Argumentation Mining from Textual Documents Combining Deep Learning and Reasoning

Filipe Cerveira do Amaral[1,2(✉)], H. Sofia Pinto[1,2], and Bruno Martins[1,2]

[1] Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
{filipe.amaral,sofia.pinto,bruno.g.martins}@tecnico.ulisboa.pt
[2] INESC-ID, Lisboa, Portugal

Abstract. Argumentation Mining (AM) is a growing sub-field of Natural Language Processing (NLP) which aims at extracting argumentative structures from text. In this work, neural learning and symbolic reasoning are combined in a system named N-SAUR, that extracts the argumentative structures present in a collection of texts, and then assesses each argument's strength. The extraction is based on Toulmin's model and the result quality surpasses previous approaches over an existing benchmark. Complementary scores are also extracted and combined with a set of rules that produce the final calculation of argument strength. The performance of the system was evaluated through human assessments. Users can also interact with the system in various ways, allowing for the strength calculation to change through user-cooperative reasoning.

Keywords: Natural language understanding · Machine learning · Neural-symbolic approaches · Argumentation mining

1 Introduction

Argumentation and debate constitute a vital part of human societies. Recent developments in artificial intelligence are being applied to analyzing and simulating this human cognitive activity. In particular, Argumentation Mining (AM) is a growing sub-field of Natural Language Processing (NLP) with many practical applications in areas that include law, healthcare, and e-government [1], aiming at the extraction of the argument structures present in text. This includes perceiving the core idea of a document, and further distinguishing relevant from irrelevant segments of text (i.e., the Argumentative Discourse Units (ADUs) [13]) with respect to that idea, together with the relations that may exist among them, according to a chosen argumentation model.

Understanding whether a given piece of evidence supports a given claim, or whether two claims attack each other, are complex problems that humans can address thanks to their ability to exploit commonsense knowledge to perform reasoning and inference. Despite the remarkable impact of deep neural networks

N. Moniz et al. (Eds.): EPIA 2023, LNAI 14115, pp. 389–401, 2023.
https://doi.org/10.1007/978-3-031-49008-8_31

in NLP, these techniques alone will likely not suffice to address the complex issues associated with AM, motivating the development of techniques that explicitly consider knowledge expressed in the form of rules and constraints.

This paper presents a Neural-Symbolic Argumentation Mining system with User-Cooperative Reasoning (N-SAUR), which uses neural technology to extract and label ADUs from texts, based on the Toulmin argumentation model [15], followed by a symbolic AI approach to reason about the previously collected information. Reasoning is performed through a set of Problog rules, which attempt to calculate how strong a given claim is, in regards to its argumentative power.

2 Related Work

We now present relevant sub-symbolic and neural-symbolic approaches, previously described in the literature.

Sub-Symbolic Methods to Argumentation Mining: In [16], two types of Argumentation Mining (AM) are defined: close-domain discourse level and information-seeking. The former refers to identifying the structure of the arguments, and the latter to extracting the implicit meaning of the arguments, as well as the claims they are defending. The authors defined an argument as an atomic piece of text, thus not requiring any context. Using BERT and FLAIR models to implement a slot-filling approach based on token-level annotations, they were able to extract ADUs. Data supporting the experiments consisted of a large amount of text on controversial topics, so that discussion was present and arguments from different points of view would appear.

Open-domain AM was tackled in [11], searching in a pool of documents for arguments on a given topic, both supporting and refuting the main topic. The results emphasize the importance of (1) using contextualized word embeddings (e.g., from BERT), and (2) the need for clustering similar arguments, to simplify the argumentative structure before classification. An approach to AM across different languages was described in [3], emphasizing the importance of training on same-domain corpora, as the context and the theme of an argumentative text are crucial for it to be understood.

Regarding argumentative models, in [16] the authors used a simple binary PRO/CON classification scheme, in which ADUs are either for or against the topic of the text at hand. In [9], the authors used a slightly more complex structure, in which an ADU could either be a premise, a claim, or a major claim. Premises support claims, which in turn either support or attack major claims, of which a text can have just a few. Other authors have instead used a modified Toulmin model [7], represented on the right-hand side of Fig. 1 and based on the original model shown at the left-hand side. This model features claims, that can be supported by either backing or grounds, and attacked by rebuttals, which can, in turn, be attacked by refutations.

