Philipp Heinisch* and Philipp Cimiano

A multi-task approach to argument frame classification at variable granularity levels

<https://doi.org/10.1515/itit-2020-0054>

Received December 5, 2020; revised February 11, 2021; accepted March 15, 2021

Abstract: Within the field of argument mining, an important task consists in predicting the frame of an argument, that is, making explicit the aspects of a controversial discussion that the argument emphasizes and which narrative it constructs. Many approaches so far have adopted the framing classification proposed by Boydstun et al. [3], consisting of 15 categories that have been mainly designed to capture frames in media coverage of political articles. In addition to being quite coarse-grained, these categories are limited in terms of their coverage of the breadth of discussion topics that people debate. Other approaches have proposed to rely on issue-specific and subjective (argumentation) frames indicated by users via labels in debating portals. These labels are overly specific and do often not generalize across topics. We present an approach to bridge between coarse-grained and issue-specific inventories for classifying argumentation frames and propose a supervised approach to classifying frames of arguments at a variable level of granularity by clustering issue-specific, user-provided labels into frame clusters and predicting the frame cluster that an argument evokes. We demonstrate how the approach supports the prediction of frames for varying numbers of clusters. We combine the two tasks, frame prediction with respect to media frames categories as well as prediction of clusters of user-provided labels, in a multi-task setting, learning a classifier that performs the two tasks. As main result, we show that this multi-task setting improves the classification on the single tasks, the media frames classification by up to +9.9 % accuracy and the cluster prediction by up to +8 % accuracy.

Keywords: framing, computational arguments, neuronal parameter sharing

ACM CCS: Computing methodologies → Artificial intelligence → Natural language processing → Information extraction

*Corresponding author: Philipp Heinisch, Universität Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany, e-mail: pheinisch@techfak.uni-bielefeld.de

Philipp Cimiano, Universität Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany, e-mail: cimiano@techfak.uni-bielefeld.de

1 Introduction

Nowadays, users share arguments on controversial discussion topics online through a plethora of argumentation and debating portals. In recent years, there has been a growing interest on automating the analysis of such arguments to structure the discussion and support deliberation, giving raise to a field called argument mining [31, 4, 13, 17]. Search engines that index and support retrieval of arguments have been developed, such as args.me [30] and the ArgumenText search engine [27]. Beyond the mere retrieval of arguments, the field of argument mining is increasingly considering tasks related to the grouping, summarization and ranking of arguments, providing more advanced functionality to support users in obtaining an overview about the arguments exchanged by users on a certain topic. When analysing arguments, an important aspect is to understand the perspective from which the argument is *framed*, that is, which aspects of the discussion it emphasizes and which narrative it constructs [8, 1]. When providing an overview of the arguments that are exchanged on the Web, a breakdown of arguments by frame and/or stakeholder type is key to understand potential argumentative tactics, hidden agendas, etc.

Consider as an example the following two arguments arguing against the lockdown, but using different frames:

1. “I think the lockdown in the COVID-19-outbreak was a wrong decision because it ruins the economy. I know some successful companies which are bankrupt now because of the lockdown.” frame: economics
2. “Yes, the lockdown decreased the infection rate, but consider mental health, too! Humanity needs (offline) interaction with each other. We’re created as social beings. Hence, the long isolation (for some among us) harms possibly persistently the total health. frame: health

So far, approaches to frame classification have mainly relied on predefined, coarse-grained and issue-independent inventories of frame categories and – moreover – are often not tailored to whole arguments but text spans in newspaper texts, for example. One popular inventory of frames is captured in the Media-Frames-set defined by Boydstun et al. [3]. This scheme consists of 15 categories that have been mainly designed to capture the different frames within political discussions. However, the frame types are too

coarse-grained to cover the potential argumentative perspectives in discussions of arbitrary topics.

On the other hand, recent work has proposed a bottom-up and data-driven approach to infer more comprehensive, issue-specific frames. Ajjour et al. [1] have proposed to rely on user labels of arguments exchanged on debatepedia.org as approximations of frames. Such user labels range from very general ones such as “economics” to overly specific ones such as “protecting non-smokers”. The number of these *issue-specific frames* is, however, very high, as they capture the very specific perspective of a user on the topic under discussion. For this reason, Ajjour et al. [1] explored an unsupervised approach in which the arguments are clustered on the basis of the user-provided labels.

Building on these previous approaches, we focus on the supervised classification of argument frames. On the one hand, we consider the classification of arguments with respect to the Media-Frames proposed by Boydstun et al. [3]. On the other hand, we consider the (supervised) prediction of a specific cluster of user-defined labels. Our approach relies on clustering the issue-specific labels as generated by users in debatepedia.org such as “protecting non-smokers” into more coarse-grained frames. We rely on semantic similarity measures to perform this clustering. The advantage of our approach is that users can switch the kind and the granularity of the frames as needed for a certain application and thus allows a user to control the granularity of the (distinguished) frame classes. These two approaches are strongly related, but can be regarded as two different tasks as they rely on different datasets as well as different classification labels. In this paper we explore whether the two tasks can benefit from each other in a multi-task setting, training a classifier that optimizes joint representations of arguments to perform well on both tasks.

Thus, in this paper, we investigate whether (subsets of) datasets that are annotated according to the Media-Frames can be successfully exploited to improve classification accuracy with respect to predicting our clusters in a multi-task setting and the other way round. We show that a multi-task setting in which a classifier is trained to solve both the classification of frames following the Media-Frames categories as well as the task of classifying an argument into our frame clusters can improve results on both tasks.

Our contributions are the following:¹

- We present a supervised frame classification system that operates at varying levels of granularity by relying on an inventory of frames induced via clustering techniques.
- We show that classification accuracy can be improved by factoring in data annotated along the Media-Frame categories and training a system in a multi-task setting to address two tasks: coarse-grained Media-Frame-classification and frame cluster classification. Performance improves up to 9.9 points depending on the multi-task-combination, the kind and number of frames considered.
- We show that results in the multi-task setting with fine-grained framing clusters are better when using a soft-parameter-sharing compared to a hard-parameter-sharing approach.

2 Related work

The formal definition of framing goes back to the work by Entman [11], who defined it as ‘*Framing essentially involves selection and salience. To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described.*’ Early work on frame identification was done by Paul and Girju [22], who developed a probabilistic model to provide a token-based view on the topic and aspect of a document. However, their notion of aspect is rather related to the party that authors the document rather than to the argumentative perspective conveyed in the text. One of the first sets of frames to support argument analyses was developed by Neuman et al. [21], who focused on the following frames: thematic frame, human impact frame, powerlessness frame, economics frame, moral values frame and conflict frame. More recent generic sets of frames have been proposed by Semetko and Valkenburg [26] as well as the comprehensive framing set of Boydstun et al. [3], which was tailored to the analysis of media coverage of political debates. This Media-Frames-set consists of the following 15 categories: (i) **economic**, (ii) capacity and resources, (iii) morality, (iv) fairness and equality, (v) **legality, constitutionality and jurisprudence**, (vi) policy prescription and evaluation, (vii) **crime and punishment**, (viii) security and defence, (ix) health and safety, (x) quality of life, (xi) cultural identity, (xii) public opinion, (xiii) **political**, (xiv) external regulation and reputation, (xv) other (we mark the most frequent frames in bold).

