Data Augmentation for Improving the Prediction of Validity and Novelty of Argumentative Conclusions

Philipp Heinisch
Bielefeld University
pheinisch@techfak.uni-bielefeld.de

Moritz Plenz
Heidelberg University
plenz@cl.uni-heidelberg.de

Juri Opitz
Heidelberg University
opitz@cl.uni-heidelberg.de

Anette Frank
Heidelberg University
frank@cl.uni-heidelberg.de

Philipp Cimiano
Bielefeld University
cimiano@techfak.uni-bielefeld.de

Abstract

We address the problem of automatically predicting the quality of a conclusion given a set of (textual) premises of an argument, focusing in particular on the task of predicting the validity and novelty of the argumentative conclusion. We propose a multi-task approach that jointly predicts the validity and novelty of the textual conclusion, relying on pre-trained language models fine-tuned on the task. As training data for this task is scarce and costly to obtain, we experimentally investigate the impact of data augmentation approaches for improving the accuracy of prediction compared to a baseline that relies on task-specific data only. We consider the generation of synthetic data as well as the integration of datasets from related argument tasks. We show that especially our synthetic data, combined with class-balancing and instance-specific learning rates, substantially improves classification results (+15.1 points in F_1-score). Using only training data retrieved from related datasets by automatically labeling them for validity and novelty, combined with synthetic data, outperforms the baseline by 11.5 points in F_1-score.

1 Introduction

Recently, there has been interest in developing approaches that can automatically generate conclusions from textual premises (Syed et al., 2021; Heinisch et al., 2022a). Many of these systems rely on language models that are fine-tuned to the task of generating argument conclusions. As the space of possible conclusions that can be generated from a textual premise is a priori not constrained, it is key for a system to understand whether a conclusion candidate is adequate. In particular, models that can predict the quality of conclusions are needed to guide a generation system towards generating suitable argumentation conclusions.

While there has been work on identifying dimensions that characterize argument quality (Wachsmuth et al., 2017b), there are very few models that actually operationalize the (automatic) scoring of the quality of a conclusion. Gurcke et al. (2021) have analyzed whether the notion of "sufficiency" of an argument can be predicted, reaching an accuracy of 90% with transformer-based language models. Heinisch et al. (2022a) have relied on the notions of "validity" and "novelty" in their manual evaluation of conclusion quality – "validity" meaning that the conclusion is justified based on its premise and "novelty" that the conclusion contains novel content which is related to the premise. They have shown that there is a weak correlation between the automatically computed similarity between generated conclusion and reference conclusion, as measured by the BERTscore, on the one hand, and the criteria of manually rated validity and novelty on the other hand. One key problem is that it is difficult to obtain sufficient training data for such tasks, which is a necessary basis for training reliable models for these tasks.

In this paper, we focus on predicting the validity and novelty of argument conclusions. We propose a multi-task classification approach that jointly predicts validity and novelty in a single model that exploits synergies between both tasks.

Our main goal is to explore to what extent data augmentation can contribute to overcome the scarcity of manually labeled argument quality data. We propose and experimentally investigate two types of approaches. On the one hand, we investigate the impact of a synthetic data generation approach that modifies existing training data by generating 'altered copies' of its instances, e.g., by shifting or extending content between premise and conclusion in view of novelty, or by paraphrasing or negating parts of the argument in view of validity. Further, we augment the data labeled for novelty and validity by considering datasets from related argument mining tasks. In particu-

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 19–33
October 17, 2022.

lar, we consider data from the ExplaGraph-dataset by Saha et al. (2021), the IBM-ArgumentQuality-dataset by Gretz et al. (2020b) and the Student-Essays-dataset, annotated for sufficiency of the conclusion by Stab and Gurevych (2017b). We describe how training data from these related tasks is mapped into a form that can be used to enhance the performance of our classifier for validity and novelty prediction. We experimentally evaluate the impact of both data augmentation strategies, showing that the generation of synthetic data outperforms a baseline system trained with only task-specific data by 15.1 points in F_1-score (38.3 vs. 23.2). Even when only using datasets from related tasks as training data, we improve results over the baseline by 11.5 points.

Our main contributions are:

- We present an approach for augmenting training data for validity and novelty, by creating synthetically generated instances. We do this by applying systematic transformations to the original, task-specific training data.
- We also explore various datasets in the field of argument mining, and show how to adapt them automatically to the task of validity and novelty prediction – in combination with specific training techniques, such as instance-adaptive learning rates.
- We perform an extensive automatic evaluation study of various combinations of datasets and training dataset sizes in combination with varying ratios of synthetic vs. non-synthetic instances. We obtain comparable classifier performances without even using the explicitly annotated validity-novelty-training split.
- To give further insight into our results, we present a case study that helps to better understand the effects of interleaving datasets, and of our adaptive training process.

2 Related Work

The task of automatic generation of arguments has received increasing attention in the last years (Gretz et al., 2020a; Schiller et al., 2021). In particular, research has considered the generation of a conclusion given a (textual) premise (Syed et al., 2021; Opitz et al., 2021; Heinisch et al., 2022a). These approaches rely on language models that are fine-tuned to the task of conclusion generation. The generation of conclusions can be seen as a search in the output space of a language model conditioned on the textual premise.

In the manual evaluation of approaches generating conclusions, Opitz et al. (2021) and Heinisch et al. (2022a) found that (generated) conclusions are often either not justified given their premise, or are often just a plain copy or paraphrase of the premise, hence lacking novelty. They conclude that validity and novelty are two main properties a conclusion should fulfill and that stand in a trade-off relation to each other.

A key question is thus how to guide the search or generation process towards i) conclusions that represent a legitimate inference from the premises, meaning that the conclusions are *valid*, and ii) conclusions that are not simple paraphrases of the premises, i.e., they are *novel* or *informative*. Having operationalized and thus automatically computable quality dimensions is key to generating high-quality conclusions.

While there is previous work that identifies quality criteria for arguments (Wachsmuth et al., 2017b; Gretz et al., 2020b), it has been shown that the annotation of such quality criteria is highly subjective (Wachsmuth et al., 2017a; Wachsmuth and Werner, 2020). Also, little work has been done on automatically rating the quality criteria for arguments. An exception is work by Gurcke et al. (2021) who – following Stab and Gurevych (2017b) – studied the operationalization of the criterion of sufficiency. Sufficiency measures whether the premises provide enough evidence for accepting or rejecting the conclusion, and is hence a criterion closely related to our notion of *validity*.

