Overview of the 2022 Validity and Novelty Prediction Shared Task

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

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

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

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

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

Abstract

This paper provides an overview of the Argument Validity and Novelty Prediction Shared Task that was organized as part of the 9th Workshop on Argument Mining (ArgMining 2022). The task focused on the prediction of the validity and novelty of a conclusion given a textual premise. Validity is defined as the degree to which the conclusion is justified with respect to the given premise. Novelty defines the degree to which the conclusion contains content that is new in relation to the premise. Six groups participated in the task, submitting overall 13 system runs for the subtask of binary classification and 2 system runs for the subtask of relative classification. The results reveal that the task is challenging, with best results obtained for Validity prediction in the range of 75% F_1 score, for Novelty prediction of 70% F_1 score and for correctly predicting both Validity and Novelty of 45% F_1 score. In this paper we summarize the task definition and dataset. We give an overview of the results obtained by the participating systems, as well as insights to be gained from the diverse contributions[1].

1 Introduction

An important challenge within the field of argument mining is the assessment of the quality of an argument. In recent years, several systems have emerged that make mined arguments accessible to an end user, either via search engines (Wachsmuth et al., 2017b), debate summarization systems (Bar-Haim et al., 2020), dialogue systems (Rach et al., 2021), or by other means. In order to establish confidence and trust on the side of the user, the ability to distinguish high-quality arguments from low-quality ones is important.

Wachsmuth et al. (2017a) investigated the notion of quality for argumentation and proposed 3 dimensions along which the quality of arguments can be rated: cogency, reasonableness, and effectiveness, introducing corresponding subcategories for each dimension. However, there have not been many attempts to operationalize the notion of quality so far, e.g., by an exact definition of a metric that assesses the quality or by means of an automatic procedure to determine the quality. Exceptions are datasets manually labeled with coarse scores denoting overall quality, which have been used for supervised learning (Toledo et al., 2019; Gretz et al., 2020b) or attempts to determine single subdimensions, such as sufficiency (a subdimension of cogency) (Stab and Gurevych, 2017b; Gurcke et al., 2021).

Motivated by this gap, the authors of this paper decided to propose a new shared task and submitted it to the 9th Argmining Workshop. Instead of tackling the entire wide field of argument quality or isolating a single quality aspect, we focus on the conclusion in the context of its argument, and assess its quality in the Validity and Novelty Prediction Shared Task. This task consists of the prediction of these two important conclusion quality dimensions.

Following Opitz et al. (2021), we define Validity as the degree to which the conclusion is justified with respect to the given premise, and Novelty as the degree to which the conclusion contains premise-related content that is not explicitly stated in the premise.

The two notions stand in a trade-off to each other as it is straightforward to maximize one of them at the expense of the other. Copying or paraphrasing parts of the premise as a conclusion will yield high validity but no novelty. Expanding a concept of the premise with commonsense knowledge as a conclusion can potentially yield high novelty but may not satisfy validity. Previous research in Opitz et al. (2021) and Heinisch et al. (2022) has indeed shown that it is difficult to generate conclusions that satisfy both criteria, which require proper inference based on and expanding the premise.

[1]The shared task website including the data and result table is located at https://phhei.github.io/ArgsValidNovel/

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 84–94
October 17, 2022.

We divide the task of predicting validity and novelty into two subtasks. The first subtask consists in the binary prediction of whether a conclusion is valid resp. novel or not. The second subtask is framed as a comparative task, tasking systems to predict which of two given conclusions is more valid resp. more novel compared to the other, or whether they form a tie. The best achieved F_1 score for binary prediction of both validity and novelty is 45.16, by van der Meer et al. (2022), and the best achieved F_1 score averaging the scores for relative validity and novelty is 41.5, by the team NLP@UIT[2] – while the best scores for predicting Validity and Novelty as single prediction targets yield substantially higher results, with up to 74.6 points F_1 score for Validity, and 70 points F_1 score for Novelty. This large contrast shows that the joint objective is challenging. Judging from the properties of the high-scoring systems for the individual quality aspects, we conclude that this challenging task requires strong text understanding capabilities, as well as (symbolic) background knowledge, which our received submissions are addressing, by taking a first step towards tackling this fundamental task for many downstream applications in Computational Argumentation.

In the following, we describe the task as organized in the context of the 9th Argument Mining Workshop. We describe the datasets used, as well as the different systems that have participated in the task. We provide an overview of the results the systems have obtained and make explicit the lessons we can learn from the shared task results, so that these observations can guide the community in their future choice of methods to address this and related tasks.