¹ The code and datasets can be downloaded at <https://github.com/phhei/FramingNN>

One problem of these fixed, generic framesets is that they assume a fixed granularity at which texts are analyzed. Further, they are limited in the topics they cover, as they were typically designed for a specific type of discourse, e. g. political discussion in the case of the Media-Frames inventory. For this reason, de Vreese [8] analysed manually a combination of fixed generic frames with a set of issue-specific frames. These issue-specific frames can range from concepts such as ‘*economic insolvency*’ through to complete claims [2] such as ‘*personal fundamental rights are more important than the public interest.*’.

One popular work in terms of annotated resources for frame identification is the dataset by Card et al. [5], containing approximately 200,000 annotated text spans in articles on the basis of the generic Media-Frameset [3]. This is one of the datasets we use in our experiments (see Section 3.2 for details). Building on this dataset, Naderi and Hirst [20] presented one of the first computational models to address the prediction of a Media-Frame class for a given text. The best performing model was a recurrent neural network with word embeddings as input. Showing that automatic frame classification is feasible, they present results for multiple settings, that is multiple one-vs-rest classification scenarios (topic-specific accuracies are up to 92.5%). However, their results showed a clear drop in accuracy by 18% (from 71.2% to 58.7%) when moving from classifying the top 5 frames to classifying 15 frames. The accuracy drops to 53.7% when using a topic-agnostic classifier that is trained on the topics “immigration” and “same-sex marriage” at the same time. Their work thus clearly showed the problem of transferability of models for one topic to another and indicated that the performance on the task is very dependent on the inventory of frame types that is adopted. This dependence is also corroborated by the work of Field et al. [12], who presented a cross-lingual approach to detect subtle media manipulation strategies. They gathered cue words for every frame and translated and generalized the resulting framing lexicon to the target dataset.

On the other side, moving beyond fixed and topic-independent frame sets, recent work proposed to use embeddings of arguments and frames to compute a more flexible association between arguments and frames. Naderi and Hirst [19] introduced an approach that starts from the seven issue-specific claims provided in the small dataset by Boltužić and Šnajder [2]. They computed embeddings for the words of each argument and each frame, and proposed different vector operations to predict the most probable frame on the basis of the given embeddings. While they achieved reasonable accuracies of up to 75.4%, as

limitations of their model they acknowledge the fact that it has difficulties in processing anecdotes and references to the debate.

More recently, Schiller et al. [25] suggested to reduce frame identification to a labelling rather than to a classification task. They fine-tune a transformer model, BERT [9], to identify frames in argumentative sentences using the IOB scheme. Their approach reaches an F_1 -score of 0.71. However, their work is limited in that the model assumes that the frame is explicitly expressed in the text, which is not always the case. Consider a rephrased version of example 1: “*Successful teams lost thousands of \$ while getting 0\$ because of the lockdown*”. This argument makes perfect sense but it does not contain an explicit mention of the frame ‘*economics*’.

Our work is inspired by the work of Ajjour et al. [1], who intend to overcome the problems of fixed and generic frame sets by considering user-assigned frames from the debatepedia.org portal in which each argument is annotated by users with a label that describes the aspect emphasized by the argument. The authors compiled a dataset of 12,326 arguments containing 1,623 distinct user labels. While about 80% of the user labels are topic-specific, the remaining 20% of the user labels occur in at least two topics. As most of the labels occur only for one topic and there is a long-tail of labels occurring very rarely, the authors opt against a supervised approach and propose an unsupervised approach to identify clusters of arguments based on a clustering of the associated user-generated topics. In our work, we directly build on the work and dataset proposed by Ajjour et al. [1], extending it to a supervised approach. We rely on a clustering of the labels into less issue-specific frames and hence as a basis to close the gap to the generic-frame-set-approaches. A clustering approach allows us to control the granularity of the data-induced frames considered. Instead of predicting single user-provided and issue-specific labels, in our approach we predict membership of an argument in a certain cluster of thematically related user-provided labels. We call these groups of thematically related user-provided labels ‘*frame clusters*’. This allows to increase or decrease the granularity of frames considered as needed by a particular application while not fixing the frames a priori, but inducing them from data, leading to better coverage of the topics in a particular dataset.

Our work attempts to investigate synergies between the two related tasks of classifying texts with respect to the Media-Frame categories and the task of classifying arguments into clusters of issue-specific user-labeled frames. It has been shown in earlier work that multi-task [24, 15] and cross-dataset [29, 18] approaches that train on diversified

tasks at the same time can overcome the lack of generalization on single datasets in which isolated classifiers might be trained to overfit a particular task rather than inducing (general) representations. For instance, Tommasi et al. [29] detected biases on the task of detecting cars in images when image classification approaches were trained on single datasets in which certain categories of cars, e. g. sports cars, are over-represented. A training on multiple datasets was shown to be beneficial in terms of generalization of the classifier. In the field of natural language processing, several approaches have demonstrated a positive impact of training classifiers on multiple tasks or datasets at the same time in terms of increased generalization abilities and robustness. Wu et al. [32], for example, developed a system that leverages joint training on the task of predicting an appropriate category for a product review while at the same time extracting the relevant aspect and opinion terms in the question-answer-review. Inspired by such works, we present an approach that combines different tasks and datasets to improve generalization on the task of predicting the frame of an argument. In our approach, one model is trained to predict both the generic Media-Frames-set as well as a frame cluster. Our results show that this multi-task and multi-dataset training can improve overall generalization, leading to an improvement on both tasks compared to an approach that is trained on a single task/dataset.

3 Datasets and preprocessing

The task of predicting the frame of an argument can be regarded as a classification task. As input, we assume that we have a (structured) argument consisting of a conclusion and a set of premises. In this section we describe the two datasets we use for our experiments, the Webis-Argument-dataset by Ajjour et al. [1] and the Media-Frames-dataset by Card et al. [5]. While the arguments in the Webis-dataset are already in the desired form, the text from the Media-Frames-dataset needs to be preprocessed using some heuristics to identify the premises and conclusion.

