it becomes apparent that the sentences in the text of Fig. 5 have many different classes, whereas the sentences in Fig. 7 have little variation in terms of classes. We have chosen such different examples to investigate how the models behave differently with low and high class variation. In contrast, Figs. 6 and 8 show the predicted labels from the *MULT ALL* model. The difference between Figs. 5 and 6 demonstrates that the model's predictions are unstable if text with many different classes is used. By comparing Figs. 7 and 8, it can be seen that the model has problems distinguishing between *background* and *outcome*. A possible explanation is that *background* and *outcome* often have a similar writing style, which means that it is not possible to differentiate it exactly and it tends to be estimated as *outcome*. For example "*this suggests*" is used inside a sentence labelled as *background*, a phrasing which might also be used to discuss an *outcome*. Figures 10 and 9 depict the results for the *MULT GRP model*. We see that *MULT GRP* has the capacity to provide more fine-grained predictions, even when the sentence classes vary. Nevertheless, misclassifications do occur. One



a tricritical point arises, when a line of second order transitions cuts the coexistence curve at the critical consolute point. 17, 51 even if this condition is not exactly met, the coupling of the two fluctuations is expected to change critical properties, e. g. the shape of the coexistence curve. 52 in ionic fluids, order transitions between an insulating and a conducting state or between a uniform fluid and a charge ordered state53 might be thought of. therefore, precise measurements of coexistence curves are required to judge the validity of those theoretical reasoning. in this work, we present measurements of the coexistence curves of solutions of $c6mim + bf4-$ in the alcohols $1 - butanol, 1 - pentanol, 2 - butanol$ and $2 - pentanol.$ in samples of critical composition we determine the refractive index in the homogeneous phase above tc and in the two phases below tc using the minimum beam deflection method. 6 the coexistence curves are calculated from the refractive index data and compared with the results obtained by determining the separation temperatures in a set of mixtures of given concentration1. the power n of $r - n$ potentials may be written $n = d + s$, where d is the dimension of the system. for $s \geq 2 - \eta sr$ the potential is termed short range and phase transitions determined by such potential belong to the ising universality class17, 51. η is the so - called fisher exponent, which corrects the classical ornstein - zernicke correlation function and assumes the value $\eta sr = 0.03$ in $(d = 3)$ - ising systems. the common scenario for fluid phase transitions driven by short - range interactions is ising criticality in the asymptotic region with crossover to vdw mean field behaviour at large separations from the critical point. 48 for $s < 0$ the thermodynamic limit does not exist. 17 potentials with $0 < s < 2 - \eta sr$ are termed long - range potentials. theory predicts for potentials with $0 < s < d/2$ the following set of critical exponents19, 20$\nu = 1/s$, $\eta = 2 - s$, $\gamma = 1$, which are termed mean field exponents. the exponents ν and γ determine the temperature dependence of correlation length ξ and susceptibility χ, respectively. the vdw mean field exponents, conventionally called mean field exponents, result from a mean field theory for fluids with particles interacting by a short - range potential. the exponents given in eqn. (2) become identical with the vdw mean field exponents for $s = 2$.

Fig. 5 True labels with background (yellow), motivation (grey), goal (blue), method (brown), model (green) (colour figure online)



a tricritical point arises, when a line of second order transitions cuts the coexistence curve at the critical consolute point. 17, 51 even if this condition is not exactly met, the coupling of the two fluctuations is expected to change critical properties, e. g. the shape of the coexistence curve. 52 in ionic fluids, order transitions between an insulating and a conducting state or between a uniform fluid and a charge ordered state53 might be thought of. therefore, precise measurements of coexistence curves are required to judge the validity of those theoretical reasoning. in this work, we present measurements of the coexistence curves of solutions of $c6mim + bf4-$ in the alcohols $1 - butanol, 1 - pentanol, 2 - butanol$ and $2 - pentanol.$ in samples of critical composition we determine the refractive index in the homogeneous phase above tc and in the two phases below tc using the minimum beam deflection method. 6 the coexistence curves are calculated from the refractive index data and compared with the results obtained by determining the separation temperatures in a set of mixtures of given concentration1. the power n of $r - n$ potentials may be written $n = d + s$, where d is the dimension of the system. for $s \geq 2 - \eta sr$ the potential is termed short range and phase transitions determined by such potential belong to the ising universality class17, 51. η is the so - called fisher exponent, which corrects the classical ornstein - zernicke correlation function and assumes the value $\eta sr = 0.03$ in $(d = 3)$ - ising systems. the common scenario for fluid phase transitions driven by short - range interactions is ising criticality in the asymptotic region with crossover to vdw mean field behaviour at large separations from the critical point. 48 for $s < 0$ the thermodynamic limit does not exist. 17 potentials with $0 < s < 2 - \eta sr$ are termed long - range potentials. theory predicts for potentials with $0 < s < d/2$ the following set of critical exponents19, 20$\nu = 1/s$, $\eta = 2 - s$, $\gamma = 1$, which are termed mean field exponents. the exponents ν and γ determine the temperature dependence of correlation length ξ and susceptibility χ, respectively. the vdw mean field exponents, conventionally called mean field exponents, result from a mean field theory for fluids with particles interacting by a short - range potential. the exponents given in eqn. (2) become identical with the vdw mean field exponents for $s = 2$.

Fig. 6 Predicted labels with background (yellow), object (purple), method (brown), model (green) of the model *MULT ALL* (colour figure online)

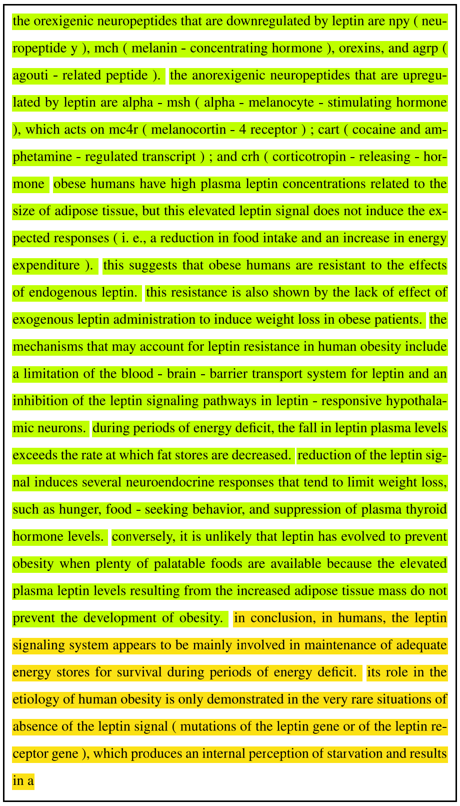
possible explanation is that *GRP ALL* is trained using all datasets with one shared layer. Accordingly, more variations might have been seen, leading to more adaption. This model, however, shows weaker results when class variation is *low* (see Fig. 9). This can be explained by the fact that due to the many texts seen in the training, it is no longer possible to pay close attention to small differences if they are very similar to other classes. This could also explain the misclassifications of the previous example.