Neural-Symbolic Methods to Argumentation Mining: An extensive analysis on the combination of learning and reasoning is described in [6]. The former

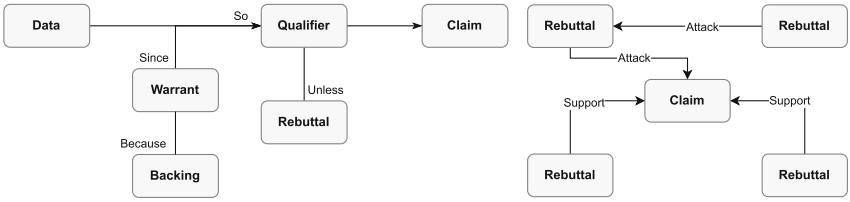

Fig. 1. Toulmin's original model and the modified Toulmin model [7], respectively on the left and right-hand sides. Both diagrams are adapted from [7].

is usually implemented with neural techniques and the latter with symbolic approaches. The authors propose the combination of both approaches, naming it neural-symbolic computing. On the other hand, the authors of [5] called for a "*leap forward*" in Argumentation Mining (AM) by integrating both sub-symbolic and symbolic approaches. They showed some examples of logical inference tools (including defeasible inference and DeepProbLog [10]).

A new task, Argument Explicitation (AE), was proposed in [9]. Besides detecting the arguments present in a text, AE also (a) tries to explain them (i.e., enhancing them with knowledge that is retrieved on-the-fly), (b) tries to make the missing premises (called enthymemes) explicit, and (c) assesses the arguments' validity, by either fact-checking or, in case of a subjective conclusion, performing enthymeme reconstruction, making it evident for the user that the argument that is being made assumes a given implicit premise, and making it the user's responsibility to assess its validity. In the same paper, a framework for the task is proposed, combining different kinds of explicitation.

3 The N-SAUR System

We propose a Neural-Symbolic Argumentation Mining with User-Cooperative Reasoning (N-SAUR) system, which receives as input a collection of argumentative texts that fall into a single topic of discussion, and then calculates the strength of each claim that is present in the documents, scoring them with a value from 0 to 1 that can be read as the likelihood of each claim being a strong one, with regard to its argumentative power.

N-SAUR is divided into three tasks (see left-hand side of Fig. 2), namely (a) recognition of all the ADUs in the document collection, and (b) complementary ADU scoring, which is itself composed of (b1) quality and natural language inference scoring, and (b2) the final logical calculation.

ADU Recognition and Expression: After receiving the text collection as input, the first task the system performs is individual to each text, and concerns detecting and classifying all relevant spans of text for the argumentation that is taking place in every single document. The encoded information includes each ADU's text content, as well as their classifications and relations, according to the modified Toulmin model (right-hand side of Fig. 1).

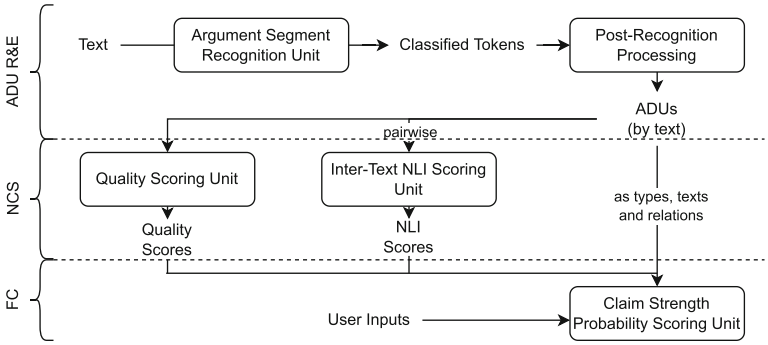

Fig. 2. The N-SAUR system architecture.

Neural Complementary Scoring: With all ADUs extracted and their classifications and relations encoded, the next components of the system aim at perceiving how strong, with regard to its argumentative power, each claim is. The local context of an argument, i.e. all the argumentative structure leading to the final idea, i.e., the claim, is represented by the expressed classifications and relations among the extracted ADUs. The more a claim is supported or attacked, the stronger or weaker it is, respectively. Each extracted ADU will be scored according to its inherent argumentative power, by the **Quality Scoring Unit**. The stronger the quality of an ADU that is supporting or attacking a claim, the stronger or weaker that claim should be, respectively.