In this paper, we are concerned with developing a computational model that can jointly predict the validity and novelty of conclusions. Given that manually annotated data is scarce, relying on the manual studies by Heinisch et al. (2022a), we consider how task-specific datasets can be augmented with synthetic data and how to repurpose data from related argument mining tasks. Our work is thus related to and encouraged by data augmentation approaches in general. One example is the field of code-mixed languages, which often lacks available annotated training data. Here, Pratapa et al. (2018) showed how to create synthetic instances of code-mixing language by merging sentences from different languages with the help of syntactic parse trees. Another task that has been

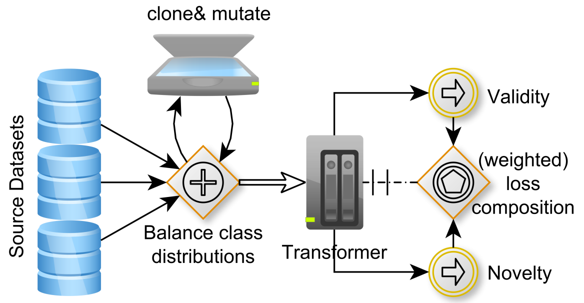

Figure 1: Architecture for validity-novelty multi-task-classification with modulated data augmentation.

shown to profit from automatically generated synthetic training data is grammatical error correction. Here, it has been shown that creating additional training data by corrupting error-free sentences leads to performance gains (Grundkiewicz et al., 2019; Stahlberg and Kumar, 2021). Finally, it has been shown that, by generating synthetic negative instances, one can bootstrap classifiers, e.g., to rate the output of a language model converting knowledge graph triples into natural language (Harkous et al., 2020). Building on prior evidence that generation of synthetic data can improve classifier performance, we investigate a clone&mutate technique that can artificially create new training instances of every class.

3 Methods

In this section, we present our methods for tackling the task of predicting validity and novelty as a classification task. We describe the learning objective and how we generate and modulate additional training data using data augmentation techniques. Figure 1 shows our proposed architecture.

3.1 Learning Objective

We adopt a multi-task classification setting to jointly predict validity and novelty. Inspired by Jin et al. (2020), our loss function includes a combined loss that controls the interaction of the separate individual task losses for novelty and validity, $L_{t_{val}}$ and $L_{t_{nov}}$, which we define by mean squared error. The interaction of the different losses is defined as follows:

$$\mathcal{L} = \alpha \mathcal{L}_{t_{val}} \mathcal{L}_{t_{nov}} + \beta \mathcal{L}_{t_{val}} + \gamma \mathcal{L}_{t_{nov}} \qquad (1)$$

where α, β, γ are scalars > 0.

If the target validity or novelty is unknown for a training instance, the related loss $\mathcal{L}_{t_{val}}$ or $\mathcal{L}_{t_{nov}}$ in Equation 1 is set to 0 to avoid random model weight adjustments.

Extending the loss function - introducing dataset- and instance-specific weights We hypothesize that not all instances have the same relevance for the task at hand, so that the impact of each training instance should not be uniform. Therefore, we introduce a fixed weight w_i for each training instance i that is multiplied with the loss computed for the specific training sample i as follows:

$$\mathcal{L}_i = w_i \left(\alpha \mathcal{L}_{t_{val}} \mathcal{L}_{t_{nov}} + \beta \mathcal{L}_{t_{val}} + \gamma \mathcal{L}_{t_{nov}} \right) \qquad (2)$$

We investigate three approaches for setting the instance weights. First, as a baseline, in the *uniform weighting* setting, we set the weight w_i uniformly to 1 for every instance. For *dataset-specific weighting* we set w_i to a value that is specific for each dataset and apply it to all instances contained therein. Finally, in the *individual weighting* setting, the weight is set individually for each sample.

3.2 Training Data

We explore the impact of using different source datasets as training data in which we represent each instance as a pair of a textual premise p and conclusion c. We test combinations of data having explicit values for validity and novelty, as well as data without such explicit values. We describe the used datasets including the procedures for setting the values for validity v, novelty n and the weights w in Section 4. To resolve the issue of class-imbalance when merging uneven source datasets, we rely on synthetic data generation as described below, to ensure a larger training dataset while maintaining class balance.

Synthetic generation of data: clone&mutate For augmenting the training data, we propose a procedure that selects training instances randomly and applies a clone&mutate operation to create new instances artificially.

The mutate-operations we apply are as follows:

- *Paraphrase ($\tilde{\bullet}$)/ Summarization ($\ddot{\bullet}$):* We apply a language model to change the wording in the premise and/or conclusion. We use the state-of-the-art model Pegasus (Zhang et al., 2020) fine-tuned on paraphrasing or summarization.

from↓ to→	**v/n**	**v/¬n**	**¬v/n**	**¬v/¬n**
v/n	$p := \tilde{p}$ ...	$c := \ddot{p}$	$c := \neg c$	$p := p + \neg c$
v/¬n		$c := \tilde{c}$ ...		$c := \neg c$
¬v/n		$c := \ddot{p}$ $p := p + \tilde{c}$	... $p := \acute{p}$	$p := p + \neg c$
¬v/¬n		$c := \ddot{p}$ $p := p + \tilde{c}$	$p \mapsto c$	... $c := \acute{c}$

Table 1: Operations for synthetic data generation. Given an instance with a known label **v**alidity and **n**ovelty (rows) and a target **v**alidity/ **n**ovelty-label (columns), each cell lists the set of available operations (Section 3.2) to perform the desired mutation. The union of the operations in the cells in the diagonal apply to any single cell along the diagonal.

- *Substitution ($\acute{\bullet}$):* We introduce synonyms and hypernyms of words in the premise or conclusion using WordNet[1] (Fellbaum, 1998). We also add non-content phrases such as *Hence* and remove punctuation cues with a certain probability. The degree to which words are substituted is determined by random choice.
- *Negation ($\neg\bullet$):* We negate the conclusion or premise by adding/ removing the word "not" while preserving grammaticality.
- *Copy-Conclusion ($+$):* We append the (paraphrased) conclusion to the premise.
- *Move-Premise ($\mapsto$):* We move the last sentence of the premise into the conclusion.