3.1 Webis-dataset

The Webis-dataset was created by Ajjour et al. [1] and derived from <http://www.debatepedia.org>. The dataset comprises of 12,326 arguments covering 465 topics. The arguments are already decomposed into their argument units

and enriched with attributes like the topic and stance. Each argument is annotated by different users with a specific label that represents the frame of the argument. We refer to these labels as *issue-specific frames*. Overall, there are 1,623 distinct issue-specific frames in the dataset. Of these, 330 are used for at least two topics, while 80 % of the issue-specific frames appear only in a single topic. The user-generated labels mostly consist of one token. In few cases they are multi-word units or phrases. As the labels are user-generated, they are not homogenized, so that the labels have a high semantic overlap. For example, there are two distinct labels ‘*sexuality*’ and ‘*sex*’. We describe in the next section how we cluster these labels to overcome these issues.

3.2 Media-Frames-dataset

The Media-Frames-dataset developed by Card et al. [5] consists of more than 10,000 news articles on three policy topics: *immigration*, *smoking* and *same-sex marriage*. Instead of having arguments with premises and conclusion, there are roughly 200,000 annotated text spans, each labeled with at least one out of the 15 frames from the Media-Frames-set [3]. An example for a text span within a news article annotated with a frame is given below (annotated text span is given in bold):

“An additional 5,954 also should not have been naturalized 5,634 for failing to reveal arrests for serious crimes, the INS said. “If we determine that a person provided false testimony about having been arrested, they would not meet the ‘**good moral character**’ standard for citizenship” said INS Commissioner Doris Meissner.”

The text span ‘*good moral character*’ is annotated with the frame *Morality* by two annotators. No further annotator marked this text span with another frame. There were 19 annotators in total. While the two annotators agree on this example in terms of the frame, we note that the inter-annotator agreement for the frame class conducted with the overlap of text span annotations on the dataset is generally low, and varies significantly across the three topics. The most stable topic in terms of inter-annotator agreement is *immigration*, which has a Fleiss-Kappa score of 0.16, corresponding to a slight agreement. While the topic *smoking* has a higher score, most annotations originate from a single annotator within this topic, so that this is a case of spurious agreement.

To ensure consistency, we keep only those annotated text spans on which at least two annotators agree, in addition to text spans that have been annotated by a single an-

notator. We further consider a subset of the above annotations that we refer to as *high-agreement-subset* from which we remove the single-annotator annotations. The resulting dataset is of much smaller size, consisting of 21,206 annotated arguments.²

Following Naderi and Hirst [20], we postprocess the data by lower-casing it and masking numbers. Furthermore, to create input comparable to the examples from the Webis-dataset, we implement three alternative heuristics to process the texts to infer the premise and conclusion of the arguments as follows:

- **noArg**-mode: we define the premise as an empty string and the conclusion as the sentence of the article containing the annotated text span. All other text is ignored in this mode. In our example we would consider *‘If we determine that a person provided false testimony about having been arrested, they would not meet the good moral character standard for citizenship, said INS Commissioner Doris Meissner.’* as the **conclusion** with an empty premise.
- **narrow-context**-mode: we define the conclusion to comprise only of the text span annotated by the annotators, while the premise consists of the part of the sentence preceding the annotated span. In our example, we would only consider *‘good moral character’* as the **conclusion** and the preceding if-clause *‘If we determine that a person provided false testimony about having been arrested, they would not meet the’* as the **premise**. If the text span starts at the beginning of a sentence, we extract the preceding one as a premise. The remainder of the article is ignored.
- **broad-context**-mode: we consider all sentences covered by at least one word by the annotated text span in the processed article. Similarly to the first mode, we define the conclusion as the (last) considered sentence. The premise comprises of the remaining considered sentences, or, if there are not any, the sentence before the conclusion in the processed article. In our example, we would regard the sentence *‘If we determine that a person provided false testimony about having been arrested, they would not meet the good moral character standard for citizenship, said INS Commissioner Doris Meissner.’* as a **conclusion** and *‘An ad-*

ditional 5,954 also should not have been naturalized 5,634 for failing to reveal arrests for serious crimes, the INS said.’ as **premise**, hence, constructing a moral argument. The rest of the article is also ignored.

4 Methods

In this section we describe the various methods that we rely on in our experiments. On the one hand, we describe the methods we use to cluster the user-provided issue-specific frames with the goal of inducing clusters that are used as class labels for the supervised classification of arguments into frames. We then describe the architecture of the model we use to predict the frames. Furthermore, we describe the settings of our multi-task/multi-dataset approach to training a classifier by predicting the fitting Media-Frame and the relevant frame cluster with the same model. Finally, we describe our method for addressing the imbalance of the data as the size of the Media-Frames-dataset is about ten times larger than the size of the Webis-dataset.

4.1 Clustering of user labels

We rely on a clustering approach to group the issue-specific frame labels into larger clusters that generalize better and abstract from the single user- and issue-specific labels. By this, we overcome the lack of coverage and brittleness of framesets with a fixed set of frames. We developed an algorithm to encode the specific characteristics of such frame labels with a small variable number of tokens, depicted in Figure 1. This algorithm is based on three main steps to encode its input, for example “good economics”. In the **first step**, we apply standard preprocessing steps including POS-tagging and removal of stop words like “the”. Then, our *semantic weighting* approach aims to capture one main characteristic of such labels by determining the semantic head, relying on the POS-tags, for a more reliable grouping in the clustering. For example, considering the labels “good economics” and “bad economics”, the head is the noun ‘economics’. Some user labels are noun-compounds such as “adult health” where we consider the last noun as head and all preceding words as modifiers of the head as we can rephrase the example to “health of adults”. Furthermore, to ensure an equal length n , we repeat tokens in shorter labels. We assume a fixed size of 4 words. We induce a representation for the user-specific label that relies on embeddings for each word, giving highest relevance to the embedding of the head word.

² Unfortunately, the provided boundaries of the annotated text spans do not follow a fixed standard. Even sometimes the boundaries were wrongly shifted by one character since some boundaries cut the first or last part of a word away. Hence, we implemented a bunch of heuristics to correct misleading text boundaries and to have a fixed standard without preceding/succeeding spaces. This increases the amount of high-agreement-subset-samples by $\approx 7,500$ samples.

Figure 1: Sketch of the three steps of our *semantic weighting* approach for embedding the user labels. Due to representation purposes, we consider only word embeddings of dimensionality 12 in this picture.

Hence, in the **second step**, we map each reordered word $i \in \{1, \dots, n\}$ to its word embedding $\vec{v}_i \in \mathbb{R}^M$. Furthermore, we map these word embeddings $\vec{v}_i = (v_{i_1}, \dots, v_{i_M})^T$ to increasingly compressed order-weighted vectors $\vec{v}_i' = (\frac{\sum_{j=1}^i v_{ij}}{i}, \frac{\sum_{j=i+1}^{2i} v_{ij}}{i}, \dots, \frac{\sum_{j=M-i+1}^M v_{ij}}{i})^T \in \mathbb{R}^{\frac{M}{i}}$ by average-pooling all disjoint word windows of size i . The **third step** then concatenates all n compressed vectors $\vec{v}_i'$, resulting in a final label embedding vector of length $\sum_{i=1}^n \frac{M}{i}$.