Additionally, the system will assess the relations between all the ADUs from documents about the same topic. For that, all possible pairs of ADUs are scored by the **Inter-Text Natural Language Inference Scoring Unit**. This information enables N-SAUR to derive additional support/attack relations, thus getting more data to assess a claim's strength. A claim will be supported or attacked by an ADU if that ADU is entailed by another one that in turn supports or attacks the claim, respectively.

Final Calculation: With all the previously gathered scores, the system is now able to calculate the argumentative strength of each claim in the collection of input documents, using a set of rules and probabilistic logic. Additionally, N-SAUR enables users' input for a claim's strength calculation. We call this ability user-cooperative reasoning.

Argument Segment Recognition Unit: The first component in the N-SAUR pipeline aims at detecting the relevant spans for the argumentative structure of each text, as well as the corresponding classifications. All word tokens are classified using a BIO encoding (Beginning, Inside, Outside) for the classes claim, grounds, backing, rebuttal, or refutation. Several deep learning BERT-based models from the HuggingFace Transformers library were tested, fine-tuning them to this task with the Argumentation in User-Generated Content Dataset [7]. The best model, BERT Large Cased [2], was chosen.

Post-Recognition Processing: With all tokens classified, the next module in the pipeline aggregates the tokens into the spans of text with a classification, expressing the ADUs. Afterwards, according to the modified Toulmin model, the system encodes the relations that exist among all ADUs from the same text. For a relation to occur, the distance between the given ADUs is not taken into consideration, and if a backing and a claim are detected in the same text they have a relation of support. If there is more than one possible relation, then all of them are encoded with equal probability for each target (e.g., two claims and one premise means that there is a 50% chance of the premise supporting either of the claims). This is the first part of the pipeline that introduces probabilistic logic, using Problog, which will then be propagated forward.

Natural Language Inference Scoring: All the extracted ADUs, as well as their classifications and relations, are organized into all possible inter-text pairs, in order to perform Natural Language Inference (NLI). This task takes two sentences and determines if there is a relation of entailment, contradiction, or neutrality between them. For that, a DeBERTa model [8] from the HuggingFace Transformers library, which is already fine-tuned for the NLI task, was used. The model produces, for each pair, a probability distribution for the three classes - neutrality, entailment, or contradiction. From all these results, the pairs for which neutrality has the highest probability are discarded. A sample of pairs is then drawn according to the highest majority probability, balanced according to the entailment and contradiction classes.

Quality Scoring Unit: In addition to the support it has, a claim's strength is also influenced by its own inherent quality. Besides that, it is also possible to consider the inherent quality of its supporters and attackers. This way, any given ADU that is a source of a relation to a given claim can have its quality measured and used as a weight of that relation.

Given the aforementioned ideas, all extracted ADUs, and not only claims, are scored for quality. Experiments were made with several other BERT-based models from the HuggingFace Transformers library (including RoBERTa) and BERT was chosen. This model was trained for quality scoring using the IBM-Rank [14] dataset, which contains statements and their corresponding scores, from 0 to 1. In order to perform the fine-tuning, a regression head was added to the output of BERT, so that the neural model outputs a score between 0 and 1. All ADUs already mined by the pipeline are scored in this way.

Claim Strength Probability Scoring: With all the previously gathered information, N-SAUR can now calculate how strong each of the extracted claims is, using Problog, which is used to encode all the information to be given as input to probabilistic reasoning. There are three sources of data to be expressed in Problog facts, namely the ADU extractions per se, and the results from the two scoring units.

Both classifications and relations are extracted simultaneously, since the latter are implicit in the former, given the argumentation model being used. The

following code snippet shows an example of the Problog encoding for a document with two claims. The predicate support(XX, YY) means XX supports YY.

```
1       claim(c1).
2       claim(c2).
3       grounds(g1).
4       0.5::support(g1, c1).
5       0.5::support(g1, c2).
```

For the other sub-processes of data, Problog facts are also used. In the following code snippet, the probability of each fact is given by the quality score and from the NLI result (i.e., the probability of the most likely label from the neural model, given that it is not neutral), respectively. The predicates entail(XX, YY) and contradict(WW, ZZ) mean XX entails YY and WW contradicts ZZ.

```
1       0.87::quality(c1).
2       0.78::quality(c2).
3       0.33::quality(g1).
4       0.85::quality(g2).
5       0.54::entail(g1, g2).
6       0.38::contradict(g2, c2).
```

The Problog facts are fed into the final calculation unit, which joins them with the strength calculation program. The following set of rules makes it possible to derive a support/attack relation from a combination of entailment and other support/attack relations, to which we add the quality score of the newly created relation's source.

```
1       support(A,C) :- support(A,B), entail(B,C), quality(A).
2       support(A,C) :- support(B,C), entail(B,A), quality(A).
```

Given all facts from previous code snippets, we can derive a support relation between $g2$ and $c1$, because of the entailment between $g1$ and $g2$.