In Table 1 we explain which of the above operations we apply, depending on the intended change of validity and novelty. In case more than one operation is applicable, we randomly select one operation. For example, if we synthesize a new instance with an unchanged label for validity and novelty, we randomly either paraphrase or substitute the premise or the conclusion.

Some cells are empty in Table 1, indicating a lack of mutation operations to accomplish the intended change in validity and novelty. In such cases we sample a new instance for augmentation. Potentially, all these mutations introduce noise to a different extent, e.g. paraphrases not being close to the source text, or substituted hypernyms affecting the validity of the argument, etc. As a kind of confidence measure, we individually scale the weight of the synthetic instances both in the dataset-specific weight mode and in the instance-individual weight mode.

4 Datasets

This section presents the four datasets we use in our work. As a baseline, we rely on the relatively small dataset provided by Heinisch et al. (2022b), in which conclusions were explicitly annotated for validity and novelty (henceforth task-internal data). We further rely on task-external data: the ExplaGraphs dataset by Saha et al. (2021), the IBM Debater datasets by Gretz et al. (2020b) and the annotated essays dataset by Stab and Gurevych (2017a) (sorted by their relatedness to the validity-novelty-classification task in descending order). Appendix A shows examples from each dataset.

4.1 Task-internal Data

We use the dataset of the shared task on predicting validity and novelty provided by Heinisch et al. (2022b) as task-internal data. This dataset is an extension of the dataset provided by Heinisch et al. (2022a) in the context of a conclusion generation approach. They used a fine-tuned language model to generate conclusions that follow a particular frame, conditioned on premises as input. The quality of the generated conclusions was rated regarding their validity and novelty by three annotators. The dataset is rather small in size, consisting of 750 manually annotated instances in the training split. The label distribution is quite imbalanced, with 55% of conclusions being valid, 16% being novel, and only 2% of conclusions being both novel and valid. Some instances (6%) have a tie in the aggregated annotations because one or all annotators indicated *"don't know"* for the aspect in question. We treat a tie in validity as a unknown label and a tie in novelty as $\frac{1}{2}$ since the conclusion seems to contain degrees of novelty[2]. By treating such potentially novel instances as being novel for our statistics in Table 2, we can double the proportion of non-valid & novel instances to 4%.

Since this dataset was manually labeled for the task of validity- and novelty-prediction, we give each instance the highest weight of $w_i = 3$ in the dataset-specific weight configuration. In the

[1] For efficiency reasons, we do not apply word sense disambiguation while selecting the synset in WordNet but give preference to the most probable first synset and prioritize replacing words having few synsets.

[2] Normally, our target values for validity and novelty are either unknown, 0, or 1, with an exception in this dataset for the novelty to model the special case of a tie.

Dataset	#	v/n	v/¬n	¬v/n	¬v/¬n
task-internal					
train	750	15%	38%	4%	39%
dev	198	19%	44%	22%	15%
test	520	25%	35%	18%	21%
task-external					
Expla	2.8k	55%	0%	45%	0%
IBM	30k	50%		50%	
Essays	749	66%		34%	

Table 2: Statistics of the processed source datasets, showing the total number (#) of retrieved instances and instance distributions for **v**alidity and **n**ovelty.

individual weighting setting for this dataset, the weight of each instance is scaled proportionally to the instance annotator agreements from $w_i = 1$ (no agreement at all) up to $w_i = 5$ (full agreement for validity and novelty).

Our development split and test split originate from the same data source as the task-internal training split annotated with validity and novelty, but cover different debate topics. For development and test data, we only consider instances that at a minimum achieved votes with majority agreement for both validity and novelty. More details on the dataset are given in Table 2.

4.2 Task-external Data

For our task-external dataset, we combine instances from the following three datasets.

ExplaGraphs by Saha et al. (2021) is a dataset for stance prediction. Given a textual belief and argument, the task is to classify the relationship between these short texts into support and attack. A belief in their setup can be seen as a conclusion, the argument as a premise. To make the link between belief and argument explicit, the authors perform a manual annotation that provides, for each sample, a conceptual explanation graph linking premise and conclusion. When reusing their data, we consider pairs linked by a support relation to have a valid conclusion and those related by an attack relation to have a non-valid conclusion. We consider all instances as novel, since the authors claim high novelty of the conclusions, which is supported by the inserted explanation graphs. Because of the high data quality due to the manual creation process we decide to double-weight each instance with $w_i = 2$ in the dataset-specific weighting. For individual weighting, we also consider the given explanation graph: if the graph is separable into non-contiguous subgraphs by deleting a single commonsense-concept node, indicating an inference which could be easy to undermine and is therefore not so representative, we subtract 0.8 from the dataset-specific weight. In case the resulting graph is linear, hinting a trivial straight-lined inference without combining different concepts or aspects, we further subtract 0.2 from the weight.

IBM Debater - IBM-ArgQ-Rank-30kArgs by Gretz et al. (2020b) (IBM) This dataset is used for determining the quality of arguments from 471 topics. Each argument consists of a topic and a premise pro or con the topic in question. For our purposes, the topic can be regarded as the conclusion. In their dataset, the support or attack of the premise towards its conclusion is manually labeled. We consider conclusions in support vs. attack as valid and invalid, respectively. Since the dataset does not contain any indicators for novelty, we set novelty to *'unknown'*. Since this dataset does not relate to the task of novelty prediction and only indirectly to validity prediction, we do not give a weight preference for instances from this dataset ($w_i = 1$), except in the individual weighting case where we allowed to consider the instance-individual annotated argument quality. We set the weight of low-quality-arguments (which are often defeasible) to $w_i = \frac{1}{2}$, and increase the weight with increasing quality up to $w_i = \frac{3}{2}$. After a manual inspection, we found support instances to be more reliable in general, such that we further add $\frac{1}{3}$ to the weight in these cases. Using the same weighting scheme, we further extend the dataset with 150 instances from arguments from non-American cultures provided by Kiesel et al. (2022) to increase the cultural diversity in this quality dataset.