5.5 Semantic relatedness of classes across annotation schemes

In this section, we first evaluate our proposed approach for the semi-automatic 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 downstream application, this multi-domain dataset with a generic set of classes could help to structure research papers in a domain-independent manner, supporting, for instance, the development of academic search engines. As a last step, we compare the semi-automatic approach to a fully automatic approach with the *k*-means algorithm.

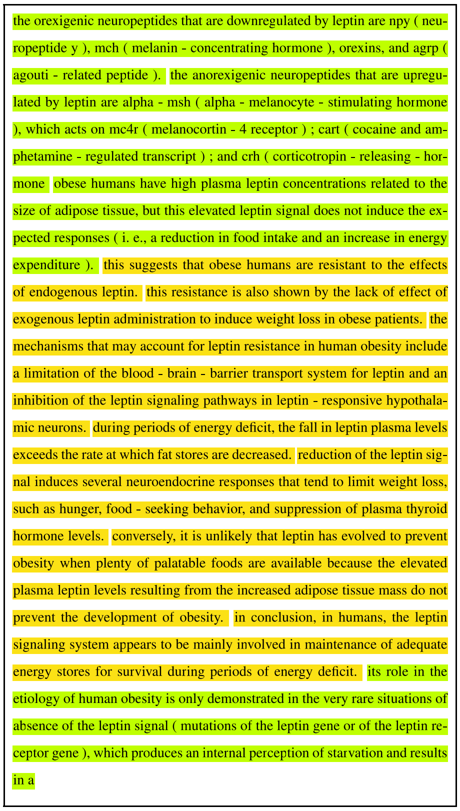
Analysis of Semantic Relatedness of Classes: Based on the annotation guidelines of the investigated datasets PMD [22], NIC [45], DRI [29] and ART [53], we identified six core clusters of semantically related classes, which are depicted in Fig. 11. The identification process of the clusters followed the 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 4 shows the set of semantic vectors for all classes present in the datasets, computed with the MULT ALL model, exemplarily. Already in the matrix representation, it



the orexigenic neuropeptides that are downregulated by leptin are npy (neuropeptide y), mch (melanin - concentrating hormone), orexins, and agrp (agouti - related peptide). the anorexigenic neuropeptides that are upregulated by leptin are alpha - msh (alpha - melanocyte - stimulating hormone), which acts on mc4r (melanocortin - 4 receptor) ; cart (cocaine and amphetamine - regulated transcript) ; and crh (corticotropin - releasing - hormone obese humans have high plasma leptin concentrations related to the size of adipose tissue, but this elevated leptin signal does not induce the expected responses (i. e., a reduction in food intake and an increase in energy expenditure). this suggests that obese humans are resistant to the effects of endogenous leptin. this resistance is also shown by the lack of effect of exogenous leptin administration to induce weight loss in obese patients. the mechanisms that may account for leptin resistance in human obesity include a limitation of the blood - brain - barrier transport system for leptin and an inhibition of the leptin signaling pathways in leptin - responsive hypothalamic neurons. during periods of energy deficit, the fall in leptin plasma levels exceeds the rate at which fat stores are decreased. reduction of the leptin signal induces several neuroendocrine responses that tend to limit weight loss, such as hunger, food - seeking behavior, and suppression of plasma thyroid hormone levels. conversely, it is unlikely that leptin has evolved to prevent obesity when plenty of palatable foods are available because the elevated plasma leptin levels resulting from the increased adipose tissue mass do not prevent the development of obesity. in conclusion, in humans, the leptin signaling system appears to be mainly involved in maintenance of adequate energy stores for survival during periods of energy deficit. its role in the etiology of human obesity is only demonstrated in the very rare situations of absence of the leptin signal (mutations of the leptin gene or of the leptin receptor gene), which produces an internal perception of starvation and results in a

Fig. 7 True labels with background (green), outcome (yellow) (colour figure online)



the orexigenic neuropeptides that are downregulated by leptin are npy (neuropeptide y), mch (melanin - concentrating hormone), orexins, and agrp (agouti - related peptide). the anorexigenic neuropeptides that are upregulated by leptin are alpha - msh (alpha - melanocyte - stimulating hormone), which acts on mc4r (melanocortin - 4 receptor) ; cart (cocaine and amphetamine - regulated transcript) ; and crh (corticotropin - releasing - hormone obese humans have high plasma leptin concentrations related to the size of adipose tissue, but this elevated leptin signal does not induce the expected responses (i. e., a reduction in food intake and an increase in energy expenditure). this suggests that obese humans are resistant to the effects of endogenous leptin. this resistance is also shown by the lack of effect of exogenous leptin administration to induce weight loss in obese patients. the mechanisms that may account for leptin resistance in human obesity include a limitation of the blood - brain - barrier transport system for leptin and an inhibition of the leptin signaling pathways in leptin - responsive hypothalamic neurons. during periods of energy deficit, the fall in leptin plasma levels exceeds the rate at which fat stores are decreased. reduction of the leptin signal induces several neuroendocrine responses that tend to limit weight loss, such as hunger, food - seeking behavior, and suppression of plasma thyroid hormone levels. conversely, it is unlikely that leptin has evolved to prevent obesity when plenty of palatable foods are available because the elevated plasma leptin levels resulting from the increased adipose tissue mass do not prevent the development of obesity. in conclusion, in humans, the leptin signaling system appears to be mainly involved in maintenance of adequate energy stores for survival during periods of energy deficit. its role in the etiology of human obesity is only demonstrated in the very rare situations of absence of the leptin signal (mutations of the leptin gene or of the leptin receptor gene), which produces an internal perception of starvation and results in a

Fig. 8 Predicted labels with background (green), outcome (yellow) of the model *MULT ALL* (colour figure online)

can be observed that some semantic vectors look similar, e.g. PMD:Background and DRI:Background.