2 Related work

Argument quality Within the growing field of Computational Argumentation, an important concern is to assess the quality of arguments. In their seminal work, Wachsmuth et al. (2017a) established important dimensions for rating the quality of arguments. They proposed three quality dimensions: *cogency* (related to logics), *effectiveness* (relating to rhetoric) and *reasonableness* (relating to dialetics) – which they sub-divided into 11 fine-grained quality aspects. In a recent survey, Vecchi et al. (2021) extend the notion of argument quality to account for their function in deliberative processes, in the sense that good arguments should "ensure the discourse to unfold productively", e.g., by bringing *new aspects* into the discussion.

Several works have proposed computational models to rate the quality of arguments. While Toledo et al. (2019) and Gretz et al. (2020b) target rather coarse overall quality scores based on single quality labels, other systems were designed to assess specific quality aspects, such as *convincingness* (Habernal and Gurevych, 2016), *relevance* (Wachsmuth et al., 2017c) or *cogency*, the logical coherence of an argument (Lauscher et al., 2020).

While these works assess the quality of an argument as a whole, Stab and Gurevych (2017b) focused on the quality of the premises of an argument, in terms of their *sufficiency*, asking whether an argument's premises provide enough evidence for accepting or rejecting its claim or conclusion. They provide annotations of argument sufficiency on the argument essays (Stab and Gurevych, 2017a) and develop a classifier that achieves 84% accuracy for detecting (in)sufficiently supported arguments as a whole, including their (in)sufficient premises. Gurcke et al. (2021) revisited this quality criterion (sufficiency) in a new task formulation: *conclusion generation* from (in)sufficient premises, where the aim is to determine the sufficiency of a premise by examining the quality of the generated conclusion including the premise.

Argument conclusion generation Follow-up research investigated argument conclusion generation from different angles, focusing on the generation of **conclusions with specific properties**, such as *plausibility* (next to stance) (Gretz et al., 2020a), *informativeness* (beyond validity) (Syed et al., 2021), or realizing a specific *frame* (Heinisch et al., 2022).

Measuring *novelty and validity* of conclusions Opitz et al. (2021) found that assessing the novelty and validity of conclusions in the context of a premise poses a challenging problem for automatic metrics. Their work aimed at assessing the similarity of arguments by taking their conclusions into account – which they generated with a fine-tuned T5 pre-trained language model. However, while the automatically generated conclusions were able to increase the similarity rating performance, the gain was rather small. In a manual evaluation study they found a key problem in the *quality* of the generated conclusions: they were often either *novel*, or *valid*, but rarely both, thus either adding little in-

[2]No description paper was submitted for this system.

formation (no novelty), or introducing misleading information (no validity). The fact that novelty and validity are complementary, and, to some degree, dueling aspects is further corroborated by Heinisch et al. (2022) who show that it is challenging to automatically generate conclusions that are *both* valid and novel.

We therefore believe that the development of methods that can assess these key quality aspects of conclusions poses a challenging and interesting task for the community. In particular, the results of the task may not only provide strong baselines and future improvement perspectives of such metrics, but also provide useful guidance about the improvement of conclusion generation methods.

3 Task Details

3.1 Task Description

Given a textual premise and conclusion candidate, the VALNOV task consists in predicting two aspects of a conclusion: its *validity* and *novelty*.

Validity is defined as the degree to which the conclusion is *justified* with respect to the given premise. A conclusion is considered to be valid if it is supported by inferences that link the premise to the conclusion, based on logical principles or commonsense or world knowledge, which may be defeasible. A conclusion will be trivially considered valid if it repeats or summarizes the premise – in which case it can hardly be considered as *novel*.

Novelty defines the degree to which the conclusion contains content that is *new in relation to the premise*. As extreme cases, a conclusion candidate that repeats or summarizes the premise or is unrelated to the premise will not be considered novel.

We structured the shared task into two subtasks. **Subtask A** considers a *coarse-grained categorization* of validity and novelty by predicting binary labels denoting whether a conclusion candidate is *valid* or *not valid* and *novel* or *not novel*. In **Subtask B** we aim at a more fine-grained analysis without losing the advantages of using discrete labels for evaluation. Here, we give two conclusion candidates instead of one and task the systems to predict whether one, and if so, which conclusion is *more* valid and novel than the other, respectively, resulting in a ternary prediction task with categories: {*Conclusion 1 is better, Tie, Conclusion 2 is better*}, for each quality aspect.