Essays dataset by Stab and Gurevych (2017a,b) This dataset is based on student essays in which annotators marked spans of premises, claims, and major claims, as well as the argumentative relation between the different spans. Hence, the data is often used for argument unit recognition and classification. In further work by Stab and Gurevych (2017b), the arguments were annotated in terms of sufficiency, to indicate whether the premises provide enough evidence for accepting/rejecting the claim. For our purposes, we consider the binary sufficiency criterion as validity, while setting novelty to *'unknown'*. Again, this dataset does not relate to the task of novelty prediction and covers only one partial aspect of validity in one specific text genre (cropped text parts from student essays). To avoid models tailoring too much on this data, we lower the weight for each instance to $w_i = \frac{3}{4}$ in the dataset-specific setting. As for individual weighting, we set the weight to $\frac{1}{2}$ in case no annotator agreement information was given for an instance and to $\frac{5}{6}$ and 1, corresponding to a majority-agreement and full-agreement, respectively.

5 Experiments and Evaluation

In this section, we present our experimental results with the goal of testing the following hypotheses:

- The available task-internal training data is not sufficient to solve the task of predicting validity and novelty in a supervised manner (without additional external knowledge).
- Augmenting the data with task-external and synthetic data improves the quality of the predictions.
- Different amounts of (synthetic) data influence the performance. We expect that an optimal mixing proportion yields high F_1-scores, even without task-internal training data.

5.1 Experimental Setup

For our experiments we use the pretrained language model `roberta-large` (Zhuang et al., 2021) as available in the transformers library (Wolf et al., 2020), predicting both validity and novelty by having two feed-forwarded classification heads post-processed by the Sigmoid-function to map the prediction into the interval of $[0, 1]$ for validity and novelty, respectively.

Evaluation metric For evaluation, we rely on the ValNov-score which is the macro F_1-score over the F_1-scores for each class as shown in Equation 3.

$$\begin{aligned} ValNov = (&F_1(\text{valid\&novel})+\\ &F_1(\text{valid\&not-novel})+\\ &F_1(\text{not-valid\&novel})+\\ &F_1(\text{not-valid\&not-novel}))/4 \end{aligned} \tag{3}$$

We also measure the macro F_1-score for Validity (Val) and Novelty (Nov) separately.

Training We use the Adam optimizer with a maximum learning rate of 3e-5, a model weight decay of 3e-7, a batch size of 8 and early-stopping, checking the model performance on the development split each quarter of an epoch with patience of 4. We balance the source dataset and class distribution, allowing up to 20% instances having unknown validity or unknown novelty. We do not clone&mutate instances with unknown validity or novelty. Regarding the loss function in Equation 2, we set $\alpha = \beta = \gamma = 0.5$. We use binary target values $\{0, 1\}$ for validity and novelty.

Model selection The performance of our models varies substantially between runs due to randomized initialization. Some runs produce models that end up predicting only one class. To circumvent this problem, we run the training with different initialization for 12 runs, selecting the model with the best performance on the development set. More details can be found in the Appendix B.1.

5.2 Results and Evaluation

We run several experiments to evaluate our three hypotheses. First, we use only task-internal training data, then consider the integration of task-external and synthetic data, and finally, we vary training set sizes and data type proportions.

Baseline results (using only task-internal data) In this setting, our training set consists of 750 instances. This size is small compared to custom training sets for fine-tuning language models. Moreover, the number of instances per class is not balanced (Table 2). Hence, the results for the fine-tuned model are slightly worse compared to a random baseline of 24.5 ValNov-score. The best-performing model on the development split yields a ValNov-score of 23.2. Despite this low score, the F_1-score for classifying valid conclusions (61.5) outperforms the random baseline (49.5) and many other experimental settings. In contrast, the model completely fails to discriminate novelty: No novel instance was correctly predicted as novel. Introducing a class balance in the training data by undersampling removes this bias, and increases the F_1-score in novelty from 36.1 to 41.5 points which is still below the random baseline (49.8). The class-balanced training set contains only 137 instances, which results in a worse overall model performance of 21.4 ValNov-score. This first set of experiments highlights the need for techniques to overcome the problem of scarce labeled data and especially for solving the task of novelty prediction. We therefore aim to address the problem through augmentation of training data.

Augmenting training data with task-external and synthetic data Table 3 shows the results with different training set mixtures and instance weighting configurations, including the discussed baseline as reference.

Task-internal + synthetic training data: Augmenting task-internal data with synthetic instances by generating instances for underrepresented classes outperforms random guessing and our baseline model. We result in overall ValNov-scores of 33.3 / 38.1 / 38.3 without weight adjustments / weight adjustments only for synthetic instances / individual weight adjustments, respectively, outperforming the baseline by between 10.1 and 15.1 points. While there is a minor decrease on the prediction of validity, the prediction of novelty nearly doubles its F_1-score, yielding scores of up to 66.2 due to the additional novel instances in the synthetic data.

Task-external training data: Using task-external training data only without any task-internal data yields low ValNov-scores between 10 and 20.7, yielding worse results than the random baseline. This seems plausible as more than 93% of the datapoints lack a novelty label, with ExplaGraph being the only dataset including novelty information by exclusively presenting novel instances. It is only through the inclusion of synthetic data that we can increase performance to a ValNov-score of 22.6.

Task-internal + task-external training data: When combining task-internal and task-external training data, we generally observe minor improvements in the ValNov-score, having ValNov-scores of up to 25.1, which outperforms random guessing and our model baseline using internal training data only. One exception is the case of dataset-specific instance weighting, in which we regress to a model classifying all instances as valid and novel due to the (weighted) overpresence of valid and novel training instances. The settings in which synthetic training data is added worsen the ValNov-scores compared to the version of the system using internal and synthetic data only.