We apply a k -means algorithm using the cosine distance to cluster the final label vectors into k clusters. As an alternative to the random seeding of the clusters in the k -means algorithm, we also explore the option that the clusters are seeded with the 15 categories from the Media-Frames-set. The algorithm then selects the closest Media-Frame class for each user-provided and issue-specific label by using the Word-Mover's-Distance [16].

4.2 Classification model

For the classification of argumentative texts into frames, we rely on a similar architecture as proposed by Naderi and Hirst [20]. We rely on a recurrent network that processes embeddings of tokens from the premises and con-

Figure 2: Architecture of our model including its input and output. In the multi-task setting, we share the weights of the recurrent layer across the tasks.

clusion as input, using padding to ensure that all inputs are of equal length. The hidden recurrent layer consists of LSTMs [14] or GRUs [7] and processes the input in a bidirectional fashion. On top of that, we put a fully-connected feed-forward dense output layer which outputs a probability for each class based on the final states of the recurrent layer. The architecture is depicted in Figure 2.

Assuming C is our set of frame classes, y the one-hot-encoded ground truth vector and $\hat{y}$ the predicted vector containing the probability distribution of the frame classes, we consider the categorical cross-entropy as loss, which is defined as follows:

$$\text{loss} = - \sum_i^C y_i \cdot \log(\hat{y}_i) \quad (1)$$

Multi-task settings

As our goal is to show that the multi-tasks/multi-dataset-learning outperforms the single-task/single-dataset-learning, we experiment with a multi-task setting in which one model is trained to perform both tasks, i. e. prediction of one of 15 frame types on the Media-Frames-dataset and predicting one of k clusters on the Webis-dataset. For the multi-task setting architecture, we thus have two output layers.

We consider two different settings described in the following:

- In the **hard-parameter sharing** mode, both tasks share the same recurrent layer and thus the same parametrization. This type of architecture is widely spread in the field of parameter sharing due to being rather straightforward to implement.
- We consider a **soft-parameter sharing** configuration as an alternative. It has been shown in previous work that an architecture based on soft-parameter sharing can deliver better results than a hard-parameter sharing architecture [33]. In this configuration, each task has its own recurrent layer, but the parameters are softly shared in the sense that the loss objective penalizes variances in the weights of the layers across both task-specific layers. We implement such a penalization for the weight matrices w and the recurrent neural layer l_M on the Media-Frames-task and l_W on the Webis-task, respectively. The modified loss is defined as follows where the λ -parameter controls how strictly the variance in parameters across layers is penalized [10] for each side (Media+Webis):

$$\text{loss}_{\text{soft}} = \text{loss} + \lambda \cdot \sum_{w_1, w_2 \in (l_M, l_W)} \left(\sum_i \sum_j (w_{1ij} - w_{2ij})^2 \right) \quad (2)$$

Overcoming dataset imbalance

Since the Webis-dataset is more than ten times smaller than the Media-Frames-dataset, there is a huge imbalance in the data. An option would be to downsample the Media-Frames-dataset or to upsample the Webis-dataset. As removing data is not a good option, we pursue the upsampling alternative. This leads to creating many duplicates of each sample in the Webis-dataset, so that each sample is processed multiple times in each training epoch in terms of gradient descent computation and thus in updating the parameters of the weight matrices. As this might lead to overfitting towards the examples of the Webis-dataset, we apply an alternative loss that additionally weighs the losses on both datasets as follows:

$$\text{loss}_{\text{total}} = \alpha \cdot \text{loss}_{\text{Media}} + \beta \cdot \text{loss}_{\text{Webis}} \quad (3)$$

We set the α and β values as follows:

$$\begin{aligned} \alpha &= 1 \\ \beta &= \frac{\# \text{Webis samples}}{\# \text{Media samples}} \end{aligned} \quad (4)$$

5 Results and discussion

In this section, we report the results of our experiments in the *single-task setting* (STS) where we predict the appropriate frame class for the Media-Frames-dataset and

Webis-dataset independently. Further, we provide the results of the *multi-task setting* (MTS) where we train a model with shared parameters (soft) or even with a common layer (hard) to yield a model that predicts both Media-Frames and frame clusters.

We rely on pre-trained 300-dimensional GloVe [23] embeddings³ to embed the input tokens and the user-label-tokens as frame labels in the Webis-dataset for further processing. Regarding the clustering, we experiment with different number of clusters $k \in \{3, 5, 10, 15, 25, 100\}$. We perform multiple runs of the k -means algorithm for each k with at least 10 iterations and a fixed random initialization. Note that the centroids are not further modified after running the k -means algorithm through the training data partition. During test, the centroids remain fixed and the test data samples are assigned to the corresponding centroid. The classification results averaged over all k -means runs are given in Tables 1 and 3–7.

In all cases, the representation learning layers are implemented by a recurrent neural network⁴ with either LSTM or GRU layers with 128 hidden units. Our preliminary experiments showed⁵ that LSTMs perform best for smaller cluster sizes so that we use a configuration where LSTMs are used for cluster sizes of $k \leq 10$ and GRUs for settings with $k \geq 15$. For training, we rely on the Adam optimizer with a batch size of 64 for the single-task setting and 32 for the multi-task setting since the latter is computationally more intensive. We apply early-stopping [6] with patience of 2 with maximally twelve epochs for parameter optimization. We do not shuffle the samples of the topic-ordered datasets and test the aspect of generalizability with a train-validation-test-split of 80-10-10 %, respectively.

5.1 Single-task setting (STS)

In this section we present the results of our experiments conducted with independent models trained for each task.

5.1.1 Media-Frames-dataset

In this section we show our results regarding the classification of the Media-Frames-dataset. We report results showing the impact of our strategies for pre-processing the data

³ <http://nlp.stanford.edu/data/glove.840B.300d.zip>

⁴ <https://tensorflow.org/> – version 2 (Keras API).

⁵ This observation coincides with the results by Naderi and Hirst [20].

Table 1: This table shows the results (accuracy) on the Media-Frames-dataset for the most frequent Media-frames-classes (acc(5)) and all Media-frames-classes (acc(15)). Results marked with * consider the additional annotation class “Irrelevant”, too. Besides the outperforming neural net with 128 GRUs by Naderi and Hirst [20] processing whole sentences (baseline), the table shows the results for our settings, too.