Split	#	v/n	v/¬n	¬v/n	¬v/¬n
train	750	14%	39%	2%	39%
dev	202	19%	43%	22%	14%
test	520	25%	35%	18%	21%

Table 1: Data statistics for subtask A, considering **v**alidity and **n**ovelty.

		Validity		
		C1	**tie**	**C2**
Novelty	**C1**	8%	4%	6%
	tie	12%	32%	10%
	C2	9%	7%	12%

Table 2: Test data statistics for subtask B, considering validity and novelty.

3.2 Data

The data used in the Validity and Novelty shared task originates from a manual annotation study by Heinisch et al. (2022). They used as a basis the argumentative dataset of Ajjour et al. (2019), which had been collected from the high-quality, mostly political arguments from `debatepedia.org`. Heinisch et al. (2022) used the topic and premises from this data and generated automatic conclusions from them, which they then presented to human annotators to judge their validity and novelty, as well as the original conclusions, or conclusions randomly sampled from the remaining instances. The annotators had a higher education entrance qualification and some experience in the field of argument mining. Each data instance was labeled by three annotators for validity and novelty, where they could choose from the options {yes, I don't know, no} and {Conclusion 1 is better, tie, Conclusion 2 is better} for Subtask A and Subtask B, respectively. The annotators labeled validity and novelty separately and independently from each other. In order to reduce the annotation workload and to offer a more fine-grained analysis for validity and novelty prediction, we presented five to ten different conclusions (Subtask A) and conclusion combinations (Subtask B) for each premise, sometimes having only minor surface differences in the presented conclusions.

Since the annotation of validity and especially novelty introduces a degree of subjectiveness, as in many annotation tasks in the field of argument mining (Gurcke et al., 2021), we published the agreements for each instance. For Subtask A, we distinguish four classes of agreement: "defeasible"

(there is no agreement due to one or three "I don't know"-labels), "majority" (two out of three annotators agree), "confident" (two out of three annotators agree and the third annotator labels "I don't know"), and "very confident" (full agreement). Defeasible instances are uncommon (1-4%) and were discarded for the test split. Two out of three samples have very confident validity labels, and every second sample yields a very confident novelty label. An exception is the test-split, with 41% very confident novelty labels and 58% majority novelty labels. We found similar agreements in Subtask B, except for a slightly increased chance (5%) to have one vote for Conclusion 1, one Vote for a Tie, and one vote for Conclusion 2 for validity and novelty, respectively. In such cases, we set the final label to "tie" in validity and novelty, respectively, instead of "unknown". For all other annotator decision distributions, we consider the label with the highest number of votes.

We split the data into train, development, and test data by avoiding a topic-overlap between train (22 overlapping topics for Subtasks A and B, respectively) and development (eight and seven overlapping topics for Subtasks A and B, respectively) data. For Subtask A and B, the development- and test data share eight topics, including the premises but different conclusions. In addition, the test split introduces seven novel topics. The train split and the test split have no topics in common. Overall, we have annotations for 750 train samples, 202 development samples, and 520 test samples for Subtask A and 600 train samples, 72 development samples, and 283 test samples for Subtask B. Further data statistics are in Table 1 and Table 2 for Subtask A and Subtask B, respectively.

We published the train- and development data split for developing the systems and released the test split without reference labels for the final prediction submissions. We revealed the test labels afterward.

3.3 Metrics

For evaluation, we consider standard metrics relying on the F_1 score measured on the predictions made on the predefined test split. For subtask A, our main metric for ranking the submissions is the macro F_1-score for predicting both validity and novelty, resulting in four different combinations (valid and novel, valid and not novel, not valid and novel, not valid and not novel). We also report the macro F_1 scores for validity and novelty separately. For subtask B, we respect the more fine-grained character and rely on the average of the separately calculated macro F_1 scores for validity and novelty.

4 Submissions and Results

In total, we received 13 submissions, from six participating teams[3] for Subtask A, and an additional submission each for Subtask B from two teams that participated in Subtask A. In addition we provide baselines for both subtasks, by fine-tuning the RoBERTa-base-language model (Zhuang et al., 2021) on the Shared Task training data, once to predict validity and and once novelty independently of each other (more details in Appendix A).

Note that some teams did not provide a system description paper. We nevertheless include their results and short descriptions based on the teams' submission information.

4.1 Subtask A

All the submitted systems rely on machine learning in some way.

Many of the submitted systems have built on large language models, mostly RoBERTa (Zhuang et al., 2021)), based on the transformer architecture. Some submitted systems fine-tuned large language models trained on the Natural Language Inference (NLI) and/or Argument Relation classification (ArgRel) task. In order to couple the predictions for both tasks (validity and novelty), it seems intuitive to propose a joint architecture based on Multi-Task-Learning which one of the submitted systems opts for. A further option is to rely on auto-regressive language model such as GPT-3, conditioning them on selected prompts to predict the quality labels as a generative task.