Effect of weighting Examining the impact of our weighting mechanisms, we see that the *dataset-specific weighting scheme* often worsens the results. For the task-internal condition in Table 3, we see no impact at all on ValNov-score. Considering the condition using internal and synthetic data, we do see an impact of dataset-specific weighting by +4.8 points in the ValNov-score by distinguishing between original and synthetic data in the impact of the learning rate. For the other conditions (external + synthetic data, internal + external + synthetic data) we see a detrimental impact of dataset-specific weighting. The *individual weighting scheme* has very mixed results in general. The internal+synthetic condition benefits from the individual weighting mechanisms as the ValNov-score increases by 0.2 points (from 38.1 to 38.3) and significantly increases the novelty score by 6.6 points (from 59.4 to 66.2), yielding the overall best result. For the other settings, the impact of individual weighting is very mixed, leading to similarly worse results compared to the dataset-specific weighting in the case of external data and internal+external+synthetic data. In the case of using external+synthetic data and internal+external data, however, individual weighting leads to higher ValNov-scores (+1.8 and +1.2 compared to disabled weight adjustments).

Effect of training data sizes for synthetic data Since our synthetic data generation method can generate an arbitrary number of instances, we explore the impact of different training data sizes on model performance. As sample sizes we consider a range from 100 instances to 100k instances (see Table 4). For all configurations, we see a clear increase in ValNov-score when moving from 100 to 1k training instances. We see improvements of between 4.1 points (internal+external data, individual weighting) to 19 points (internal data, individual weighting). Moving from 1k to 10k instances has a mixed impact. For some settings based on a large merged dataset of non-synthetic instances or individual weighting we see a further improvement (+1.6 for internal data with individual weighting, and +14.3 for in-&external with dataset-fixed weights). For other conditions we see a worsening of results moving from 1k to 10k instances. Interestingly, when moving from 10k to 100k instances, we see a worsening for nearly all conditions compared to the best results at 1k or 10k. Overall, the sweet-spot thus lies around 1k to 10k instances.

Data	w/o weight			dataset-specific weight			individual weight		
components	ValNov	Val	Nov	ValNov	Val	Nov	ValNov	Val	Nov
internal	23.2	**61.5**	36.1	23.2	**61.5**	36.1	22.7	57.7	36.1
+ synthetic	33.3	57.4	59.0	38.1	60.2	59.4	**38.3**	57.2	**66.2**
external	20.7	58.5	36.4	10.0	37.7	30.3	10.0	37.7	30.3
+ synthetic	21.8	50.5	42.6	15.8	41.8	36.0	22.6	41.9	57.1
internal+external	23.9	53.8	41.5	10.0	37.7	30.3	25.1	59.3	43.2
+ synthetic	32.7	57.9	51.0	13.1	37.7	36.1	13.1	37.7	36.1

Table 3: F_1-score-results for augmenting the training data with task-external and synthetic data. Synthetic data (based on the given data components) includes the class-balance, providing data for underrepresented classes. Using synthetic data does not change the number of training instances here, only the instance class distribution.

Config	100	1k	10k	100k
internal (ind. w.)	17.4	36.4	38.0	29.4
external (set w.)	18.9	34.7	32.9	23.9
external (ind. w.)	19.7	33.8	30.3	25.4
int-+external (w/o w.)	18.8	23.0	17.9	34.4
int-+external (set w.)	19.2	23.8	38.0	33.8
int-+external (ind. w.)	21.7	25.8	25.6	26.9

Table 4: ValNov-scores for training sizes (+synthetic data) without instance weighting (w/o w.), w/ dataset-specific (set w.) and w/ individual weighting (ind. w.)

Summary of results Using the task-internal data without augmenting it with synthetic or external data is insufficient to solve the validity-novelty-prediction task (ValNov-score of 23.2). Augmenting the task-internal data with synthetic data, including the class-balancing effect, improves the prediction performance. In fact, our best configuration is the one using the task-internal data class-balanced by the synthetic data, reaching the overall best ValNov-score of 38.3 and a high novelty F_1-score of 66.2, in addition to a above-average validity prediction score of 57.2 that is only seven points away from the overall maximum (64.5 with 10,000 dataset-specific weighted internal-external-synthetic instances).

Adding additional external or more synthetic data does not improve performance in general. In fact, we see the different data proportions heavily influence the performance, especially the right amount of synthetic data seems to be crucial. While we see some improvements in having 1k and 10k instances, the performance is often negatively affected when adding further synthetic training data instances.

A quite remarkable result, however, is that in spite of not seeing improvements in the ValNov-score when using external data *in addition* to task-internal data, we observe that by using task-external data *instead of* task-internal data, we can get comparable results to training with task-internal data. Using 1,000 task-external and synthetic instances with dataset-specific weighting, we obtain a model with only 3.6 points less in the ValNov-score and an F_1-score of 65.2 in the novelty aspect, which is only 2.4 points below the overall maximum (10,000 individual weighted internal-synthetic instances).

5.3 Case Study

In a case study, we compare the predictions made by the *task-internal model* (trained with task-internal training data without any changes), the *task-internal-synthetic model* (750 individual-weighted task-internal instances class-balanced with synthetic instances), the *task-internal-external-synthetic model* (10,000 dataset-specific weighted task-internal and task-external instances class-balanced with synthetic instances) and *task-external-synthetic model* (1,000 dataset-specific weighted task-external instances class-weighted with synthetic instances). We consider different conclusion candidates for the premise:

> *"**Year-round school**: Many districts are finding that year-round schools are not cost-effective to operate unless the student population substantially exceeds traditional school capacity"*.

The conclusion *"Many districts find year-round schools are not cost-effective"* is a valid but not novel summary of the premise – which is easy to detect by paraphrase-recognition capabilities. All our four models succeed in predicting the validity and lack of novelty of this conclusion.

In order to further understand the behavior of our models, we consider a conclusion that incorporates an inconsistency with respect to the above premise. However, the inconsistency is subtle and not trivial to detect. Consider the conclusion: *"Year-round schools are ineffective when student populations exceed capacity"* which contradicts the statement that "year-round schools would be cost-effective if student population would exceed capacity", which follows from the above premise. The conclusion thus represents a non-valid-non-novel example. All models with the exception of the task-internal-external-synthetic model fail to recognize the contradiction and classify the example as valid. We hypothesize that the task-internal-external-synthetic model captures this example because it has been largely trained with antonym-substitution (cost-effective vs. ineffective in the above example). However, the model slightly misclassifies the novelty with a probability of 56% being novel due to a tendency to classify non-valid instances as novel. We consider a more obviously inconsistent conclusion with an explicit negation: *"Year-round schools are not cost-effective for large schools"*. All models misclassify this example as valid, showing a general lack of logical reasoning capabilities. In particular, there is an obvious element of commonsense-knowledge (large school = school with high student capacity) that the models are lacking.