The following code snippet shows the next abstraction level in the strength calculation. It is possible to compute how much a given claim is opposed by other ADUs, as given by how much it is attacked, and also according to whether there is a contradiction in the ADUs that support it (Lines 1–2). This is then used to calculate how much the claim is endorsed (Line 3). Furthermore, a claim can also be supported by user input, under a support form, being expressed as a different kind of endorsement (Line 4).

```
1       oppose(C) :- claim(C), attack(A,C), not(attack(_,A)).
2       oppose(C) :- claim(C), support(A,C), contradict(A,C).
3       endorse(C) :- claim(C), support(_,C), not(oppose(C)).
4       user_endorse(C) :- claim(C), user(A), support(A,C).
```

Finally, the claim's strength is given by combining the endorsement, the quality of the claim itself, and, if there is any, the users' inputs:

```
1       strength(C) :- endorse(C), quality(C); endorse(C),
            quality(C), user_endorse(C).
```

It is worth noting that although the previous rules are not assigned to any probabilities, the probabilistic facts will propagate them, yielding very expressive results. As such, each claim's strength will be calculated with regard to several probabilistic facts. Moreover, a user can change each fact's probability in order to check the results yielded by the change. It is also possible to vote directly on how strong a claim is. With several votes, expressed as a number from 0 to 1, N-SAUR will, arguably, get closer to the real strength of a given claim.

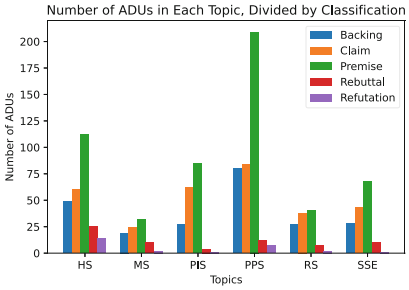

Fig. 3. The AUGC dataset distribution of ADU classes.

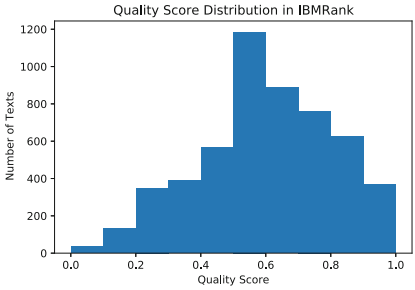

Fig. 4. Quality score distribution on the IBMRank dataset [14].

4 Experimental Setup

This section provides a characterization of the data supporting our experiments, together with the evaluation methodology and the corresponding results.

4.1 Data

N-SAUR has three neural units, two of which were fine-tuned before their usage: the Argument Segment Recognition and the Quality Scoring units. The Argumentation in User-Generated Content (AUGC) dataset [7] was used to fine-tune the BERT Large model used for argument span recognition. The dataset is divided into three parts: an unlabeled raw corpus, a version with annotations regarding persuasiveness, and a version with annotations regarding argument structures. For this work, only the last part was used. It contains 340 small texts about six different controversial topics relating to education: homeschooling, mainstreaming (including children with special needs into regular classes), prayer in schools, public versus private schools, redshirting (postponing children's kindergarten entrance), and single-sex education. All these texts were collected from internet users' posts on blogs or forums, comments, or articles. The distribution of the ADU classifications is represented in Fig. 3.

In turn, the IBMRank dataset [14] was used to train the model for quality scoring. To create this dataset, human annotators were asked to label sentences

as good or bad, as well as to decide what was the best one out of a random pair. These results were then converted into a numerical label and the data was cleansed according to factors such as the annotator agreement and the difference in quality between arguments. The IBMRank dataset contains 5297 sentences with lengths between 36 and 275 characters, which were split into training and testing folds of 90% and 10% of sentences, respectively. The quality score distribution across the data is shown in Fig. 4.

4.2 Evaluation Methodologies

N-SAUR can be evaluated both holistically and in terms of its specific underlying components, namely the neural models that were fine-tuned.