Beyond applying state-of-the-art machine learning architectures and models on the task, some participants have looked into the question how to incorporate background knowledge into the task. Two participating teams have looked in particular into how to extract paths from background knowledge resources such as ConceptNet (Speer et al., 2017) or WikiData (Vrandečić and Krötzsch, 2014) and incorporate these paths as features into a classifier.

We describe the participating systems in more detail in the following.

[3]We allowed each team to submit up to five different system runs.

Team submissions	Short Description	ValNov	Validity	Novelty
CLTeamL-3	GPT-3$_{\boldsymbol{Val}\&Nov}$+$_{NLI}$RoBERTa$_{Val\&\boldsymbol{Nov}}$	**45.16**	**74.64**	61.75
AXiS@EdUni-1	FFNN$_{Val\&Nov}$ w/ $_{NLI}$BART & WikiData	43.27	69.80	62.43
ACCEPT-1	SVM$_{Val\&Nov}$ w/ ConceptNet & SBERT	43.13	59.20	**70.00**
CLTeamL-5	GPT-3$_{\boldsymbol{Val}\&Nov}$+$_{ARC}$RoBERTa$_{Val\&\boldsymbol{Nov}}$	43.10	**74.64**	58.90
CSS*	$_{NLI}$RoBERTa$_{Val\&Nov}$	42.40	70.76	59.86
AXiS@EdUni-2	FFNN$_{Val\|Nov}$ w/ $_{NLI}$BART & WikiData	39.74	66.69	61.63
CLTeamL-2	$_{NLI}$RoBERTa$_{Val\&Nov}$	38.70	65.03	61.75
CLTeamL-1	GPT-3$_{Val\&Nov}$	35.32	**74.64**	46.07
CLTeamL-4	$_{ARC}$RoBERTa$_{Val\&Nov}$	33.11	56.74	58.95
ACCEPT-3	SVM$_{Val\&Nov}$ w/ ConceptNet	30.13	58.63	56.81
ACCEPT-2	SVM$_{Val\|Nov}$ w/ ConceptNet & SBERT	29.92	56.80	48.10
NLP@UIT	SBERT	25.89	61.72	43.36
Baseline	RoBERTa$_{Val\|Nov}$	23.90	59.96	36.12
Harshad	BERT$_{Val}$ + novelty := validity	17.35	56.31	39.00
-	overall system-average excluding the baseline	35.94	62.74	52.97

Table 3: Results (macro-F1-scores) for subtask A including short descriptions for each system. A "&"-sign indicates a jointly trained Validity-Novelty-Predictor, a "|"-sign validity and novelty predictions independent of each other.
*The CSS team revised their predictions after the submission deadline due to detecting a formatting failure of their previously submitted prediction file

Team CLTeamL described in van der Meer et al. (2022), submitted five system runs. They experimented with prompting GPT-3 in a few-shot scenario for both prediction targets (validity and novelty). They combine prompting with in-context learning, providing four samples from the training data that obtained majority annotator agreement, and a test sample to be classified. They also experimented with fine-tuning a multi-task RoBERTa-model on i) the NLI task, or ii) argument relation classification (ArgRel). The fine-tuned models are optionally further refined by contrastive learning.

Submission *CLTeamL-1* uses the validity and novelty predictions obtained from GPT-3 prompts. While GPT-3 performs well in predicting validity (F_1-score of 74.64), it fails in predicting novelty (F_1-score of 46.07) and, therefore, achieves a modest ValNov score of 35.32. Submission *CLTeamL-2* only uses the fine-tuned NLI RoBERTa model. This yields reverse results, with a lower score for validity (65.03) but a better score for novelty prediction (61.75). Submission *CLTeamL-3* combines GPT-3 prompting for validity and the NLI-based fine-tuned RoBERTa, further enhanced with contrastive learning for novelty. With this, the system achieves the *overall best shared task results* for ValNov (45.16), as well as the best results for validity (74.64) and the 3rd best score for novelty (61.75). Relying on a RoBERTa model fine-tuned on the ArgRel instead of the NLI task makes the overall results wose (submission *CLTeamL-4* without GPT-3, submission *CLTeamL-4* with GPT-3 for validity).