Approach	accuracy	
	5 classes	15 classes
only immigration topic		
GRU-Baseline [20]	70.7	57.1
noArg-mode	65.5	50.9
narrow-context-mode	72.0	59.0
broad-context-mode	66.7	51.7
immigration+smoking topic		
GRU-Baseline [20]	68.7	53.7*
noArg-mode	62.3	50.1
narrow-context-mode	72.2	61.2
broad-context-mode	62.3	50.3
all 3 topics		
GRU-Baseline [20]	n. a.	n. a.
noArg-mode	62.4	50.6
narrow-context-mode	73.2	62.3
broad-context-mode	62.9	50.9
all 3 topics, high inter-annotator-agreement-subset		
GRU-Baseline [20]	n. a.	n. a.
noArg-mode	71.7	57.5
narrow-context-mode	78.3	63.6
broad-context-mode	70.9	58.5

(*noArg* vs. *narrow-context* vs. *broad-context*, described in Section 3.2) as well as the impact of training on subsets of the data of different quality (single-topic vs multiple-topics with different inter-annotator-agreement-subsets). To contextualize our results, we compare our single-task-results to the model proposed by Naderi and Hirst [20], using the same bidirectional recurrent architecture with GRUs. We present the results in Table 1. We see that accuracy levels range from 62.3 % to 78.3 % when only considering the five most frequent Media-frames classes and from 50.6 % to 63.6 % when considering all the 15 Media-Frames classes including the “Other”-class but excluding the outside “Irrelevant”-class. The best performing setting in these experiments is the *narrow-context-mode*, yielding improvements by up to 3.5 accuracy points compared to the results reported by Naderi and Hirst [20]. Our results license three observations:

1. The *narrow-context-mode* outperforms the other modes (*noArg* and *narrow-context*) due to the proper balance of inferring cues from the context without get-

ting distracted by too much not-frame-related content from the broader context.

2. Naderi and Hirst [20] observed that performance of their approach deteriorates when considering more topics, that is three topics instead of one. They report best results when training their classifier on a single topic only. In contrast, our results show that there is indeed a positive effect of training the model on multiple topics, which we hypothesize is due to the larger sample of training data including our preprocessing, which seems to effectively counteract the potential risk of confusion by the classifier when working with heterogeneous topics.
3. Although the filter regarding high inter-annotator-agreement samples discards $\approx 85\%$ of all samples, we observe higher accuracies with this small subset. We gain up to 9.6 percent points in comparison to Naderi and Hirst [20]. The quality of the annotations is thus critical and has more impact than the absolute size of the dataset.

Building on these results, in the further experiments described in this paper, we rely on the high-inter-annotator agreement subset of the Media-Frames-dataset for training. We also adopt the *narrow-context-mode*.

5.1.2 Webis-dataset

In this section we present the results on the single task consisting of predicting clusters of frames for different numbers of clusters k . The classification accuracies averaged over all the runs for a given cluster size k are shown in Figure 3. We provide insights into the manual quality assessment of the clusters produced by the clustering algorithm. We also investigate the impact of our *semantic weighting* approach, described in Section 4.1.

The first observation is that the accuracies decrease with increasing number of the clusters. This is as expected as the number of class labels corresponds to the number k of clusters. The accuracy starts at 66.7 % at $k = 3$ and drops to 19.1 % at $k = 100$. In general, we observe a remarkable accuracy bandwidth for each cluster size k . This variation is explained by the fact that each run of the k -means algorithms is initialized with k randomly selected seeds and converges mostly to local optima determined by the random seeds [28]. The variance is obviously higher for smaller ks .

In order to get insights into the nature of the generated clusters, we manually investigate the clusters produced by a run of the k -means algorithm with 15 means. We notice

Figure 3: Results of the Webis frame cluster prediction task over the number k of clusters. The points show the mean, the bars represent the range of values for the 12 k -means runs.

that the number of labels assigned to each cluster displays a high variation, showing that there are frames that are more prominent than others. The variation reaches from slightly above 0 % of the labels assigned through to approx. 11 % of labels assigned. One cluster is an artefact of our method, consisting of all the words which have no embeddings and are thus assigned to the embedding for the ‘unknown’ word. Many abbreviations are in this cluster. Table 2 shows 15 clusters with the five labels that are closest to the corresponding centroid for a particular run of the k -means algorithm.

A number of clusters could be identified as corresponding to the Media-Frames by Boydston et al. [3]. One cluster for instance includes topics related to *morality* and *ideology* and corresponds to the Media-Frame (iii) *morality*. Another cluster contains many expressions related to *law* and *authority*, corresponding to Media-Frame (v) *legality, constitutionality and jurisprudence*. In addition, the clustering comprises of clusters that do not correspond to any Media-Frame. For example, one cluster contains labels related to *ecology* and *renewable energy*. This is a good example of how the bottom-up, clustering-based approach we have proposed can extend traditional frame category systems by emerging topics and alternative aspects. Our approach can in particular also contribute to accommodating shifts of concerns and aspects of a discussion over time [3].

We also test the impact of the *semantic weighting* approach (Section 4.1) by skipping the sub-step of the semantic reordering. The results of this ‘ablation’ is shown in Table 7. Except for two border cases, the results show the positive impact of the semantic clustering by increasing the accuracy results up to 8.6 % points.

Table 2: Top-5 nearest issue-specific labels for each of 15 centroids induced by one run of the k -means algorithm. Overlapping user labels are aggregated in this table by using brackets (e. g. “(getting/ purpose of) marriage” implies the user labels “marriage”, “getting marriage” and “purpose of marriage”).

≥ 5 nearest labels to centroid	
0	child-hazards, food-source, us-specific, moon-to-mars, eezs
1	(saving) lives , (dignity / debaters attitude towards/ quality of) life, (getting/ purpose of) marriage, parents
2	morality , ideology, individualism, fundamentalism, secularism
3	(impact of/ family/ reducing sectarian/ anti-gay) violence , (sexual/ child) abuse, war and crimes, crime
4	(economic/ Kosovo) viability , impacts, (economic) feasibility, effectiveness
5	social (aspects), society, importance, (political) education, interests
6	(wind/ net) energy, natural gas, (reducing) emissions , renewable, water supply
7	(global/ us) economy , economic growth, governance, (health of) concerns
8	(regime) change , (regional) impact, characteristics, individual
9	(rule/ us) law , laws, precedent, authority, legal
10	(public/ safety/ adult) health , (health) care, (international) security
11	russia, iran, iraq, israel, europa
12	pay, (state/ benefits of assessing) costs (of healthcare), taxes
13	(pretext) war , military, (military/ terrorist) threat
14	social democratic, (economics/ direct) democracy , (political/ state) rights, government

Fitting the Media-Frames-set with our clustering approach

In a further set of experiments, we tested if we can fit the clusters to the Media-Frames. For this, instead of relying on a randomized seed, we fix the centroid of each of the k -means to each of the Media-Frame expressions and calculate the nearest pre-specified centroid to each user-label by the Word-Movers-Distance [16]. We do this for the 14 frames of the Media-Frames-set, excluding the open ‘Other’ category and perform with 32.9 % accuracy. We then add a 15-th cluster corresponding to the ‘Other’ category to which we assign all labels having a distance greater than $\frac{\max(d) - \min(d)}{2} \cdot \text{len}(d)$ to all other centroids, where d is a distance-vector which contains all distances to the 14 Non-Other-frame-classes. In this setting, we get a classification accuracy of 31.8 %, which is higher than the averaged accuracy reached with the randomized seeds (30.7 %). This corroborates the fact that the Media-Frames-set is semantically well-defined, as it facilitates the task of classifying arguments into frames in our clustering setting. However, in comparison to the whole bandwidth of the random-seeded clusters, we observe also some outlier-

Table 3: This table shows the results (accuracy) on the high-agreement subset of the Media-Frames-dataset. For each cell, the left number shows the accuracy of the Media-Frames-dataset-task (15 frames) and the right number shows the accuracy of the Webis-task (k -means-clustering). All results are averaged over the 12 different runs. The first row shows results for the single task setting for reference. ‘soft’ represent the soft-parameter-sharing and ‘hard’ represents the hard-parameter-sharing approach.