For an easier-to-inspect representation, we computed the semantic vectors for SciBERT-HSLN, MULT GRP and MULT ALL, and projected them onto a 2D space using principal component analysis (PCA) [42]. The resulting 2D representations are shown in Fig. 11. The results for all classifiers enable the identification of semantically related classes. For instance, already the results for the SciBERT-HSLN classifier (see Fig. 11a) yield a rather clear *Results* and *Conclusions* cluster. From all the proposed models, the MULT ALL model creates the most meaningful clusters. Except *Problem*, all clusters for semantically related classes are well identifiable in Fig. 11c. Although MULT GRP performs best, the clusters are not consistent in Fig. 11b. The semantic vector for ART:Hypothesis is an outlier in the *Problem* cluster in Fig. 11c, because ART:Hypothesis is confused mostly with ART:Conclusion and ART:Result (see Fig. 4) and has also a very low $F1$ score (see Fig. 3).

Table 7 Silhouette scores per cluster and overall computed for the semantic vectors with the SciBERT-HSLN, MULT GRP and MULT ALL classifiers

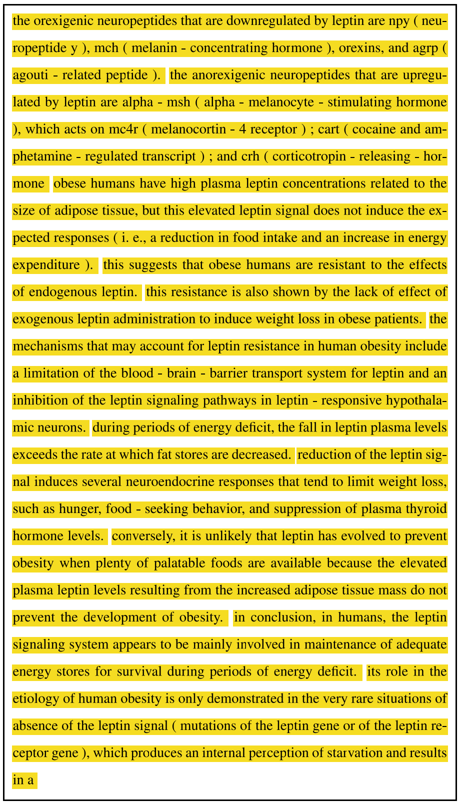
	SciBERT-HSLN	MULT GRP	MULT ALL
Background	0.45	0.18	**0.48**
Problem	−0.27	**−0.04**	−0.29
Methods	0.19	−0.03	**0.31**
Results	−0.38	0.01	**0.32**
Conclusions	**0.92**	−0.49	0.02
Future Work	0.00	0.00	0.00
Overall	0.10	−0.02	**0.20**

Bold depicts the best overall result

To quantify the consistency of the clusters provided by the different classifiers, we calculate Silhouette coefficients [68] for each cluster. Let $\mathbf{v} \in C_l$ be a semantic vector (see Eq. 14) in cluster C_l. First, we define $a(\mathbf{v})$ as the mean distance between $\mathbf{v}$ and all other data points in the same cluster, and $b(\mathbf{v})$ as the mean distance of $\mathbf{v}$ to the nearest other cluster as follows:

$$a(\mathbf{v}) = \frac{1}{|C_l| - 1} \sum_{\mathbf{k} \in C_l, \mathbf{k} \neq \mathbf{v}} d(\mathbf{v}, \mathbf{k}) \tag{16}$$

$$b(\mathbf{v}) = \min_{l' \neq l} \left\{ \frac{1}{|C_{l'}|} \sum_{\mathbf{k} \in C_{l'}} d(\mathbf{v}, \mathbf{k}) \right\} \tag{17}$$



the orexigenic neuropeptides that are downregulated by leptin are npy (neuropeptide y), mch (melanin - concentrating hormone), orexins, and agrp (agouti - related peptide). the anorexigenic neuropeptides that are upregulated by leptin are alpha - msh (alpha - melanocyte - stimulating hormone), which acts on mc4r (melanocortin - 4 receptor) ; cart (cocaine and amphetamine - regulated transcript) ; and crh (corticotropin - releasing - hormone obese humans have high plasma leptin concentrations related to the size of adipose tissue, but this elevated leptin signal does not induce the expected responses (i. e., a reduction in food intake and an increase in energy expenditure). this suggests that obese humans are resistant to the effects of endogenous leptin. this resistance is also shown by the lack of effect of exogenous leptin administration to induce weight loss in obese patients. the mechanisms that may account for leptin resistance in human obesity include a limitation of the blood - brain - barrier transport system for leptin and an inhibition of the leptin signaling pathways in leptin - responsive hypothalamic neurons. during periods of energy deficit, the fall in leptin plasma levels exceeds the rate at which fat stores are decreased. reduction of the leptin signal induces several neuroendocrine responses that tend to limit weight loss, such as hunger, food - seeking behavior, and suppression of plasma thyroid hormone levels. conversely, it is unlikely that leptin has evolved to prevent obesity when plenty of palatable foods are available because the elevated plasma leptin levels resulting from the increased adipose tissue mass do not prevent the development of obesity. in conclusion, in humans, the leptin signaling system appears to be mainly involved in maintenance of adequate energy stores for survival during periods of energy deficit. its role in the etiology of human obesity is only demonstrated in the very rare situations of absence of the leptin signal (mutations of the leptin gene or of the leptin receptor gene), which produces an internal perception of starvation and results in a

Fig. 9 Predicted labels with background (green), outcome (yellow) of the model *MULT GRP*. The true labels are shown in Fig. 7 (colour figure online)

Then, the Silhouette score for $\mathbf{v} \in C_l$ is defined as:

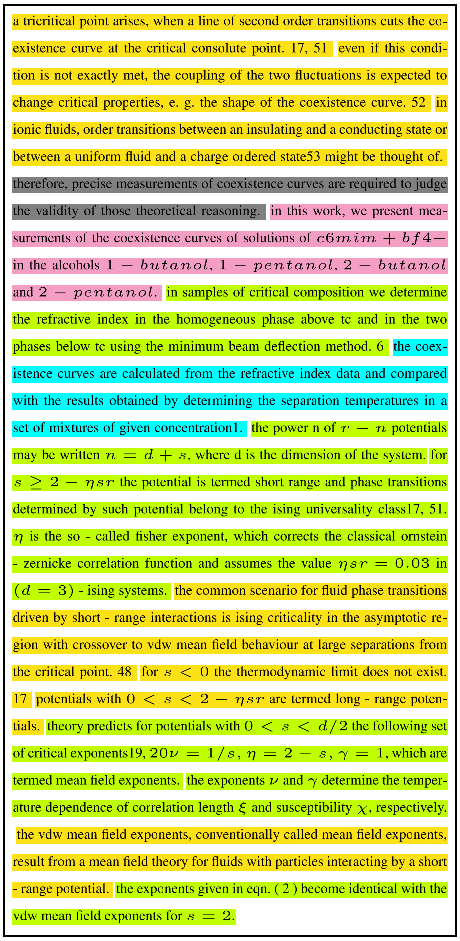
$$s(\mathbf{v}) = \begin{cases} \frac{b(\mathbf{v})-a(\mathbf{v})}{max\{a(\mathbf{v}),b(\mathbf{v})\}} & \text{if } |C_l| > 1 \\ 0 & \text{if } |C_l| = 1 \end{cases} \tag{18}$$

As a distance metric d, we use semantic_relatedness as defined in Eq. 15. Now, we can compute the Silhouette score for a cluster C_l as the arithmetic mean of all Silhouette scores in this cluster:

$$s(C_l) = \frac{1}{|C_l|} \sum_{\mathbf{v} \in C_l} s(\mathbf{v}) \tag{19}$$

A positive Silhouette coefficient indicates that objects homogeneously lie well within the cluster and do not interfere with other clusters, while a negative score indicates that the objects are merely somewhere in between clusters.

Table 7 shows the Silhouette scores [68] for each cluster. This evaluation uses our assignment of the datasets' annotation classes to one of the six core clusters identified based on the respective annotation guidelines. It can be seen that MULT ALL has the highest Silhouette coefficient (SC) and thus forms better clusters than SciBERT-HSLN and MULT GRP. The best overall result for MULT ALL is mainly caused by the relatively good cluster quality for the classes *Background* (SC = 0.48), *Results* (SC = 0.32) and *Methods* (SC = 0.31). In contrast, the quality of the three other clusters is not good. For SciBERT-HSLN, the cluster quality is relatively good for the class *Background* (SC = 0.45) and even very good for the class *Conclusions* (SC = 0.92), but the SC scores for the other four clusters are between −0.38 and 0.19. For MULT GRP, the results are the worst with an overall SC score of −0.02, whereby the best SC score (among MULT GRP clusters) is obtained for *Background* (SC = 0.18). The class *Background* achieved relatively good scores across all three methods. We hypothesise that MULT ALL can capture the semantic relatedness of classes better than the other approaches since it is enforced to learn a generic feature extractor across multiple datasets.

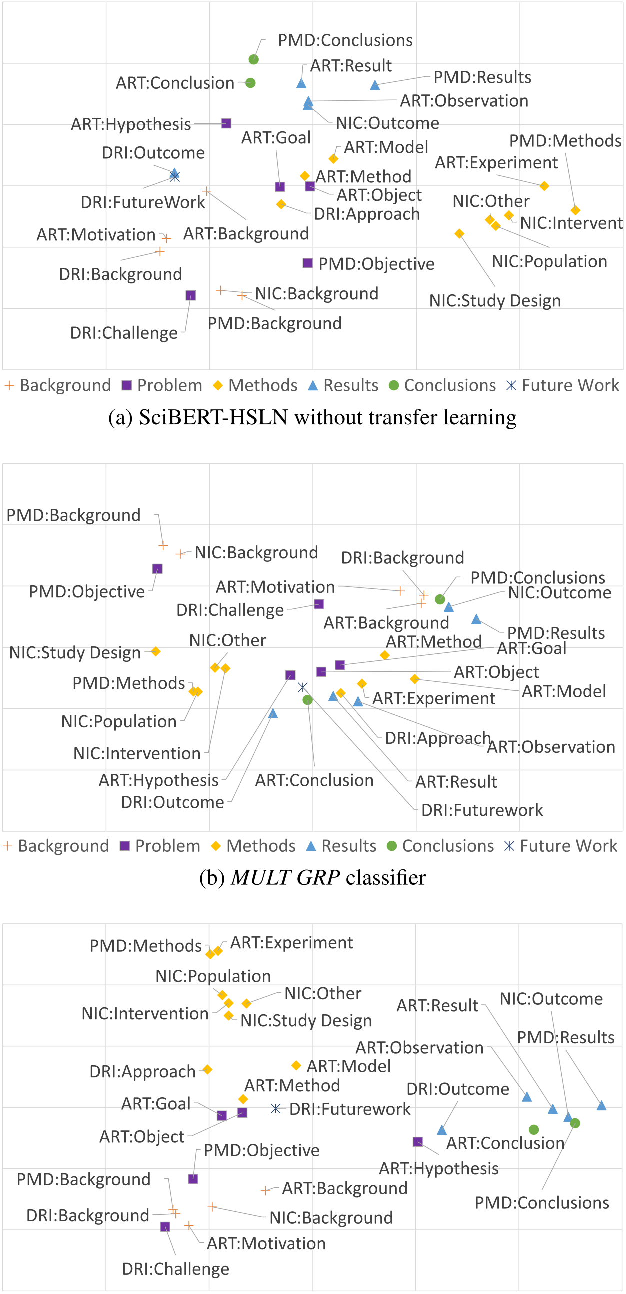


a tricritical point arises, when a line of second order transitions cuts the coexistence curve at the critical consolute point. 17, 51 even if this condition is not exactly met, the coupling of the two fluctuations is expected to change critical properties, e. g. the shape of the coexistence curve. 52 in ionic fluids, order transitions between an insulating and a conducting state or between a uniform fluid and a charge ordered state53 might be thought of. therefore, precise measurements of coexistence curves are required to judge the validity of those theoretical reasoning. in this work, we present measurements of the coexistence curves of solutions of $c6mim + bf4-$ in the alcohols $1 - butanol$, $1 - pentanol$, $2 - butanol$ and $2 - pentanol.$ in samples of critical composition we determine the refractive index in the homogeneous phase above tc and in the two phases below tc using the minimum beam deflection method. 6 the coexistence curves are calculated from the refractive index data and compared with the results obtained by determining the separation temperatures in a set of mixtures of given concentration1. the power n of $r - n$ potentials may be written $n = d + s$, where d is the dimension of the system. for $s \geq 2 - \eta sr$ the potential is termed short range and phase transitions determined by such potential belong to the ising universality class17, 51. η is the so - called fisher exponent, which corrects the classical ornstein - zernicke correlation function and assumes the value $\eta sr = 0.03$ in $(d = 3)$ - ising systems. the common scenario for fluid phase transitions driven by short - range interactions is ising criticality in the asymptotic region with crossover to vdw mean field behaviour at large separations from the critical point. 48 for $s < 0$ the thermodynamic limit does not exist. 17 potentials with $0 < s < 2 - \eta sr$ are termed long - range potentials. theory predicts for potentials with $0 < s < d/2$ the following set of critical exponents19, $20\nu = 1/s$, $\eta = 2 - s$, $\gamma = 1$, which are termed mean field exponents. the exponents ν and γ determine the temperature dependence of correlation length ξ and susceptibility χ, respectively. the vdw mean field exponents, conventionally called mean field exponents, result from a mean field theory for fluids with particles interacting by a short - range potential. the exponents given in eqn. (2) become identical with the vdw mean field exponents for $s = 2$.