Team AXiS@EdUni-1 submitted two system runs (Saadat-Yazdi et al., 2022). The system combines diverse components in a joint prediction system: i) NLI knowledge via a fine-tuned BART NLI system, which computes NLI prediction scores in two directions from premise to conclusion and vice versa; ii) neural models predicting a) semantic distance of premise and conclusion via SBERT, and b) validity and novelty by fine-tuning BERT on the training set; finally iii) knowledge from the WikiData knowledge graph, by extracting knowledge paths between premise and conclusion concepts to determine a) the semantic distance of premise and conclusion (average path length), and b) an irrelevancy score from unconnected conclusion concepts.

The features obtained from each component are fed to a small FFNN to jointly predict validity and novelty (*AXiS-1*). Submission *AXiS-2* combines the predictions of two separately trained FFNNs for validity and novelty. *AXiS-1*, with an overall F_1-score of 43.27, clearly outperforms the *AXiS-2* system with separate validity and novelty predictions (39.47). With this, *AXiS-1* ranks 2nd in the overall task, 2nd for novelty and 3rd for validity. Notably, AXiS-1 obtains the first place when con-

sidering all systems that do not leverage GPT-3.

System ablations show that i) NLI from premise to conclusion has stronger impact on results, while the reversed direction also contributes. Semantic distance has a stronger impact on validity, while irrelevancy mostly contributes for the joint ValNov score. Comparing the impact of features from neural vs. knowledge graph resources indicates that neural features have stronger impact, while both feature types contribute to the overall system score.

Team ACCEPT submitted three system runs.[4] *ACCEPT-1* is based on a contextualized graph construction connecting the premise and the conclusion using commonsense knowledge from ConceptNet (Speer et al., 2017). The algorithm to construct the connecting commonsense graph extracts concepts from the premise and conclusion and searches ConceptNet for shortest paths between premise and conclusion concepts, using SBERT to ensure semantic relatedness of the extracted paths to the argument. 13 classic graph features extracted from the constructed knowledge graph, as well as the SBERT similarity between premise and conclusion form a feature vector. This feature vector is fed to a linear SVM classifier for joint ValNov prediction.

Submission *ACCEPT-1* yields the 3rd-best shared task results (43.1), with the overall best novelty score of 70, while validity ranks close to the baseline NLI RoBERTa model (59.2). Two additional runs ablate i) the SBERT component for graph construction (*ACCEPT-3*), which incurs a large drop for novelty and a slight reduction for validity, and ii) separate feature extraction and prediction of validity and novelty scores (*ACCEPT-2*), which decreases the overall ValNov-score by 13 points (from 43.1 to 30.1).

Team CSS submitted one approach (Alshomary and Stahl, 2022). The system relies on a large RoBERTa model fine-tuned for NLI. In a transfer learning setting, this model is further fine-tuned on the training data of the shared task. Two prediction heads, one for validity and one for novelty, are used. For each prediction head, the other metric is used as an auxiliary task. Each prediction head is trained with its own set of hyper-parameters, but the RoBERTa model is shared. *CSS* ranks fifth in Subtask A with a ValNov score of 42.4. The model achieves a strong Validity score of 70.8, and a Novelty score of 59.9.

[4]No description paper was submitted for these systems.

Submission	Val/Nov	Validity	Novelty
NLP@UIT	41.50	44.60	38.39
AXiS@EdUni	29.16	32.47	25.86
Baseline	21.46	19.82	23.09

Table 4: Results (avg/ macro-F1-scores) for subtask B.

Remaining submissions Team **NLP@UIT** and team **HARSHAD** submitted one submission each, using fine-tuned transformer models. Team **NLP@UIT** has minor success with training an SBERT (Reimers and Gurevych, 2019) system (25.9 ValNov-score), while team **HARSHAD** underperforms the baseline with a BERT model fine-tuned for validity (56.31), which they also use to rate the novelty aspect (39.00), a result that underpins the dueling nature of the two aspects.

Combining the best approaches for each aspect Copying the highest-ranked validity predictions from the third submission of Team **CLTeamL** (van der Meer et al., 2022) and the highest-ranked novelty predictions from the first approach of Team **ACCEPT**, we compute a ValNov-score by joining their respective independent predictions, which represents an increase of 8.1 macro-F_1 points in predicting both validity and novelty correctly (53.3). This combination of these two systems' outputs performs best for correctly identifying valid and non-novel samples, with an F_1 score of 66.2 for this class.