	λ	3	5	10	15	25	100
STS	n. a.	63.6/66.7	63.6/55.6	63.6/36.9	63.6/30.7	63.6/26.7	63.6/19.1
soft	0.01	67.4/67.3	67.7/ 57.3	66.3/39.1	67.4/33.3	67.8/28.3	68.7 /20.8
soft	0.05	67.4/67.6	65.4/56.3	66.9/38.4	66.8/33.4	66.7/28.7	65.7/20.7
soft	0.1	65.2/67.7	66.1/57.0	65.6/38.4	65.8/33.6	65.4/28.5	66.0/20.9
soft	0.5	64.3/ 68.2	65.7/56.1	64.7/ 40.0	66.2/ 34.2	65.3/ 28.8	65.6/20.8
hard	n. a.	67.7/66.6	65.9/56.4	67.9/38.8	67.5/33.3	68.3/28.8	68.0/ 21.0

forming random-seeded clusters which capture recent topics not covered by the Media-Frames. When restricting ourselves to the five most frequent Media-Frames as in Naderi and Hirst [20], using the above procedure we get an accuracy of 45.9 %.

5.2 Multi-task setting (MTS)

In this section, we present the results of the multi-task setting where we have a shared model for both tasks. We investigate in particular the impact of relying on hard vs. soft-parameter sharing between the recurrent layers for each classification task. Table 3 provides the results for the subset of the high-inter-annotator-agreement subset of the Media-Frames-dataset. The table plots the single-task-baseline together with the results of the soft-parameter sharing approaches with $\lambda \in \{0.01, 0.05, 0.1, 0.5\}$ in addition to the result of the hard parameter sharing approach – once for the Media-Frames classification task, and once for the Webis-task. As in the single-task setting, the number of clusters ranges within $k \in \{3, 5, 10, 15, 25, 100\}$. This number impacts the prediction and results strongly on the Webis-task and influences only indirectly the Media-Frames-task via the shared parameters. We can observe improvements resulting from the multi-task learning approach for all cluster sizes on both tasks. An interesting observation is that the most substantial relative improvement of +11 % accuracy (up to +3.5 absolute improvement) compared to the single-task setting on the Webis-dataset are obtained when the number of clusters ($k = 15$) is comparable to the number of frames in the Media-Frames-set. In spite of the fact that the frame sets differ, it is remarkable that one task can clearly profit from the other. In most cases, a soft-parameter-sharing with $\lambda = 0.5$ achieves the best results, although the results show a quite high variance. Another interesting observation is that the results using a large number of clusters tend to improve more with

Table 4: Table showing the accuracies in % of applying the small portion of the Webis-dataset (covers only the three Media-Frames-topics). The results are arithmetic means out of twelve runs on the Webis-task, having a large deviation ($\pm \approx 50\%$). The columns represent the cluster size k . 100 clusters are not applicable since there are only 73 distinct user labels in the training data partition. The first row represents the results from the full Webis-dataset for reference. The second represents the single-task mode and the last two the multi-task mode with soft- ($\lambda = 0.1$) and with hard-parameter-sharing, respectively, using the high-agreement subset of the Media-Frames-dataset.

	3	5	10	15	25
STS-full	66.7	55.6	36.9	30.7	26.7
STS-3topics	56.6	42.7	35.1	25.5	20.5
MTS-S-3topics	66.7	58.7	43.8	36.5	27.6
MTS-H-3topics	63.6	55.7	43.4	38.7	30.4

a higher λ or even with the hard-parameter-sharing with $k = 100$. We hypothesize that this is due to the fact that, as the number of clusters increase, there are less datapoints per cluster. In contrast, the number of datapoints per frame is constant for the Media Fames dataset, leading to more stable results.

The improvements on the Media-Frames-task ranges from +0.7 to +5.1 percent points with respect to the STS, applying the soft-parameter-sharing-learning ($\lambda = 0.01$) with the fine-granular frame cluster structure ($k = 100$) of the Webis-Frames, supporting the thesis that both tasks can benefit from each other. It seems that relaxing the parameter sharing improves results with increasing value k . This can be explained by the fact that the fewer the clusters we consider, the more distinct the two classification tasks become, so that a hard-parameter sharing is detrimental. Overall, our results clearly license the conclusion that the multi-task setting is beneficial.

In a further set of experiments (see Table 4), we evaluate the impact of the ability of our model to generalize across the different topic distributions between the

two tasks. We manually filtered the Webis-dataset for related topics to the three topics which occur in the Media-Frames-dataset, resulting in only 473 arguments. Although we show that frame prediction can be generalized between different topics to some extent, we observe frame patterns which tend to be topic-specific and, hence, should be trained on specific topics. Thus, restricting the topics leads to better results in the Webis-task in combination with the massive amount of training samples of the Media-Frames-dataset, yielding a highly remarkable result at $k = 15$. Here we observe an accuracy of 38.7 % in the hard-parameter-sharing-mode, which is 13.2 % percent points higher than the results of the single-task setting. Besides, this small portion increases the accuracy up to 68.3 % in the hard-parameter-sharing-case on the Media-Frames-task for $k = 15$. The performance gain clearly shows the positive impact of learning joint representations across both tasks, in spite of the fact that there are substantial differences in the number and structure of topics across both datasets.

Impact of training with a less quality controlled portion of the Media-Frames-dataset

Table 5 shows the results of training with the less quality-controlled subset of the Media-Frame-dataset in the multi-task-setting. In this dataset, some arguments have only been annotated by one annotator (especially for the smoking topic) and this set thus features a lower reliability. In this setting the number of training data is increased by one order of magnitude, that is from 12,326 to 139,304 training samples. It is thus interesting to investigate the tradeoff between training with a small set of reliable training data examples compared to training with a much larger but less quality controlled dataset. The results show that, while there is always an improvement by parameter sharing on the high-agreement subset of the Media-Frames-dataset (see Table 3), in this case the multi-task setting with hard-parameter sharing has inferior performance compared to the STS results, especially on the Media-Frames-dataset. One reason for this might be the smaller reliability of the annotations in the dataset because of controversial parts which might not belong to any or several classes of the Media-Frames-set. Thus, the incorporation of such unreliable examples harms the performance of the model.

