Binary Validity-Novelty-Classification

Hannes Ill
Deep Learning for Natural Language Processing – SS23 – Philipp Cimiano and Philipp Heinisch

September 29, 2023

1 Introduction

The systematic evaluation of arguments is an important task in the field of argument mining. Recent endeavors, such as the framework proposed by Wachsmuth et al., underscore the academic interest in understanding the dimensions of argument quality, with logic being a prime focus. With the surge in automated content generation, there's a pressing need to assess the quality of arguments automatically to establish trust in digital content.

Heinisch, Frank, Opitz, Plenz, et al. created a shared task on validity and novelty prediction, with the idea of reducing the assessment of an argument's overall quality with whether its conclusion is valid and novel given its premise. They provide a dataset and introduce different approaches in their overview paper. Subtask A, titled "Binary Validity-Novelty-Classification", demands classifying the validity and novelty of arguments as either valid or not valid, as well as novel or not novel. While the former assesses if a conclusion logically ensues from its premise, the latter ensures that the conclusion doesn't merely paraphrase the premise but offers novel insight.

In this paper, I present my research where I harness the power of deep learning architectures to tackle Subtask A. [1] I implement two distinct baseline models: a Long Short-Term Memory (LSTM) network and a RoBERTa transformer. These models serve as a foundational benchmark to compare against the enhanced approach.

The primary contribution of this paper is an enhanced approach to Subtask A that makes use of data preprocessing. By eliminating class imbalances and filtering out samples with a low confidence, I aim to boost the robustness and performance of my model.

2 Related work

Argument quality assessment is at the heart of this study. As, generally, it is not easy to assess an argument's quality, the assessment can be simplified to classify whether the argument's conclusion is valid and novel given its premise. This is related to evaluating the cogency dimension of an argument's quality, as introduced by Wachsmuth et al. It is important to note that validity and novelty can counteract each other. It is difficult to formulate arguments that satisfy both validity and novelty. In contrast, for example, it is rather simple to generate a valid but not novel argument by simply repeating or paraphrasing the premise in the conclusion (Heinisch, Frank, Opitz, and Cimiano, 2022).

To test different approaches for tackling binary validity-novelty-classification, Heinisch, Frank, Opitz, Plenz, et al. presented a shared task regarding this. The best approaches submitted to them and published by them in their overview paper of this task achieved macro-F1 scores of around 75% for predicting validity, 70% for predicting novelty, and 45% for predicting both correctly.

RoBERTa an optimized variant of the BERT model, was developed by Zhuang et al. While maintaining the core architecture of BERT, RoBERTa made improvements in training strategy, including using more data, larger mini-batches, and a revised masking approach. These tweaks allowed

[1] The code can be found at `https://github.com/hannesill/Validity-Novelty-Classification`

RoBERTa to outperform BERT in multiple tasks. In the context of binary validity-novelty-classification, understanding context is crucial. By comparing subtleties across the training samples, it has a chance to approximate the notions of validity and novelty of an argument.

For comparing different NLP models, both LSTM networks and the RoBERTa transformer model have emerged as promising candidates. This research tests out both of them to see how they perform at the task of binary validity-novelty-classification.

3 Method and Model

3.1 Data Preparation

Achieving good scores in the task of binary validity-novelty-classification with NLP models requires data preparation. As I test two different neural network architectures on this data, it is important to make the data compatible with the models' input and structure requirements.

The provided dataset was structured into the components "Topic," "Premise," "Conclusion," and labels for both validity and novelty. Also, the data contained confidence measures for both the validity and novelty label. The confident levels are "defeasible", "majority", "confident", and "very confident". What they mean gets explained in 3.4. For the data preparation it is only important to know that for labels with the confidence level "defeasible" the three human annotators could not find consensus. Therefore, I filtered out all samples with confidence levels of "defeasible". This corresponds to filtering out all samples with labels of "0", as all samples where the annotators could find some consensus were either labeled with "-1" or "1".

Because filtering out samples were any of the two labels is "0" would eliminate perhaps important information from the other label, I split the two dataset into a validity and novelty dataset. This way, the most amount of information can be retained after filtering out samples with labels of "0".

Then, labels with a value of "-1" were transformed to "0" to make sure that the data samples work with binary classification methods.