4.2 Subtask B

For subtask B (Table 4) we got only two submissions. Team **NLP@UIT** was successful by training SBERT (Reimers and Gurevych, 2019) with a triplet loss objective function. It obtains the highest F_1-scores for validity (44.6) and novelty (38.39). Team **AXiS@EdUni**, with the second best system in Subtask A, reuse their system to predict validity and novelty for both conclusions presented in a sample. Since the output of their system is continuous, mapped to one specific class for subtask A, they can compare the validity and novelty predictions for each conclusion, taking the conclusion with a higher predicted score as superior in validity and novelty, respectively. Hence, they never assign a sample as a tie for validity and novelty, respectively, which lowers their results to the second best ValNov-score (29.2) in this subtask.

5 Discussion

The results for the submitted systems suggest that the prediction of validity seems to be an easier task compared to the prediction of novelty, as many submitted system reach higher scores on validity than on novelty prediction for both subtasks (computed mean scores of 62.74 for validity vs. 52.97 points F_1-scores for novelty, across all system submissions in Subtask A). Most of the submitted systems (CLTeamL, AXiS@EdUni-1, CSS, HARSHAD) rely on large pre-trained languages models (e.g. GPT-3, RoBERTa, BART) that are i) fine-tuned on task-specific data, ii) pre-trained on related tasks (NLI, ArgRel), iii) or are used as generators conditioned on selected prompts, as well as combinations of these.

Systems relying on large language models achieve strong results in terms of validity prediction. The fact that the best results are achieved with the huge GPT-3 system, pretrained with a massive amount of textual data, and relying on prompts to condition the generation without being fine-tuned for the specific task is remarkable. Pre-training on the related task of NLI has been shown, in many submissions, to be beneficial for the task, whereas Argument Relation Classification was not found to be similarly effective (see results for CLTeamL). Further, Multi-Task-Learning, aimed towards exploiting interactions between both quality labels has been shown to improve performance, having a joint instead of separate prediction of the two target labels, which corroborates their complementary nature.

Analysis of validity prediction In general, the results demonstrate that systems relying on large language models can achieve reasonable results in terms of *validity* predictions, hinting at the fact that they are capable of recognizing some sort of inference. This is supported by the fact that such models are familiar with coherence due to their pretraining process and have been shown to yield good results on popular natural language inference tasks in general (Raffel et al., 2020). Nevertheless, recent work has shown that models tend to rely on statistical cues rather than actually learning valid and general rules of inference (Niven and Kao, 2019; Zhang et al., 2022).

Analysis of novelty prediction Regarding the prediction of *novelty*, the systems based on large language models show worse performance. The best result in terms of novelty is achieved by a system from Team ACCEPT that integrates symbolic knowledge from external commonsense knowledge sources, followed by Team AXiS@EdUni, which uses the WikiData knowledge graph. This suggests that the prediction of novelty requires deeper reasoning abilities in combination with background and common sense knowledge.

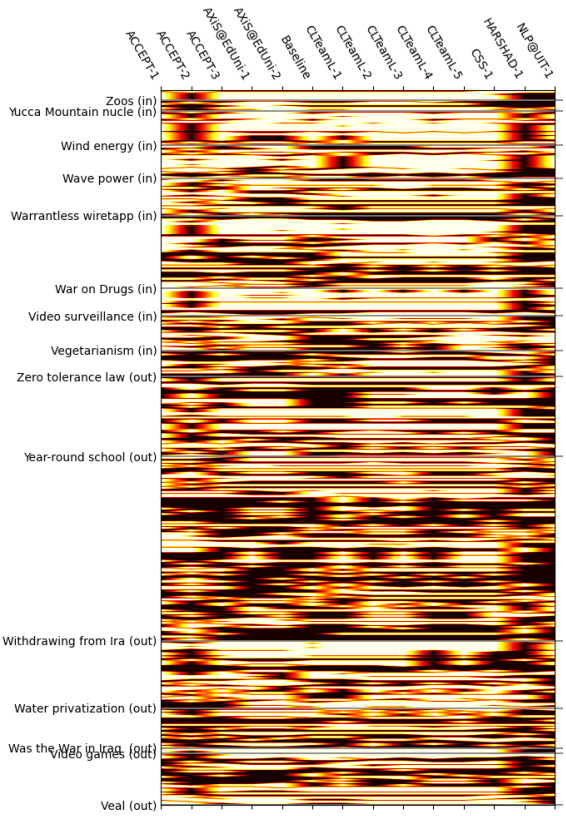

Figure 1: Error heatmap of each prediction and submitted system. The x-axis lists the submitted systems and the y-axis the instances grouped by topics. A topic marked with *out* does not occur in the other splits of the dataset, *in*-topics are also included in the validation-split. Red and dark areas represent misclassified instances.