However, even in this scenario, the performance improves with respect to the single-task setting (an exception are the results for cluster sizes $k = 3$ and $k = 5$) by between 0.4 accuracy points at $k = 10$ and 1.8 accuracy points at $k = 25$ on the Webis-dataset using soft-parameter-sharing. Our interpretation here is that due to the fact that the layers for different tasks are not shared in this setting, the model

Table 5: Table showing the accuracies in % of the STS and MTS, trained with the less quality controlled subset of the Media-Frames-dataset in the **narrow-context** mode. The structure of the table is the same as in Table 3 for the Webis-task. The results are arithmetic means out of six runs instead of twelve due to the high computational cost ($\approx$ two hours per run with 24xIntel-CPU-E5-2620 and 256 GB RAM). The deviation among the results of the single runs are similar to the deviations showed in Figure 3 concerning the k .

	λ	3	5	10	15	25	100
STS	n. a.	66.7	55.6	36.9	30.7	26.7	19.1
soft	0.01	65.8	51.2	37.2	31.0	27.8	20.4
soft	0.1	65.4	51.1	37.3	32.0	28.5	20.3
hard	n.a	64.4	51.0	36.8	32.0	27.6	19.3

can optimize parameters on the Webis-task without being too strongly affected by the lack of annotation consistency in the Media-Frames-dataset. This setting seems to tolerate noise in the training data better. In fact, in the case of soft-sharing it seems that due to the higher flexibility that comes with optimizing the two layers for two different tasks, the classification on the Webis-dataset can still profit from the multi-task regime in spite of a less quality controlled dataset. This shows also that the multi-task setting with soft-parameter sharing is generally more robust.

Revisiting the Media-Frames-set with our clustering approach

As an additional analysis, to provide another approach bridging the gap between generic frames and issue-specific user-labels, we applied the multi-task setting mapping the issue-specific user-labels to the Media-Frames-seeded clusters. Interestingly, when considering only the 5 most prominent topics from the Media-Frames-dataset, the number of data samples (13,094) become comparable to the ones from the Webis-dataset (12,326).

The results are listed in Table 6. We observe improvements by relying on the high-inter-annotator-agreement-Media-Frames-subset. With $\lambda = 0.05$ in the soft-parameter-sharing-mode, we ensure a stable accuracy gain of 9.9 % with the top-5 frames in the Media-Frames-dataset. This gain is the most substantial increase observed so far on the classification on the Media-Frames-dataset compared to the single-task-setting. Besides this outstanding case, we observe improvements in all other cases. We note that the highest accuracy is achieved for all 15 generic frames with 68.4 % (+4.8 % with $\lambda = 0.1$) on the Media-frames dataset. The Webis-task gains 4.1 % and 2.8 % point for 5 and 15 frames, respectively. These results clearly corroborate the impact of sharing parameters between the two

Table 6: Table showing the accuracies in % of the STS and MTS grounded on the Media-Frames, using the high-agreement subset of the Media-Frames-dataset. All results are averaged over 12 different runs of Word-movers-distance-clustering (Media-Frames) and training on the Webis-dataset. The structure of the table is the same as in Table 3.

	λ	most-freq. (5)	all frames (15)
STS	n/a	78.3 / 45.9	63.6 / 31.8
soft	0.01	86.6 / 46.6	68.3 / 34.6
soft	0.05	88.2 / 48.1	66.6 / 34.0
soft	0.1	87.5 / 47.4	68.4 / 34.0
soft	0.5	87.8 / 48.6	68.3 / 32.8
hard	n.a	87.9 / 50.0	67.1 / 32.6

Table 7: Table showing the deviation of the results of the non-semantic clustering approach from the mode of a semantic weighting on the user labels in %-accuracy-points. All results are averaged over the 12 different runs of k -means-clustering and training. The first row shows results for the single task setting, the others the multi-task setting on the high-agreement subset of the Media-Frames-dataset. ‘soft’ represent the soft-parameter-sharing, and ‘hard’ represents the hard-parameter-sharing approach.

	λ	3	5	10	15	25	100
STS	n. a.	-2.3	-4.1	-2.4	-1.1	-1.3	-3.3
soft	0.01	-0.1	-2.7	-1.2	-0.2	-2.8	-1.2
soft	0.05	-4.0	-5.6	-0.8	-4.3	-4.8	-0.1
soft	0.1	-8.6	-4.8	-1.8	-0.8	-1.6	-0.3
soft	0.5	-7.5	-5.0	+0.5	-0.4	-3.8	-1.6
hard	n. a.	+2.4	-3.4	-1.7	-1.1	-4.7	0.0

tasks and the ability to close the gap between the classical Media-Frames and issue-specific frames in addition to the transferability of our multi-task-approach to generalize from three to 465 topics by sharing parameters (to a certain degree).

6 Conclusion

In this paper we have attempted to address the limitations of current frame classification approaches that rely on fixed and generic frame classes and attempted to bridge the gap between very generic frame sets on the one hand, and overly specific frames as found in user-provided labels in current online argumentation portals on the other. Towards finding a bridge, namely a middle-ground between these generic frame categories and the overly issue- and user-specific categories in the Webis-dataset, we have proposed an approach to classify frames at various level

of granularities, introducing a clustering approach that groups the user-generated labels into more general frame categories and that allows to choose the level of granularity of the frames considered. A manual evaluation has shown that the clusters seem reasonable, in some cases corresponding to frames in the Media-Frames-set, in other cases representing emerging aspects of online discussions that are not included in the Media-Frames inventory. We have presented results of a standard recurrent architecture for different cluster sizes. As main result, we have shown that a multi-task setting in which one model is trained to identify both our cluster frames as well as the Media-Frames in parallel can improve classification results on both tasks. In particular, our results shows the highest performance gain is observable when using a similar amount of frame classes on both tasks by applying a soft-parameter-sharing in the multi-task setting. While we have applied some simple heuristics to infer argument structure, we expect to improve results by using state-of-the-art architectures for argument unit analysis.

Future work should be devoted to understand better how our clusters are aligned with the Media-Frame classes. Settings in which arguments can be mapped to different frames to a certain degree rather than in a crisp fashion should also be investigated. Finally, an interesting avenue for future work lies in the inclusion of topic-specific background or common-sense knowledge with the goal of increasing frame detection accuracy.