Fig. 10 Predicted labels with background (yellow), motivation (grey), object (purple), method (brown), model (green) of the model *MULT GRP*. The true labels are shown in Fig. 5 (colour figure online)

The multi-task learning approach sharing all possible layers can 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 clus-



(a) SciBERT-HSLN without transfer learning

(b) *MULT GRP* classifier

(c) *MULT ALL* classifier

Fig. 11 Semantic vectors of classes computed by **a** SciBERT-HSLN model without transfer learning, **b** *MULT GRP* model, and **c** *MULT ALL* model and projected to 2D space using PCA. The semantic vectors are assigned to generic clusters of semantically related labels

Table 8 Characteristics of the domain-independent dataset *G-PNDA* that was compiled from the origin datasets *PMD*, *NIC*, *DRI* and *ART*

Text Type	G-PMD Abstract	G-NIC Abstract	G-DRI Full paper	G-ART Full paper
# Papers	1000	1000	40	67
# Sentences	11,738	9771	8777	9528
∅ # Sentences	11	10	219	142
Background	1220	2548	1760	1657
Problem	953	0	449	529
Methods	3927	2700	5038	2752
Results	3760	4523	1394	3672
Conclusions	1878	0	0	918
Future work	0	0	136	0

ters in Fig. 11. Table 8 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 9 depicts our experimental settings and 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 Sect. 3.2). For training and evaluation, we split each dataset part 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 Sect. 4.2).

Table 9 shows that the proposed MULT GRP model outperforms all other settings. Surprisingly, sharing the output layer impairs 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 to capture the individual rhetorical structure of the domains (#RQ3, #RQ6).

Automatic Domain-Independent Clustering: To further validate the quality of the calculated semantic vectors for different annotation schemes, we conducted an additional experiment apart from the visual analysis performed before. For this, the vectors as shown in Fig. 4 are used for automatic clustering with the k-means algorithm. We initialise the algorithm with 1000 different random seeds and vary the number of clusters from two to ten to find the optimal clustering (i.e. highest Silhouette coefficient) and compare it with the semi-automatic approach. The *MULT GRP*, *SCIBERT-HSLN* and *MULT ALL* models perform similarly with the best Silhouette scores ranging from 0.51 to 0.59, respectively. To investigate the difference between the best automatic clustering and the semi-automatically assigned clusters, we compare the best automatic clusters of the MULT ALL model to the semi-automatic ones in Table 10. The k-means clusters for three, five and seven clusters have the best cluster quality.

To evaluate the semi-automatic and automatic clustering results, we compare the resulting clusters. For this purpose, we count the number of overlapping cluster assignments: The automatically computed cluster that has the largest overlap to

Table 9 Experimental results in terms of $F1$ scores (in per cent) for our proposed approaches for the generic 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

	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**

Bold depicts whether the approach improves the baseline, underline the best overall result

Table 10 Clusters for MULT ALL via k-means for different numbers of clusters (n)

n (SilSc)	Clusters	Assigned labels
6 (0.20)	ART:Background, ART:Motivation, DRI:Background, PMD:Background, NIC:Background	Background
	ART:Hypothesis, ART:Goal, ART:Object, DRI:Challenge, PMD:Objective	Problem
	ART:Experiment, ART:Model, ART:Method, DRI:Approach, PMD:Methods, NIC:Intervention, NIC:Study Design, NIC:Population, NIC:Other	Methods
	ART:Observation, ART:Result, DRI:Outcome, PMD:Results, NIC:Outcome	Results
	ART:Conclusion, PMD:Conclusions	Conclusions
	DRI:FutureWork	Future Work
3 (0.59)	**PMD:Methods**, **NIC:Intervention**, **NIC:Study Design**, **NIC:Population**, **NIC:Other**, **DRI:Approach**, DRI:FutureWork, ART:Goal, ART:Object, **ART:Experiment**, **ART:Model**, **ART:Method**	Methods
	PMD:Results, PMD:Conclusions, **NIC:Outcome**, **DRI:Outcome**, ART:Hypothesis, **ART:Observation**, **ART:Result**, ART:Conclusion	Results
	PMD:Background, PMD:Objective, **NIC:Background**, **DRI:Background**, DRI:Challenge, **ART:Background**, **ART:Motivation**	Background
5 (0.57)	**ART:Experiment**, **NIC:Other**, **PMD:Methods**, **NIC:Population**, **NIC:intervention**, **NIC:Study Design**	Methods
	NIC:Outcome, **ART:Result**, PMD:Conclusions, ART:Hypothesis, **DRI:Outcome**, ART:Conclusion, **ART:Observation**, **PMD:Results**	Results
	PMD:Objective, **NIC:Background**, **PMD:Background**, **DRI:Challenge**	Background / Problem
	ART:Goal, DRI:FutureWork, **ART:Method**, **DRI:Approach**, **ART:Model**, ART:Object	Methods
	ART:Motivation, **ART:Background**, **DRI:Background**	Background
7 (0.57)	**ART:Experiment**, **NIC:Population**, **NIC:Other**, **PMD:Methods**, **NIC:Intervention**, **NIC:Study Design**	Methods
	NIC:Outcome, **ART:Result**, PMD:Conclusions, ART:Hypothesis, **DRI:Outcome**, ART:Conclusion, **ART:Observation**, **PMD:Results**, **NIC:Background**, **PMD:Background**, DRI:Challenge	Results
	ART:Goal, **ART:Object**	Problem
	ART:Motivation, **ART:Background**, **DRI:Background**	Background
	NIC:Background, **PMD:Background**, DRI:Challenge	Background
	ART:Model, DRI:FutureWork, **DRI:Approach**, **ART:Method**	Method
	PMD:Objective	Problem

This Silhouette score is shown in the first column in parentheses (*SilSc*). To compare the clusters with the semi-automatic ones, the cluster with the biggest overlap is indicated in the "Assigned Labels" column with overlapping classes indicated in bold. The first row (n = 6) contains the results of the semi-automatic clustering

a semi-automatically determined cluster is assumed to be its correspondence and will be assigned the respective label. In the case of the k = 3 clustering, for instance, cluster 0 shares nine assigned classes with cluster 2 ("Methods") from the semi-automatic clustering. We thus assume that cluster 0 is the correspondent of the semi-automatic "Methods" cluster. Table 10 shows this assignment in the "Assigned labels" column. The overlapping classes are highlighted in bold.