Analysis of the difficulty of test topics and test instances for Subtask A We investigate the effect of individual instances and topics on the performance of the submitted systems with respect to the ValNov-score in Figure 1. In general, we observe that some instances seem more challenging than others. While 14% of all instances are correctly classified by at least 11 systems (out of 14), 23% of all instances are hard to classify (three or fewer systems correctly classifying them). 5% of all instances are never correctly classified. While detecting off-topic conclusions as neither valid nor

novel is easy for all systems, detecting many non-valid but novel instances is challenging. Also, some non-novel-non-valid instances are always misclassified in case of topic-related conclusions. One of those challenging examples is "*Economically speaking, using unwanted calves for veal is more efficient and socially desirable result than simply wasting this good and valuable meat.*" with the conclusion "*Veal is more economical than wasting good meat*". On the surface level, the conclusion looks like a valid-non-novel summarization, but it does not make sense in this wording for us humans. This example highlights the risks of relying on statistical cues. We also observe that the prediction success also depends on the complexity of the premise, explaining the larger misclassified areas in Figure 1. However, besides a common ground of difficulty shared by all systems, 3% of all instances are mostly correctly classified by the systems integrating background knowledge (ACCEPT-1 to AXis@EdUni-2 in Figure 1) but consistently misclassified of those that focus on large language models (CLTeam-1 to CSS-1 in Figure 1) and 2% of all instances for the reverse case.

Looking at the topic level, we observe that some topics are more challenging than others. For example, the discussion about "Withdrawing from Iraq" requires lots of (expert) background knowledge about US foreign policy and is, in addition, not a current topic anymore.[5] Looking at this topic, only 4.6 out of 14 systems correctly classify an instance on average. On the other hand, "Wind energy" is a much more common and current topic, with 7.3 systems correctly predicting the instances in this topic on average. The fact that "Withdrawing from Iraq" is an important topic in the test split that does not occur in the other splits intensifies the effect the low performance of some systems on novel topics (5.3 systems on average classify examples in novel test topics correctly versus 7.2 systems in test topics shared with the development set), by also showing that systems have difficulties in generalizing to unseen topics. A large amount of topics shared between train and test would surely increase results on test data, but would provide a misleading picture regarding the ability of systems to generalize across topics.

[5]Outdated discussions or topics with fading relevance can harm the performance of modern language models due to the pressure of keeping them in sync with the real world (Lazaridou et al., 2021) and the phenomena of catastrophic forgetting.

6 Conclusion

In this paper we have described the shared task on validity and novelty prediction that has been carried out as part of the 9th Argument Mining Workshop. Six groups participated in the task, submitting 15 system runs overall, with a preference for the more course-grained first subtask with binary labels for validity and novelty. The results suggest that validity is easier to predict compared to novelty. Large language models which are few-shot prompted or fine-tuned on the provided task-specific data, especially by applying transfer-learning from natural inference tasks, perform reasonably well on the task of predicting validity. However, such systems have a notably worse performance on predicting novelty. Systems that complement large pre-trained language models with external commonsense or world knowledge, by contrast, perform much better for novelty. This suggests that the recognition of novel content is challenging, requiring deeper understanding and inference involving background, common sense or even domain-specific knowledge.

Acknowledgements

This work has been funded by DFG within the project ACCEPT, which is part of the priority program "Robust Argumentation Machines" (RATIO).

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. Association for Computational Linguistics.

Milad Alshomary and Maja Stahl. 2022. Argument novelty and validity assessment via multitask and transfer learning. In *Proceedings of the 9th Workshop on Argument Mining*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 39–49, Online. Association for Computational Linguistics.

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.

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.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

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

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomás Kociský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 29348–29363.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. 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.

Niklas Rach, Carolin Schindler, Isabel Feustel, Johannes Daxenberger, Wolfgang Minker, and Stefan Ultes. 2021. From argument search to argumentative dialogue: A topic-independent approach to argument acquisition for dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2021, Singapore and Online, July 29-31, 2021*, pages 368–379. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ameer Saadat-Yazdi, Xue Li, Sandrine Chausson, Vaishak Belle, Björn Ross, Jeff Z. Pan, and Nadin Kökciyan. 2022. Kevin: A knowledge enhanced validity and novelty classifier for arguments. In *Proceedings of the 9th Workshop on Argument Mining*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

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.

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.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets

and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Michael van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Báez Santamaría. 2022. Will it blend? mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. 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, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark. Association for Computational Linguistics.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017c. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain. Association for Computational Linguistics.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. 2022. On the paradox of learning to reason from data. ArXiv.