References

1. Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. Modeling frames in argumentation. In *Proc. of Conference on Empirical Methods in Natural Language Processing and Intl. Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932. ACL, 2019.
2. Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proc. of First Workshop on Argumentation Mining*, pages 49–58. ACL, 2014.
3. Amber E. Boydston, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. Tracking the development of media frames within and across policy issues. *APSA Annual Meeting Paper*, 2014.
4. Elena Cabrio and Serena Villata. Five years of argument mining: a data-driven analysis. In *Proc. of Twenty-Seventh Intl. Joint Conference on Artificial Intelligence (IJCAI)*, pages 5427–5433. IJCAI, 2018.
5. Dallas Card, Amber E. Boydston, Justin H. Gross, Philip Resnik, and Noah A. Smith. The media frames corpus: Annotations of frames across issues. In *Proc. of 53rd Annual Meeting of Association for Computational Linguistics and 7th Intl. Joint Conference on Natural Language Processing (ACL-IJCNLP) (Volume 2)*, pages 438–444. ACL, 2015.

6. Rich Caruana, Steve Lawrence, and Lee Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Proc. of the 13th Intl. Conference on Neural Information Processing Systems (ICONIP) NIPS' 00*, pages 381–387. MIT Press, 2000.
7. Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. ACL, 2014.
8. Claes de Vreese. News framing: Theory and typology. *Information Design Journal*, 13, 2005.
9. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (NAACL)*. ACL, 2019.
10. Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proc. of 53rd ACL and the 7th IJCNLP (Volume 2)*, pages 845–850. ACL, 2015.
11. Robert M. Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43 (4), 1993.
12. Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3570–3580. ACL, 2018.
13. Michael Fromm, Evgeniy Faerman, and Thomas Seidl. Tacam: Topic and context aware argument mining. In *Proc. of IEEE/WIC/ACM Intl. Conference on Web Intelligence, WI '19*, pages 99–106. ACL, 2019. ISBN 9781450369343.
14. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8): 1735–1780, 1997. ISSN 0899-7667. doi:10.1162/neco.1997.9.8.1735.
15. N. Jin, J. Wu, X. Ma, K. Yan, and Y. Mo. Multi-task learning model based on multi-scale cnn and lstm for sentiment classification. *IEEE Access*, 8: 77060–77072, 2020. ISSN 2169-3536.
16. Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In Francis R. Bach and David M. Blei, editors, *Proc. of 32nd Intl. Conference on Machine Learning, (ICML)*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org, 2015.
17. Mirko Lenz, Premtim Sahitaj, Sean Kallenberg, Christopher Coors, Lorik Dumani, Ralf Schenkel, and Ralph Bergmann. Towards an argument mining pipeline transforming texts to argument graphs. *CoRR*, 2006.04562, 2020.
18. Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *Proc. of 29th British Machine Vision Conference (BMVC)*, pages 1–13. British Machine Vision Association, 2018.
19. Nona Naderi and Graeme Hirst. Argumentation mining in parliamentary discourse. In Matteo Baldoni, Cristina Baroglio, Floris Bex, Floriana Grasso, Nancy Green, Mohammad-Reza Namazi-Rad, Masayuki Numao, and Merlin Teodosia Suarez, editors, *Principles and Practice of Multi-Agent Systems*, pages 16–25. Springer Intl. Publishing, 2016. ISBN 978-3-319-46218-9.
20. Nona Naderi and Graeme Hirst. Classifying frames at the sentence level in news articles. In *Proc. of the Intl. Conference Recent Advances in Natural Language Processing, (RANLP)*, pages 536–542. INCOMA Ltd., 2017.
21. W. Russell Neuman, Russell W. Neuman, Marion R. Just, and Ann N. Crigler. *Common Knowledge: News and the Construction of Political Meaning*. American Politics and Political Economy. University of Chicago Press, 1992. ISBN 0226574407.
22. Michael Paul and Roxana Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proc. of 24th AAAI Conference on Artificial Intelligence, AAAI' 10*, pages 545–550. AAAI Press, 2010.
23. Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. ACL, 2014.
24. Sebastian Ruder. An overview of multi-task learning in deep neural networks. *ArXiv*, 2017.
25. Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Aspect-controlled neural argument generation. *ArXiv*, 2005.00084, 2020.
26. Holli A. Semetko and Patti M. Valkenburg Valkenburg. Framing European politics: A content analysis of press and television news. *Journal of Communication*, 50 (2), 2006.
27. Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. ArgumenText: Searching for arguments in heterogeneous sources. In *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL)*, pages 21–25. ACL, 2018.
28. Douglas Steinley and Michael J. Brusco. Initializing k-means batch clustering: A critical evaluation of several techniques. *Journal of Classification*, 24 (1): 99–121, 2007.
29. Tatiana Tommasi, Novi Quadrianto, Barbara Caputo, and Christoph H. Lampert. Beyond dataset bias: Multi-task unaligned shared knowledge transfer. In Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu, editors, *Computer Vision – ACCV 2012*, pages 1–15. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-37331-2.
30. Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an argument search engine for the Web. In Kevin Ashley, Claire Cardie, Nancy Green, Iryna Gurevych, Ivan Habernal, Diane Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern Walker, editors, *Proc. of 4th Workshop on Argument Mining*, pages 49–59. ACL, 2017a.
31. Henning Wachsmuth, Benno Stein, and Yamen Ajjour. “PageRank” for argument relevance. In *Proc. of the 15th Conference of the European Chapter of the ACLs, Volume 1 (EACL)*, pages 1117–1127. ACL, 2017b.
32. Hanqian Wu, Siliang Cheng, Zhike Wang, Shangbin Zhang, and Feng Yuan. Multi-task learning based on question–answering style reviews for aspect category classification and aspect term extraction on GPU clusters. *Cluster Computing*, 23 (3): 1973–1986, 2020.
33. Yongxin Yang and Timothy M. Hospedales. Deep multi-task representation learning: A tensor factorisation approach. In *Proc. of 5th Intl. Conference on Learning Representations, (ICLR)*. OpenReview.net, 2017.

Bionotes

Philipp Heinisch
Universität Bielefeld, Postfach 10 01 31,
33501 Bielefeld, Germany
pheinisch@techfak.uni-bielefeld.de

Philipp Heinisch, born in 1994, studied Computer Science at the Paderborn University (Master, 2019) and is a Ph.D. student in the semantic computing group of Philipp Cimiano. His main research interests include natural language processing with deep learning techniques in the field of argumentation.

Philipp Cimiano
Universität Bielefeld, Postfach 10 01 31,
33501 Bielefeld, Germany
cimiano@techfak.uni-bielefeld.de

Prof. Dr. Philipp Cimiano studied computer science and computational linguistics at Stuttgart University. He obtained his PhD and habilitation in Applied Computer Science from the University of Karlsruhe. He is full professor for computer science at Bielefeld University since 2009. His main research topics include natural language processing, text mining, knowledge engineering and management, question answering, linked data and knowledge representation. He was nominated as one of top 10 researchers to watch for the future of AI by the IEEE Intelligent Systems Magazine. He is editorial board member of the Semantic Web Journal, the Journal of Web Semantics, the Semantic Computing Journal and the Journal of Applied Ontology.