Next, the topic, premise, and conclusion were concatenated to form one sentence. Special tokens are used as prefixes at the beginning of the topic, premise, and conclusion. This helps the model differentiate the topic, premise, and conclusion in the constructed sentence. To maintain a clean and consistent format, special characters, excluding points and colons, were deleted. Also, all characters were transformed into lowercase characters.

Afterward, the vocabulary was constructed, comprising unique words encountered across the dataset, along with a few special tokens.

To ensure that the data was suitably formatted for the LSTM network, a custom collator was employed. This collator tokenized and padded each sentence, making them suitable for batch processing. For the RoBERTa transformer, a custom, pre-defined tokenizer was employed.

My enhanced approach consisted of additional data preprocessing, which I explain in 3.4 in detail.

3.2 LSTM Baseline

For the first baseline of this study, I implemented a RNN with LSTM cells. It offers a comparison for the enhanced approach. The foundational layer of this model is an embedding layer, adeptly transforming integer-encoded tokens from input sentences into embedding vectors. The embedded sentences are then funneled through a stack of eight LSTM layers. Additionally, dropout regularization is used between these layers to lower the chance for overfitting.

Afterward, the final hidden state of the last LSTM layer is passed through a fully connected linear layer. This layer serves as the classifier that returns the classes' probability values.

For the training, I employed the Binary Cross-Entropy Loss, optimized via the Adam optimizer with a learning rate of 0.01. The training process is monitored by evaluating the model after each epoch on the c, using the F1 score as the chosen metric. This consistent checkpointing offered a real-time perspective into the model's evolution and generalization capabilities. Upon completion of training, the model was further assessed on the test set. The predictions were saved to a .csv file for

future analysis. For this, the labels with "0" were transformed back to "-1" to match the labels of the samples in the provided datasets.

3.3 Transformer Baseline

The second baseline is a pre-trained transformer that is fine-tuned on the training set of the validity-novelty-classification task.

The implementation employs the RoBERTa transformer model for text classification in order to classify the validity and novelty of arguments. The process is initiated by setting a deterministic computational environment to make the models comparable across different training runs.

A pre-trained RoBERTa tokenizer is extended with the special prefix tokens. Following the tokenization and encoding of the training, validation, and test datasets, a fresh instance of the RoBERTa sequence classification model is initialized for each task. Afterward, the two models are trained separately for validity classification and novelty classification. Therefore, they also get two different datasets as described in 3.1.

Finally, predictions are saved to a .csv file in the same fashion as they were saved for the LSTM model.

3.4 Transformer Enhanced

For my goal to enhance the transformer model for the task of validity-novelty-classification, I employed further data preprocessing. This contains two preprocessing steps.

First, I selectively filtered the dataset based on annotators' confidence levels during their validity judgments. These were categorized as "defeasible", "majority", "confident", and "very confident". A label of "very confident" was attributed when all three annotators were in unison. Conversely, "defeasible" indicated a lack of consensus, suggesting the sample's subjective nature. A confidence level of "confident" means that two annotators agree and one is unsure, and "majority" means that two annotators agree and one disagrees. The same confidence levels hold for novelty judgments. By excluding "majority" or "defeasible" labeled samples, my hypothesis is that the model would encounter more straightforward examples, potentially simplifying its learning trajectory.

Secondly, after this confidence-based filtering, I observed class imbalances, particularly in the novelty category (397 not novel against 40 novel instances). I also observed a mild disparity in the validity classification (289 valid versus 259 invalid). To eliminate these class imbalances, I adopted oversampling to ensure the balance between classes. My hope was that these modifications would lead to more nuanced learning.

3.5 Evaluation

In my study, I have chosen the macro F1 score to be the measure for evaluating the model's performance. This is because Heinisch, Frank, Opitz, Plenz, et al. also use the macro F1 score to compare different models' performances.

Three distinct datasets were provided by their shared task: one for training, another for validation, and the last for testing. After some adjustments and finetuning based on the validation set's feedback, I tested the models on the test dataset. I computed the macro F1 scores for three areas: validity classification, novelty classification, and combined validity-novelty classification. Notably, in the combined assessment, both validity and novelty predictions must be correct for the overall prediction to be deemed accurate.

4 Results

In this chapter, I will analyze and interpret the results of the binary validity-novelty classification experiments. Their evaluation scores of the models are shown in Table 1.