Finally, we consider a clearly off-topic conclusion: *"Offshore drilling is very valuable to the US economy"*, which is neither valid nor novel. All models successfully predict the non-validity of the conclusion, including the task-internal model that otherwise consistently votes for validity in our case study. Regarding the novelty aspect, only the task-external-synthetic model misclassified the example as novel because it never saw such completely unrelated conclusions in its training data.

We further analyze the models in Appendix B.3.

6 Conclusion

Predicting the validity and novelty of a given conclusion based on its premise is a challenging task. Using 750 class-unbalanced training instances annotated with validity and novelty does not provide enough evidence for tuning a large language model. Augmenting the task-internal training data to 10,000 instances using task-external and synthetic data increases the ValNov-score up to 38.0. Using task-internal and synthetic data to balance the training data increases this score to 38.3. However, the results achieved by data augmentation techniques are still very modest, showing that massive training data and modern language models alone are not sufficient for solving the task. While valid but non-novel instances can, to a large part, be detected using paraphrase recognition tests, many instances require logical inference and commonsense knowledge to properly classify validity and novelty. None of these capabilities are supported in the subsymbolic approach we chose in this work. In future work, we aim to investigate the impact of incorporating commonsense knowledge and deeper logical reasoning into the task of validity and novelty prediction.

References

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020a. The workweek is the best time to start a family – a study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 528–544, Online. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020b. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Philipp Heinisch, Anette Frank, Juri Opitz, and Philipp Cimiano. 2022a. Strategies for framing argumentative conclusion generation. In *Findings of the Association for Computational Linguistics: ACL-INLG 2022*. Association for Computational Linguistics.

Philipp Heinisch, Anette Frank, Juri Opitz, Moritz Plenz, and Philipp Cimiano. 2022b. Overview of the validity and novelty prediction shared task. In *Proceedings of the 9th*

Workshop on Argument Mining, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ning Jin, Jiaxian Wu, Xiang Ma, Ke Yan, and Yuchang Mo. 2020. Multi-task learning model based on multi-scale cnn and lstm for sentiment classification. *IEEE Access*, 8:77060–77072.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*. Association for Computational Linguistics.

Juri Opitz, Philipp Heinisch, Philipp Wiesenbach, Philipp Cimiano, and Anette Frank. 2021. Explainable unsupervised argument similarity rating with Abstract Meaning Representation and conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *EMNLP*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017a. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating informative conclusions for argumentative texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3482–3493, Online. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 2020 Thirty-seventh International Conference on Machine Learning*, pages 1–55, Vienna, Austria,. ICML.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Examples from the Task-internal and Task-external Datasets

In this section we give examples for each dataset presented in Section 4, showing how the examples have been mapped to our premise-conclusion schema with novelty and validity indicators.

A.1 Task-internal Dataset

Example 1:

(1) Premise: *Twin Towers reconstruction: Pentagon, hardly a symbol of peace, has been rebuilt. Twin towers weren't. The message that this sends to the public is hardly positive.*
Conclusion: *Pentagon rebuild sends wrong message of peace*

Validity: yes / Novelty: yes
Weight: dataset-specific: 3 / individual-weighted: 1.5

Example 2:

(2) Premise: *Twin Towers reconstruction: Pentagon, hardly a symbol of peace, has been rebuilt. Twin towers weren't. The message that this sends to the public is hardly positive.*
Conclusion: *Twin towers are hardly a symbol of peace*
Validity: no / Novelty: no
Weight: dataset-specific: 3 / individual-weighted: 3.25

A.2 Task-external Datasets

A.2.1 ExplaGraphs

Example 1:

(3) Premise: *It is not realistic to abandon television, as many people still get current new information from it.*
Conclusion: *Television viewing should be moderated, not banned.*
Validity: yes / Novelty: yes
Weight: dataset-specific: 2 / individual-weighted: 2

Example 2:

(4) Premise: *Intelligence tests lower self esteem.*
Conclusion: *Intelligence tests are harmless.*
Validity: no / Novelty: yes
Weight: dataset-specific: 2 / individual-weighted: 1.8

A.2.2 IBM-ArgQ_Rank-30kArguments

(5) Premise: *A country with a diverse population is better represented by a multi-party system.*
Conclusion: *We should adopt a multi-party system*
Validity: yes / Novelty: unknown
Weight: dataset-specific: 1 / individual-weighted: 1.27

Example 2:

(6) Premise: *telemarketers have to earn a living wage somehow. it is better than government assistance*
Conclusion: *We should ban telemarketing*
Validity: no / Novelty: unknown
Weight: dataset-specific: 1 / individual-weighted: 0.53

A.2.3 Essay dataset

(7) Premise: *All the living creatures live together on our mother Earth and she is the only one.*
Conclusion: *First , environmental protection is far more urgent than economic developments.*
Validity: yes / Novelty: unknown
Weight: dataset-specific: 0.75 / individual-weighted: 0.5

Example 2:

(8) Premise: *Arts include many forms and music as well as cinema are the most typical . These two art forms not only provide the public with entertainment but also contribute significantly to the economy .*
Conclusion: *But our standard of living also depend on another factor - spiritual life which is related closely with arts .*
Validity: no / Novelty: unknown
Weight: dataset-specific: 0.75 / individual-weighted: 0.5

B Further Details of the Experimental Setup and Results

We give further details about the model selection process for each experiment (B.1) and give further insights into the model performance (B.2) and test prediction (B.3). For additional information about the implementation consult our code located at `https://github.com/phhei/ValidityNoveltyRegressor`.

B.1 Model Selection

In our experiments, we observed a high variance of results across runs. The deviations are mainly caused by the random factors introduced in the synthetic data generation and partially caused by the random initialization of weights for the classification heads. We observed in particular that often fine-tuned models get stuck in local optima in some runs, often over-focusing on one specific class (e.g., valid&not-novel) and failing completely in all other three classes. We thus ran each configuration 12 times per default, reducing the number of runs further for increasing training data sizes, that is, six runs for 10,000 - 50,000 instances, and three runs in the case of 100,000 instances. We select the model achieving the highest ValNov-score on the development split among all runs.