The automatic clustering with k = 3 clusters can differentiate well between the concepts "Background", "Methods" and "Results". The majority of concepts have been correctly assigned; all incorrect assignments are due to the small number of clusters, i.e. concepts which do not fit the three "main" clusters had to be assigned to one of the available clusters. For k = 5 clusters, the two labels "Methods" and "Background" were assigned to two clusters each (in one case being on par with "Problem"), resulting in a total of three different labels considered. There are two possible interpretations of this behaviour: The predominance of the three classes could be caused by an imbalance in the training set (that might impact the semantic vectors): Table 8 shows that these three labels are the most frequently assigned to sentences in the used datasets. Another issue could be the usage of k-means clustering, an approach that tends to find clusters of similar size [89], In our unified annotation schema, the clusters "Conclusions" and "Future Work" have a smaller size than the three above.

When using $k = 7$ clusters, we observe a similar effect. However here, the "Problem" class is split as well. The results indicate that k-means has trouble correctly identifying (or unifying) semantically similar classes for a higher number of five or more clusters. Semantically meaningful classes such as "Conclusions" or "Future Work" are not identified. Nevertheless, the results of our exploration of k-means clustering indicate that the semantic vectors correctly encode the semantic difference between "Background", "Methods" and "Results", and to less extent for "Problem".

5.6 Limitations

The datasets used in this paper depend on the annotation schemes and the distribution of classes in research papers. As sections on "Background" are typically longer than for, for example, "Hypothesis", they do not have the same number of sample phrases (e.g. 4290 "ART:Background" samples and 488 "ART:Hypothesis" in the training data set). This bias is present in the model as well. Figure 4 illustrates this problem. The values in the "ART:Hypothesis" column are generally low and often even zero. Thus, this class is rarely predicted. The row for that class shows that even sentences with that ground truth label are often predicted as other, more common ART classes such as "ART:Conclusion" or "ART:Result".

The classes for the generic dataset described in Sect. 5.5 are based on the general rhetorical structure of research papers and are derived from ontologies such as [19, 35, 82]. While different scientific disciplines use different formats, we could not test all of them. Nevertheless, the presented approach allows for the easy adoption and validation of other annotation schemes. The comparison with the k-means clustering shows that the semi-automatic clusters which had a high Silhouette score in Table 7 were also well detected automatically (e.g. for k-means with three clusters in Table 10). This supports the general structure of the semi-automatic approach which is semantically more fine-grained in terms of classes like "Future Work".

The examples in Sect. 5.4 show that the models often identify the correct boundary of topics but occasionally misclassify the labels. It also demonstrates that the distinction between classes is not always obvious, even to humans.

As described in Sect. 4.1, the four datasets were chosen as they cover different scientific domains, annotation schemes, and include full-text papers as well as abstracts. Four datasets were chosen for practical reasons but other datasets presented in Table 1 could be used to extend the study.

6 Conclusions

In this paper, we have presented a unified deep learning architecture for sequential sentence classification. The unified approach 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 (#RQ7).

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 (#RQ2). Our comprehensive experimental evaluation with four different datasets offers useful insights under which conditions transferring or sharing of specific layers is beneficial or not (#RQ3). 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. 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, i.e. DRI [29] with 84.4% $F1$ (+11.9% absolute improvement) and ART [53] with 58.8% accuracy (+7.2% absolute improvement) (#RQ4). In particular, models for tasks with small datasets and classes with few labelled examples benefit significantly from models of other tasks. We investigated the differences and problems of *MULT ALL* and *MULT GRP* through multiple examples. Our analysis 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 (#RQ1). This semantic relatedness is an important prerequisite for transfer learning in NLP tasks [58, 62, 69].

Finally, we have presented 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 and compared it to fully automatic clustering using k-means. The results showed that the cross-domain categories (as defined by us) with more than two concepts are also represented in clusters with relatively high precision by k-means (in a single cluster for

k = 3; in some cases in two clusters). The semi-automatic approach can support the investigation of annotation schemes across disciplines without re-annotating datasets (#RQ5). From the analysis, we have 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 (#RQ6).

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 and to evaluate its impact on retrieval performance in academic search engines.