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 Details about the baseline model for both subtasks

We fine-tuned a RoBERTa-base-model (`https://huggingface.co/roberta-base`) once for validity prediction and once for novelty prediction by using the training data for Subtask A and Subtask B, respectively. In Subtask A, we ignored the training data with *"unknown"* labels. We tuned each RoBERTa model for three epochs and loaded the best performing model regrading the loss score on the development split at the end. The baseline can be reproduced by running the python script located at `https://github.com/phhei/ArgsValidNovel/blob/gh-pages/BaselinePrediction/main.py`.

B Further analysis of the test predictions for Subtask A

Besides Figure 1 presenting misclassfied areas with respect to the ValNov-score (a instance is misclassified if the prediction for validity or for novelty is incorrect), we show Figure 2 for classification errors in validity predictions and Figure 3 for classification errors in the novelty predictions.

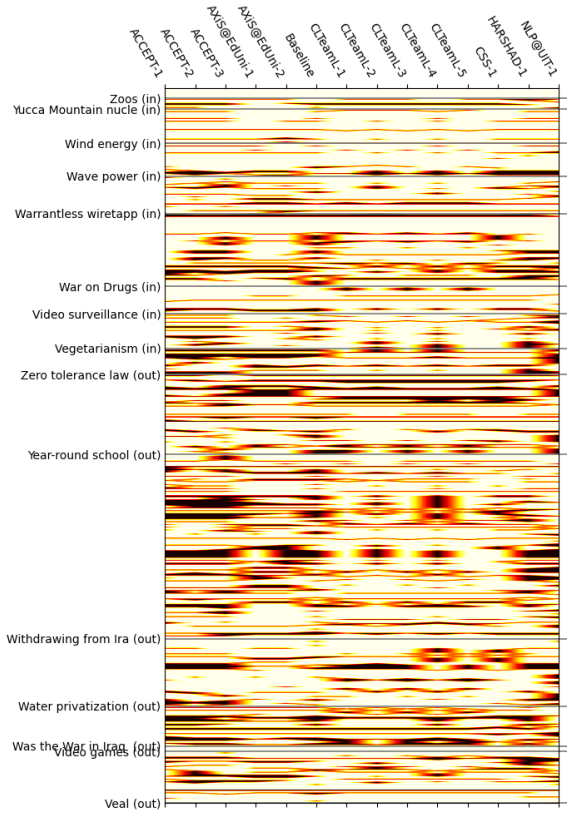

Figure 2: Error heatmap of each prediction and submitted system. The x-axis lists the submitted systems and the y-axis instances grouped by topic. A topic marked with *out* does not occur in the other splits of the dataset, *in*-topics are also included in the validation-split. Red and dark areas represent misclassified instances validity.

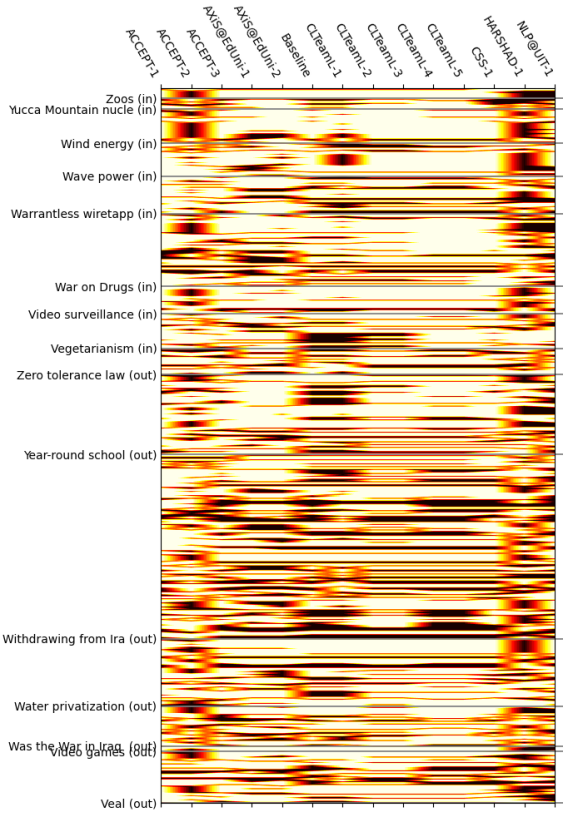

Figure 3: Error heatmap of each prediction and submitted system. The x-axis lists the submitted systems and the y-axis the instances group by topic. A topic marked with *out* does not occur in the other splits of the dataset, *in*-topics are also included in the validation-split. Red and dark areas represent misclassified instances novelty.