	100		1k		10k		100k	
Config	Val	Nov	Val	Nov	Val	Nov	Val	Nov
internal (ind. w.)	45.3	38.5	61.4	59.6	54.7	67.6	58.0	57.9
external (set w.)	47.6	40.8	57.0	65.2	57.3	53.9	44.8	51.6
external (ind. w.)	43.4	47.7	58.0	63.4	49.3	58.1	51.2	49.0
int-+external (w/o w.)	50.0	39.3	49.0	46.2	48.5	40.3	53.0	60.0
int-+external (set w.)	46.2	49.5	40.4	39.6	64.5	57.2	57.7	55.0
int-+external (ind. w.)	54.2	40.0	60.8	39.1	43.0	61.8	52.8	49.5

Table 5: F_1-scores for **val**idity and **nov**elty for different training sizes (+synthetic data) without instance weighting (w/o w.), w/ dataset-specific (set w.) and w/ individual weighting (ind. w.). For the ValNov-scores see Table 4.

We observed that selecting the final model based on the performance on the development split is a good indicator, especially for models trained on large training sets. In 58% of all cases, the best performing model on the development split was also the best performing model on the test split. In all other cases, the selected model achieves $\varnothing 88.8\%$ of the ValNov-score that would have been achieved based on model selection on test data.

B.2 Further Details regarding Effect of Training Data Sizes for Synthetic Data

Table 5 shows the F_1-scores in addition to the ValNov-scores given in Table 4. Table 4 and 5 omit some source-data-weight-combinations, e.g. task-internal data in combination with the uniform weighting setting. We omit these combinations because they do not outperform the other weight settings given the same training data in any data set size. Table 3 hints at this trend already, with instance-individual weighting as the outperforming weighting setting when using only task-internal data in combination with synthetic data.

B.3 Further Analyses of the Test-predictions

We carried out a further analysis of the predictions on the task-internal test set of the baseline model (Section 5.2), the *task-internal model* (trained with the task-internal training data without any changes), the *task-internal-synthetic model* (750 individual-weighted task-internal instances class-balanced with synthetic instances), the *task-internal-external-synthetic model* (10,000 dataset-specific weighted task-internal and task-external instances class-balanced with synthetic instances), and *task-external-synthetic model* (1,000 dataset-specific weighted task-external instances class-weighted with synthetic instances). Figures 2-5 show the heatmaps and histograms for validity and novelty of the predictions and prediction errors of these four models.

The baseline model (Figure 2) succeeds in distinguishing between valid and non-valid conclusions in some cases. However, it fails completely in the case of novelty, as every instance is classified as non-novel (the predicted probability of a conclusion being novel is between 1% and 5%). This leads to very low scores regarding novelty prediction, yielding an F_1-score of 36.1. The baseline model is thus biased to detect valid but non-novel conclusions, for example repetitions of the premise.

The model trained on data augmented with synthetic instances (Figure 3) is more diverse in its predictions, mostly predicting examples as being valid, both novel and not novel. The model learns successfully to discriminate between novel and non-novel conclusions with an an F_1-score of 66.2, thus being a good summarization detector. However, the model avoids to classify a conclusion as not valid but novel, with an F_1-score of only 15.1 in this case. By avoiding such difficult cases, the model correctly predicts at least one of the two quality dimensions (novelty, validity) in many cases.

The task-internal-external-synthetic model (Figure 4) succeeds very well in recognizing conclusions that are valid but not novel (66.2 F_1-score). The corresponding training data includes a high number of examples which vary in terms of their validity label. The performance of the model on novelty prediction, however, remains weak.

When discarding the task-internal data and thus applying a model trained on task-external data to task-internal test data (Figure 5), this leads to high diversity and thus uncertainty in the predicted labels. In spite of this, it is quite remarkable that the model predicts at least one of the two quality dimensions correctly in many cases. However, the model has lower F_1-scores in recognizing valid-non-novel conclusions (59.9) and especially non-valid-non-novel conclusions (11.0). We hypothesize that this is due to the fact that the model has only seen synthetic instances in the latter class. Hence, the model rarely saw random off-topic conclusions which are not valid and not novel and part of the task-internal test data. The performance of the model on recognizing non-valid and novel conclusions (28.7 F_1-score) is however above the baseline. This is likely due to the many non-valid but novel instances in the ExplaGraphs dataset.

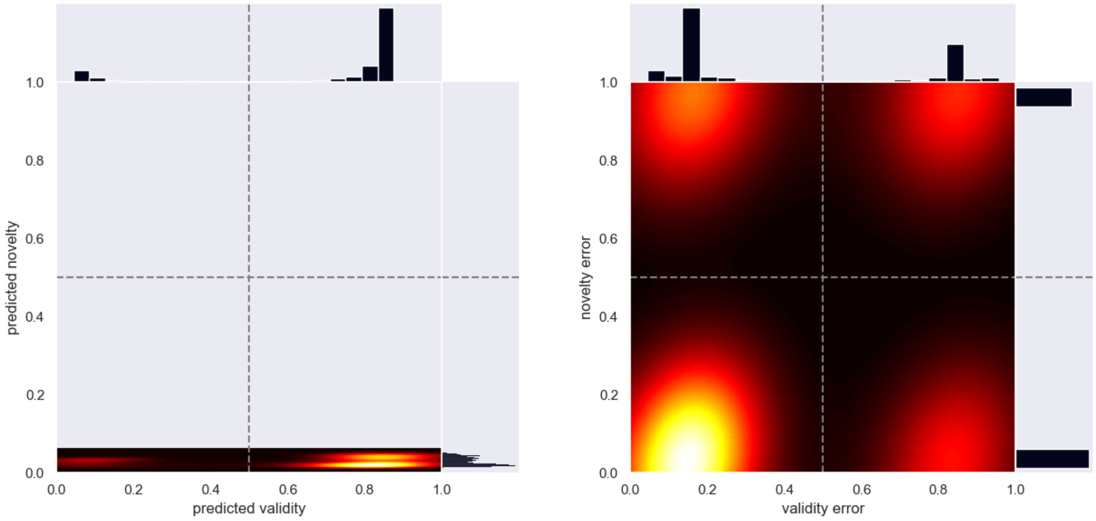

Figure 2: Heatmaps for the baseline-model *(task-internal model)*. The highest predicted value for novelty is 0.05. Therefore, the plots contain gray areas.