Since its first presentation during the Joint Conference on Digital Libraries'22 [2], this work has also been adopted for legal sequential sentence classification [44]. Kalamkar et al. present a new corpus for the automatic structuring of legal documents and evaluate several baseline algorithms for sentence classification, of which our SciBERT-HSLN performs best. This result shows that the here-presented approaches can potentially be adopted for further sentence classification tasks.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. AbuRa'ed, A., Saggion, H., Shvets, A., Bravo, À.: Automatic related work section generation: experiments in scientific document abstracting. Scientometrics (2020). https://doi.org/10.1007/s11192-020-03630-2
2. Aizawa, A., Mandl, T., Carevic, Z., Hinze, A., Mayr, P., Schaer, P. (eds.): JCDL '22: The ACM/IEEE Joint Conference on Digital Libraries in 2022, Cologne, Germany, June 20 - 24, 2022. ACM (2022). https://doi.org/10.1145/3529372
3. Asadi, N., Badie, K., Mahmoudi, M.T.: Automatic zone identification in scientific papers via fusion techniques. Scientometrics (2019). https://doi.org/10.1007/s11192-019-03060-9
4. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: 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, pp. 546–555. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/S17-2091
5. Badie, K., Asadi, N., Mahmoudi, M.T.: Zone identification based on features with high semantic richness and combining results of separate classifiers. J. Inf. Telecommun. (2018). https://doi.org/10.1080/24751839.2018.1460083
6. Banerjee, S., Sanyal, D.K., Chattopadhyay, S., Bhowmick, P.K., Das, P.P.: 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, pp. 429–432. ACM (2020). https://doi.org/10.1145/3383583.3398598
7. Beltagy, I., Lo, K., Cohan, A.: 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, pp. 3613–3618. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1371
8. Bornmann, L., Mutz, R.: Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. J. Assoc. Inf. Sci. Technol. (2015). https://doi.org/10.1002/asi.23329
9. Brack, A., D'Souza, J., Hoppe, A., Auer, S., Ewerth, R.: 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, pp. 251–266. Springer (2020). https://doi.org/10.1007/978-3-030-45439-5_17
10. Brack, A., Hoppe, A., Buschermöhle, P., Ewerth, R.: Cross-domain multi-task learning for sequential sentence classification in research papers. In: JCDL '22: The ACM/IEEE Joint Conference on Digital Libraries in 2022, Cologne, Germany, June 20–24, 2022, p. 34. ACM (2022). https://doi.org/10.1145/3529372.3530922
11. Brack, A., Hoppe, A., Stocker, M., Auer, S., Ewerth, R.: Analysing the requirements for an open research knowledge graph: use cases, quality requirements, and construction strategies. Int. J. Digit. Libr. (2022). https://doi.org/10.1007/s00799-021-00306-x
12. Brack, A., Müller, D.U., Hoppe, A., Ewerth, R.: 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, pp. 79–97. Springer (2021). https://doi.org/10.1007/978-3-030-72113-8_6
13. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: 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 (2020). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
14. Changpinyo, S., Hu, H., Sha, F.: 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, pp. 2965–2977. Association for Computational Linguistics (2018). https://www.aclweb.org/anthology/C18-1251/
15. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: 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, pp. 1724–1734. ACL (2014). https://doi.org/10.3115/v1/d14-1179

16. Cohan, A., Ammar, W., van Zuylen, M., Cady, F.: 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), pp. 3586–3596. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1361
17. Cohan, A., Beltagy, I., King, D., Dalvi, B., Weld, D.S.: 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, pp. 3691–3697. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1383
18. Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W., Goharian, N.: 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), pp. 615–621. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-2097
19. Constantin, A., Peroni, S., Pettifer, S., Shotton, D.M., Vitali, F.: The document components ontology (doco). Semantic Web (2016). https://doi.org/10.3233/SW-150177
20. Dayrell, C., Jr., A.C., Lima, G., Jr., D.M., Copestake, A.A., Feltrim, V.D., Tagnin, S.E.O., Aluísio, S.M.: 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, pp. 1604–1609. European Language Resources Association (ELRA) (2012). http://www.lrec-conf.org/proceedings/lrec2012/summaries/734.html
21. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
22. Dernoncourt, F., Lee, J.Y.: 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, pp. 308–313. Asian Federation of Natural Language Processing (2017). https://www.aclweb.org/anthology/I17-2052/
23. Dernoncourt, F., Lee, J.Y., Szolovits, P.: 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, pp. 694–700. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/e17-2110
24. Devlin, J., Chang, M., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
25. DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B., Wang, L.L.: 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, pp. 7494–7513. Association for Computational Linguistics (2021). https://aclanthology.org/2021.emnlp-main.594
26. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. (1998). https://doi.org/10.1162/089976698300017197
27. Edwards, A., Camacho-Collados, J., de Ribaupierre, H., Preece, A.D.: Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In: Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8–13, 2020, pp. 5522–5529. International Committee on Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.coling-main.481
28. Fellbaum, C., Miller, G.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
29. Fisas, B., Saggion, H., Ronzano, F.: 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, pp. 42–51. The Association for Computer Linguistics (2015). https://doi.org/10.3115/v1/w15-1605
30. Forney, G.D.: The viterbi algorithm. Proc. IEEE (1973). https://doi.org/10.1109/PROC.1973.9030
31. Friedrich, A., Adel, H., Tomazic, F., Hingerl, J., Benteau, R., Marusczyk, A., Lange, L.: 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, pp. 1255–1268. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.116
32. Gábor, K., Buscaldi, D., Schumann, A., QasemiZadeh, B., Zargayouna, H., Charnois, T.: 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, pp. 679–688. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/s18-1111
33. Ghosh Roy, S., Pinnaparaju, N., Jain, R., Gupta, M., Varma, V.: Summaformers @ LaySumm 20, LongSumm 20. In: Proceedings of the First Workshop on Scholarly Document Processing, pp. 336–343. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.sdp-1.39
34. Gonçalves, S., Cortez, P., Moro, S.: A deep learning classifier for sentence classification in biomedical and computer science abstracts. Neural Comput. Appl. (2020). https://doi.org/10.1007/s00521-019-04334-2
35. Groza, T., Handschuh, S., Möller, K., Decker, S.: SALT—semantically annotated latex for scientific publications. In: The Semantic Web: Research and Applications, 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3–7, 2007, Proceedings, pp. 518–532. Springer (2007). https://doi.org/10.1007/978-3-540-72667-8_37
36. Gupta, K., Ahmad, A., Ghosal, T., Ekbal, A.: 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, pp. 436–452. Springer (2021). https://doi.org/10.1007/978-3-030-91669-5_34
37. He, P., Liu, X., Gao, J., Chen, W.: 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 (2021). https://openreview.net/forum?id=XPZIaotutsD

38. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. (1997). https://doi.org/10.1162/neco.1997.9.8.1735
39. Howard, J., Ruder, S.: 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, pp. 328–339. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/P18-1031
40. Jia, R., Wong, C., Poon, H.: 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), pp. 3693–3704. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1370
41. Jin, D., Szolovits, P.: 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, pp. 3100–3109. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/d18-1349
42. Jolliffe, I.T.: Principal component analysis. In: International Encyclopedia of Statistical Science, pp. 1094–1096. Springer (2011). https://doi.org/10.1007/978-3-642-04898-2_455
43. Kabongo, S., D'Souza, J., Auer, S.: 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, pp. 453–470. Springer (2021). https://doi.org/10.1007/978-3-030-91669-5_35
44. Kalamkar, P., Tiwari, A., Agarwal, A., Karn, S., Gupta, S., Raghavan, V., Modi, A.: Corpus for automatic structuring of legal documents. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20–25 June 2022, pp. 4420–4429. European Language Resources Association (2022). https://aclanthology.org/2022.lrec-1.470
45. Kim, S., Martínez, D., Cavedon, L., Yencken, L.: Automatic classification of sentences to support evidence based medicine. BMC Bioinform. (2011). https://doi.org/10.1186/1471-2105-12-S2-S5
46. Kingma, D.P., Ba, J.: 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 (2015). arXiv:1412.6980
47. Kunnath, S.N., Pride, D., Gyawali, B., Knoth, P.: Overview of the 2020 WOSP 3C citation context classification task. In: Proceedings of the 8th International Workshop on Mining Scientific Publications, pp. 75–83. Association for Computational Linguistics, Wuhan, China (2020). https://www.aclweb.org/anthology/2020.wosp-1.12
48. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: 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, pp. 282–289. Morgan Kaufmann (2001)
49. Lauscher, A., Glavas, G., Eckert, K.: 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, pp. 22–28. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/w18-5203
50. Lauscher, A., Glavas, G., Ponzetto, S.P., Eckert, K.: 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, pp. 3326–3338. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/d18-1370
51. Lee, J.Y., Dernoncourt, F., Szolovits, P.: 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. European Language Resources Association (ELRA) (2018). http://www.lrec-conf.org/proceedings/lrec2018/summaries/878.html
52. Liakata, M., Saha, S., Dobnik, S., Batchelor, C.R., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics (2012). https://doi.org/10.1093/bioinformatics/bts071
53. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.R.: 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. European Language Resources Association (2010). http://www.lrec-conf.org/proceedings/lrec2010/summaries/644.html
54. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
55. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: 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, pp. 3219–3232. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/d18-1360
56. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika (1947). https://doi.org/10.1007/bf02295996
57. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: 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, pp. 3111–3119 (2013). https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html
58. Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., Jin, Z.: 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, pp. 479–489. The Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/d16-1046
59. Nasar, Z., Jaffry, S.W., Malik, M.K.: Information extraction from scientific articles: a survey. Scientometrics (2018). https://doi.org/10.1007/s11192-018-2921-5
60. Neves, M.L., Butzke, D., Grune, B.: 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, pp. 124–135. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/w19-4515
61. Oelen, A., Stocker, M., Auer, S.: Crowdsourcing scholarly discourse annotations. In: IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13–17, 2021, pp. 464–474. ACM (2021). https://doi.org/10.1145/3397481.3450685
62. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. (2010). https://doi.org/10.1109/TKDE.2009.191

63. Park, S., Caragea, C.: 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, pp. 5409–5419. International Committee on Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.coling-main.472
64. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019). https://dl.acm.org/doi/10.5555/3454287.3455008
65. Pennington, J., Socher, R., Manning, C.D.: 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, pp. 1532–1543. ACL (2014). https://doi.org/10.3115/v1/d14-1162
66. Pruksachatkun, Y., Phang, J., Liu, H., Htut, P.M., Zhang, X., Pang, R.Y., Vania, C., Kann, K., Bowman, S.R.: 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, pp. 5231–5247. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.467
67. Reimers, N., Gurevych, I.: 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, pp. 338–348. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/d17-1035
68. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. (1987). https://doi.org/10.1016/0377-0427(87)90125-7
69. Ruder, S.: Neural transfer learning for natural language processing. Ph.D. thesis, National University of Ireland, Galway (2019)
70. Safder, I., Hassan, S.: Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. Scientometrics (2019). https://doi.org/10.1007/s11192-019-03025-y
71. Safder, I., Hassan, S., Visvizi, A., Noraset, T., Nawaz, R., Tuarob, S.: Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. Inf. Process. Manag. (2020). https://doi.org/10.1016/j.ipm.2020.102269
72. Sanh, V., Wolf, T., Ruder, S.: 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, pp. 6949–6956. AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.33016949
73. Schulz, C., Eger, S., Daxenberger, J., Kahse, T., Gurevych, I.: 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), pp. 35–41. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/n18-2006
74. Semwal, T., Yenigalla, P., Mathur, G., Nair, S.B.: 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, pp. 513–521. SIAM (2018). https://doi.org/10.1137/1.9781611975321.58
75. Shang, X., Ma, Q., Lin, Z., Yan, J., Chen, Z.: 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, pp. 198–203. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-short.26
76. Spangher, A., May, J., Shiang, S., Deng, L.: 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, pp. 498–517. Association for Computational Linguistics (2021). https://aclanthology.org/2021.emnlp-main.40
77. Stead, C., Smith, S., Busch, P.A., Vatanasakdakul, S.: 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, pp. 120–125. Australasian Language Technology Association (2019). https://aclweb.org/anthology/papers/U/U19/U19-1016/
78. Su, X., Li, R., Li, X.: 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, pp. 457–469. Springer (2020). https://doi.org/10.1007/978-3-030-60450-9_36
79. Teufel, S.: Argumentative zoning: Information extraction from scientific text. Ph.D. thesis, University of Edinburgh (1999)
80. Teufel, S., Siddharthan, A., Batchelor, C.R.: 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, pp. 1493–1502. ACL (2009). https://www.aclweb.org/anthology/D09-1155/
81. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: 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, pp. 5998–6008 (2017). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
82. de Waard, A., Tel, G.: The ABCDE format enabling semantic conference proceedings. In: SemWiki2006, First Workshop on Semantic Wikis—From Wiki to Semantics, Proceedings, co-located with the ESWC2006, Budva, Montenegro, June 12, 2006. CEUR-WS.org (2006). http://ceur-ws.org/Vol-206/paper8.pdf
83. Wei, Z., Jia, Y., Tian, Y., Hosseini, M.J., Steedman, M., Chang, Y.: Joint extraction of entities and relations with a hierarchical multi-task tagging model. CoRR (2019). arXiv:1908.08672
84. Weiss, K.R., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data (2016). https://doi.org/10.1186/s40537-016-0043-6
85. Xiong, C., Power, R., Callan, J.: 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, pp. 1271–1279. ACM (2017). https://doi.org/10.1145/3038912.3052558
86. Yamada, K., Hirao, T., Sasano, R., Takeda, K., Nagata, M.: Sequential span classification with neural semi-Markov CRFs for

biomedical abstracts. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 871–877. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.77

87. Yang, Z., Salakhutdinov, R., Cohen, W.W.: 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 (2017). https://openreview.net/forum?id=ByxpMd9lx
88. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A.J., Hovy, E.H.: 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, pp. 1480–1489. The Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/n16-1174
89. Zhou, K., Yang, S.: Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering. Pattern Anal. Appl. (2020). https://doi.org/10.1007/s10044-019-00783-6