Assessment of the Argument Segment Recognition Unit: Topic-wise cross-validation was performed (see Sect. 4.1), and F1 scores were measured over all possible token classifications. Macro-F1 scores, across all topics by class and also across all classes by topic, were also calculated.

Assessment of the Quality Scoring Unit: We used the Mean Absolute Error (MAE) for model assessment, given by:

$$\mathrm{MAE}(r, p) = \frac{1}{n} \sum_{i=1}^{n} |r_i - p_i|, \tag{1}$$

where r_i and p_i refer to each n real and predicted values, respectively.

Holistic Assessment: To evaluate the final results, the performance of N-SAUR was compared to real people's assessments. To do that, the *public vs private schools* topic was chosen, since it was thought to be a topic most people have already an opinion about, and requires no further explanation. A sample of 30 claims was drawn from the total set of claims collected by N-SAUR. The participants were asked to compare pairs of claims and indicate which one was the best. Many times, however, the claim text itself is not enough, so a context was added, composed of all the ADUs that had a relation (attack or support) with the claim. Given this information, the goal was to get the ranking of 30 claims with pairwise comparisons. A reasonable number of pairs to sample was between 86 and 138, to have correct results with a probability of 2/3 [12]. Hence, 104 unique pairs were divided into 13 surveys of 8 questions each.

Two experiments were made with the gathered data. The first one aimed at contrasting the performance of N-SAUR with the annotators in regard to pairwise comparisons. For each pair of claims, the annotators' and N-SAUR's picks as the stronger claim were compared. The annotators' pick is determined by a majority vote, and the system's pick is the claim with the highest strength score. To refine the results, the pairs are organized by how much consensus there was among the annotators, namely from 50 to 66.7%, from 66.7 to 83.3%, and from 83.3 to 100%. For each bin, an accuracy score can be calculated by considering pairs in which both the annotators' and the system's picks for the stronger claim are the same, as successful.

The second experiment started by producing a ranking of all 30 claim pairs. To do so, each of the pair's majority stance was used for a Borda count method [4], in order to score each claim. On the other hand, N-SAUR also produces a ranking, by means of its assigned strength scores. The two rankings were compared with the Normalized Discounted Cumulative Gain (NDCG) [17], considering the result of the Borda count method as a relevance score. The NDCG value for a list of results p is given by:

$$nDCG_p = \frac{\sum_{i=1}^{p} \frac{2^{rel_i}-1}{\log_2(i+1)}}{\sum_{i=1}^{|REL_p|} \frac{2^{rel_i}-1}{\log_2(i+1)}}, \tag{2}$$

where rel_i is the relevance of the given item, and $|REL_p|$ is the ordered list of relevant items.

Table 1. Fine-tuning results of BERT Large Cased using the AUGC dataset.

Tags	MS	RS	HS	PPS	SSE	PIS	Macro-F1
O	0.957	0.999	0.994	0.988	0.993	0.999	0.988
B-Backing	0.872	0.966	0.898	0.922	0.967	0.947	0.929
I-Backing	0.941	0.997	0.994	0.983	0.988	0.95	0.976
B-Claim	0.788	0.857	0.949	0.841	0.929	0.903	0.878
I-Claim	0.853	0.918	0.986	0.933	0.939	0.927	0.926
B-Premise	0.886	0.932	0.934	0.959	0.919	0.942	0.929
I-Premise	0.957	0.996	0.986	0.984	0.985	0.989	0.983
B-Rebuttal	0.583	0.400	0.557	0.615	0.700	0.800	0.609
I-Rebuttal	0.929	0.847	0.841	0.869	0.952	0.729	0.861
B-Refutation	0.000	0.000	0.000	0.000	0.000	0.000	0.000
I-Refutation	0.000	0.442	0.806	0.879	0.065	0.000	0.365
Macro-F1	0.706	0.759	0.813	0.816	0.767	0.744	0.768
AUGC	0.188	0.257	0.197	0.203	0.194	0.166	0.201

4.3 Experimental Results

We now present and discuss the results obtained for each of the parts.

Assessment of the Argument Span Recognition Unit: The results of the fine-tuning for argument span recognition are presented in Table 1, divided into the 6 topics of the AUGC dataset, that constituted the folds for cross-validation training of a BERT Large Cased model. It is worth pointing out that our approach yielded better results than those presented for the original benchmark, which used an SVM classifier [7] in the same cross-domain setting.