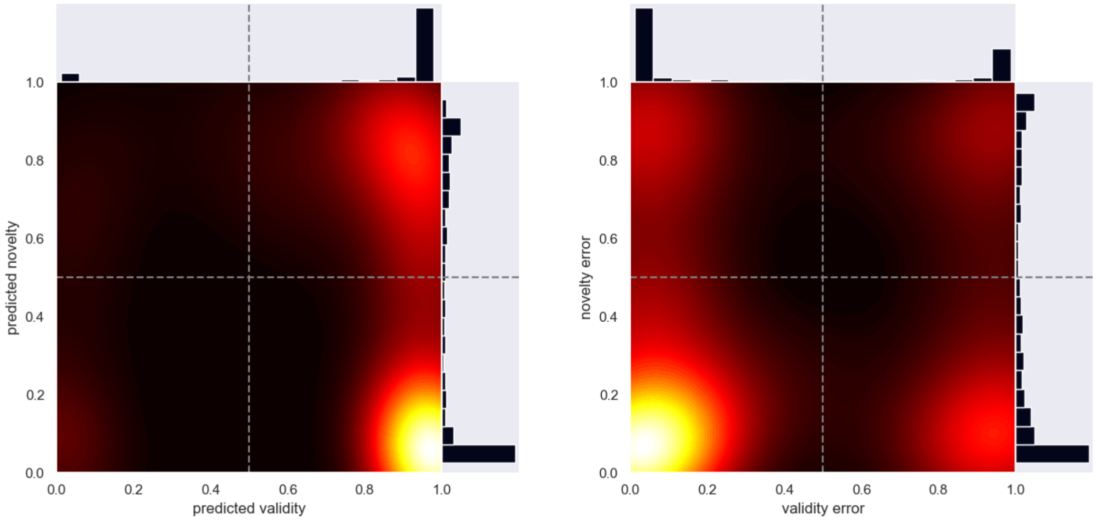

Figure 3: Heatmaps for the *task-internal-synthetic model.*

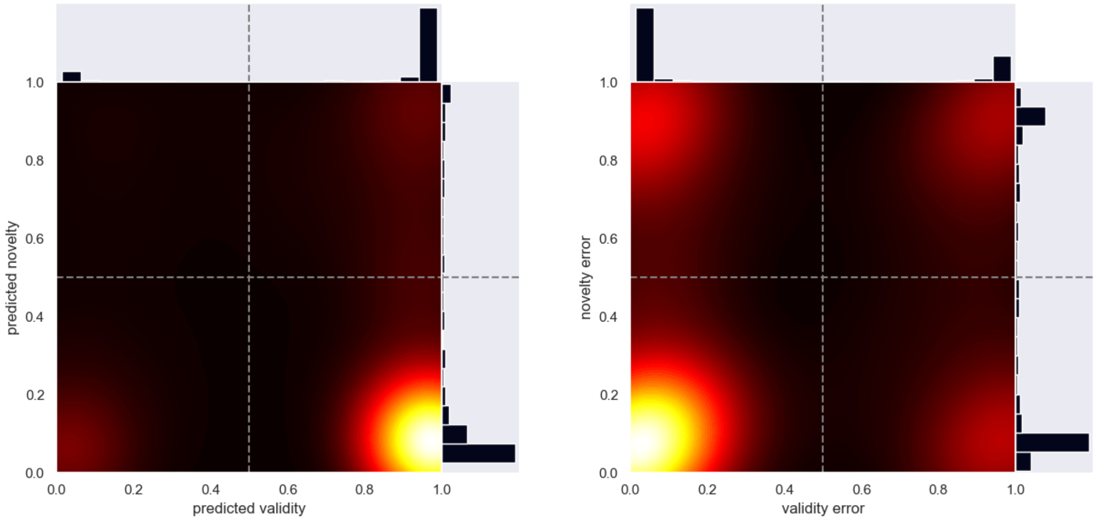

Figure 4: Heatmaps for the *task-internal-external-synthetic model.*

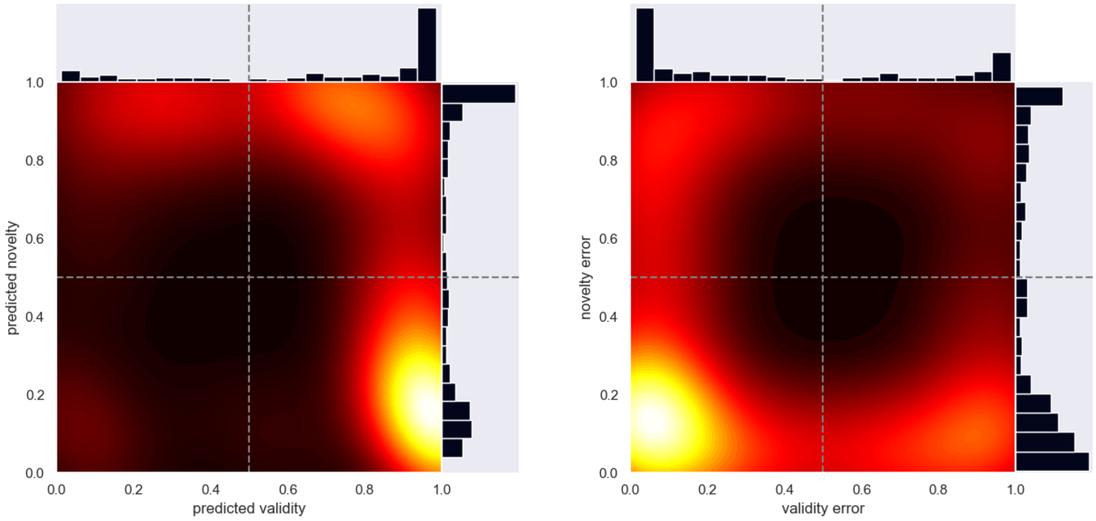

Figure 5: Heatmaps for the *task-external-synthetic model.*