Both the LSTM and the transformer baseline yielded the same macro F1 scores for validity, novelty, and the combined validity-novelty classification. Precisely, they achieved an F1 score of 0.3765

Model	Validity F1	Novelty F1	Both F1
LSTM	0.3765	0.3612	0.1307
Transformer (Baseline)	0.3765	0.3612	0.1307
Transformer (Enhanced)	**0.5830**	**0.4334**	**0.2804**

Table 1: Comparison of macro F1 scores for validity, novelty, and combined validity-novelty classification across different model architectures. The enhanced transformer model outperforms the baseline and LSTM variants.

for validity, 0.3612 for novelty, and a notably lower score of 0.1307 when considering both tasks simultaneously. On closer examination, it is evident that these models effectively predicted the majority class from the training set, failing to capture the intricate patterns and relationships within the data. This behavior was expected for the LSTM as it only trained for three epochs, and in order to learn to classify the validity and novelty of an argument, a thorough text understanding is necessary. This is simply not possible with so little training. However, the results of the transformer baseline are a little surprising as it is pre-trained and should already have some text understanding. The score of the novelty class can be explained by the huge class imbalance, but I expected the validity F1 score to be higher. My hypothesis is that more extensive hyperparameter tuning would bring the validity F1 score of the baseline transformer up.

In contrast, the transformer model in the enhanced approach achieved a significant improvement over the baselines. For validity classification, its F1 score surged to 0.5830, and for novelty, it rose to 0.4334. Most importantly, when assessing both tasks together, the F1 score achieved was 0.2804, more than double the score of the baselines. This enhancement in the model's performance can be attributed to the confidence-based filtering and class-balancing data preprocessing. By only seeing samples labeled confidently the learning is more straightforward. And by ensuring a more even representation of classes, the model became better at identifying less frequent patterns and thus outperformed the baselines, which were previously skewed by the majority class.

Regarding the task of binary validity-novelty classification, these results indicate that data preprocessing is important and necessary. An enhanced preprocessing approach, like the one implemented and tested in this study, is an example for this.

5 Conclusion

This study shows the significance of data preprocessing in the domain of binary validity-novelty classification. Starting from the foundational benchmarks presented by LSTM and transformer baselines, it was observed that even sophisticated models like the transformer, which possess a high degree of text understanding from pre-training, can fall short without proper preprocessing. The equal F1 scores of the baselines indicate that models might gravitate towards predicting the majority class, especially when confronted with uncertain samples or class imbalances.

Also, this study tried different data augmentation approaches like paraphrasing the sentences by using pre-trained paraphrasing models, back translation with translation models, or synonym replacement. All of these approaches did not increase the model's performance. Hence, they were not adapted for the enhanced approach, but indicate that it is not trivial to augment argument datasets.

In summary, this study shows how important data preprocessing is for the task of argument quality assessment. As the amount of digital content continues to rise drastically, the findings presented here might help to guide future research to gain trust in automatically generated content. Future research could explore more advanced preprocessing techniques and consider how to combine them with beneficial data augmentation.

References

Heinisch, Philipp, Anette Frank, Juri Opitz, and Philipp Cimiano (July 2022). "Strategies for framing argumentative conclusion generation". In: *Proceedings of the 15th International Conference on Natural Language Generation*. Waterville, Maine, USA and virtual meeting: Association for Computational Linguistics, pp. 246–259. URL: https://aclanthology.org/2022.inlg-main.20.

Heinisch, Philipp, Anette Frank, Juri Opitz, Moritz Plenz, et al. (Oct. 2022). "Overview of the 2022 Validity and Novelty Prediction Shared Task". In: *Proceedings of the 9th Workshop on Argument Mining*. Online and in Gyeongju, Republic of Korea: International Conference on Computational Linguistics, pp. 84–94. URL: https://aclanthology.org/2022.argmining-1.7.

Wachsmuth, Henning et al. (Apr. 2017). "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*. Valencia, Spain: Association for Computational Linguistics, pp. 176–187. URL: https://aclanthology.org/E17-1017.

Zhuang, Liu et al. (Aug. 2021). "A Robustly Optimized BERT Pre-training Approach with Post-training". English. In: *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Huhhot, China: Chinese Information Processing Society of China, pp. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108.