The *public vs private schools* topic was chosen for subsequent analysis, and a confusion matrix for its results is presented in Fig. 7.

It is also possible to draw a comparison between the number of examples for each ADU class (see Fig. 3) and result quality. Even though state-of-the-art neural models surpassed the previous benchmark, the argumentation model is perhaps too complex to perform this kind of automated recognition, especially regarding the refutation class, thus justifying the tendency for choosing simpler models. On the other hand, the poor performance on some classes may be due to a lack of enough data (e.g., for the neural models to distinguish premises and refutations/rebuttals effectively).

Assessment of the Quality Scoring Unit: We measured an MAE of 0.173, which represents a small difference between predictions and real scores, yielding a good determinant of each ADU's quality, and consequently an accurate contribution to the pipeline.

Holistic Assessment: For the human vs N-SAUR performance comparison, 50 people responded to the aforementioned surveys. Almost all the participants (94%) were between 18 and 29 years old, and only a small minority (8%) claimed not to have previously watched nor participated in any debates. All pairs of claims were evaluated at least three times.

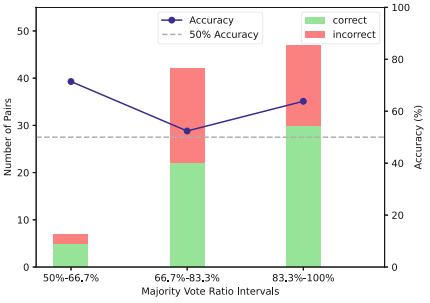

Fig. 5. N-SAUR pairwise performance by consensus.

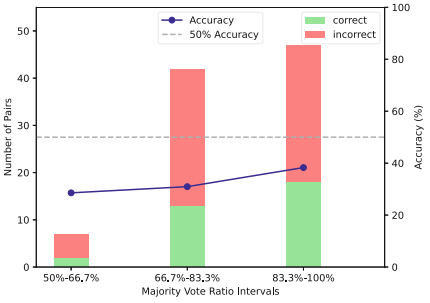

Fig. 6. N-SAUR pairwise performance without quality scoring.

Accuracy for the first experiment is presented in Fig. 5. It is possible to see that for all bins, there was a majority of correctly classified pairs, which is then reflected in the accuracy (curve in blue), being always above the 50% mark (grey dashed line). This measure is higher on the first bin, perhaps because of its small amount of pairs. On the other hand, with the remaining two bins, the system's performance goes up with the consensus of annotators' votes, as can be seen by the accuracy increase from the second to the third bin. In the second experiment, the overall ranking of N-SAUR, contrasted with the one produced by the annotations as the ground-truth, yielded an NDCG score of 0.830.

In order to further validate the previous results, another experiment was carried out, in which quality scoring was removed and the aforementioned assessment metrics were calculated. Rankings by the system were in this case performed according to rules without a quality predicate. The pairwise comparison results are represented in Fig. 6, which shows that the results are worse than the original N-SAUR. The NDCG experiment, however, measured a value of 0.853, being slightly better. These results show that sometimes a single measure may not be enough, and perhaps the NDCG is better because the sample size of 30 claims may have been too little. Nevertheless, further analysis with a larger quantity of annotations is required.

Overall, the experiments suggest that N-SAUR was capable of perceiving some of the same patterns humans use to make their judgments. The results also open the door for further investigation, aiming at establishing benchmarks considerably higher than 50% on the claim's strength prediction task.

5 Conclusions and Future Work

We presented a Neural-Symbolic Argumentation Mining with User–Cooperative Reasoning (N-SAUR) system, following the argument made in [5] for combining both symbolic and sub-symbolic techniques (with the additional optional user interaction) to perform Argumentation Mining (AM) and further reasoning with the results. We used state-of-the-art neural methods for NLP to extract the argumentative structures from text documents and classify them formally, followed by the gathering of natural language inference relations and quality scores. All

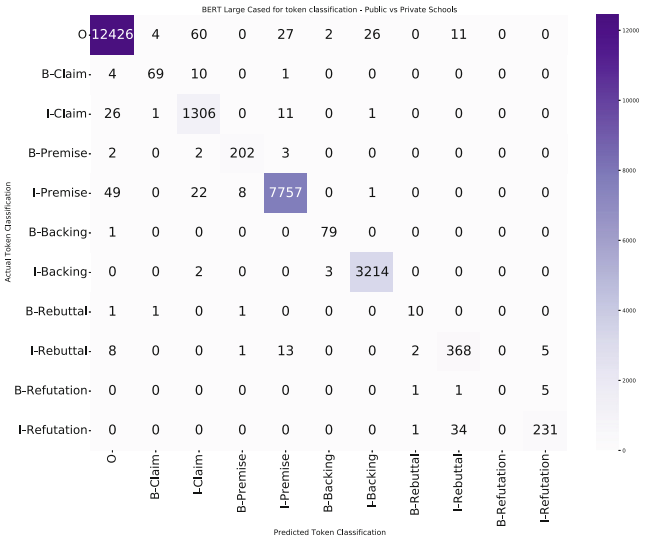

Fig. 7. Confusion matrix for the public vs private schools topic.

this information is encoded in Problog, and claim strength scores are calculated through probabilistic reasoning, with the option of user participation.

N-SAUR is innovative in various ways: it combines two approaches of AI, which are usually separate, applying them to an NLP sub-field of growing importance. It also used a complex argumentative model, surpassing existing benchmarks. Moreover, it embodies a framework in which more models, complementary scores, and even user votes can be added, with the end goal of coming increasingly closer to the real strength prediction of an argument clause.

Future work in neural-symbolic AM should consider tackling the problem with a more complex set of rules, attempt to cluster similar claims [11], and add more complementary neural scores (e.g., probabilities associated with ADU recognition), so as to yield more fine-grained results. Future work should also further investigate the calibration of the probability scores output by the neural models and add the user-cooperative aspect to the experiments. Moreover, a full integration of DeepProbLog [10], allowing for all the probabilities used in Problog to be computed end-to-end by a neural network, should be attempted.

Acknowledgements. This research was supported by Fundação para a Ciência e Tecnologia (FCT), through the INESC-ID multi-annual funding with reference UIDB/50021/2020 and by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003) which corresponds to the FCT reference CHIST-ERA/0001/2019.

References

1. Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G., Thimm, M., Villata, S.: Towards artificial argumentation. AI Mag. **38**(3), 25–36 (2017)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (2019)
3. Eger, S., Daxenberger, J., Stab, C., Gurevych, I.: Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In: Proceedings of the International Conference on Computational Linguistics (2018)
4. Emerson, P.: The Original Borda Count and Partial Voting. Soc. Choice Welfare **40**, 353–358 (2013)
5. Galassi, A., Kersting, K., Lippi, M., Shao, X., Torroni, P.: Neural-symbolic argumentation mining: an argument in favor of deep learning and reasoning. Frontiers in Big Data 2 (2020)
6. d'Avila Garcez, A.S., Gori, M., Lamb, L.C., Serafini, L., Spranger, M., Tran, S.N.: Neural-symbolic computing: an effective methodology for principled integration of machine learning and reasoning. J. Appl. Log. **6**(4), 611–632 (2019)
7. Habernal, I., Gurevych, I.: Argumentation mining in user-generated web discourse. Comput. Linguist. **43**(1), 125–179 (2017)
8. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with disentangled attention. In: Proceedings of the International Conference on Learning Representations (2021)

9. Hulpus, I., Kobbe, J., Meilicke, C., Stuckenschmidt, H., Becker, M., Opitz, J., Nastase, V., Frank, A.: Towards explaining natural language arguments with background knowledge. In: Joint Proceedings of the International Workshop on Dataset PROFlLing and Search, and the Workshop on Semantic Explainability, co-located with the International Semantic Web Conference (2019)
10. Manhaeve, R., Dumančić, S., Kimmig, A., Demeester, T., De Raedt, L.: Neural probabilistic logic programming in DeepProbLog. Artif. Intell. **298** (2021)
11. Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., Gurevych, I.: Classification and clustering of arguments with contextualized word embeddings. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2019)
12. Ren, W., Liu, J., Shroff, N.: Sample complexity bounds for active ranking from multi-wise comparisons. In: Proceedings of the Annual Meeting on Neural Information Processing Systems (2021)
13. Stede, M., Schneider, J.: Argumentation mining. Synth. Lect. Hum. Lang. Technol. **11**, 1–191 (2018)
14. Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., Slonim, N.: Automatic argument quality assessment-new datasets and methods. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2019)
15. Toulmin, S.E.: The Uses of Argument. Cambridge University Press (2003)
16. Trautmann, D., Daxenberger, J., Stab, C., Schutze, H., Gurevych, I.: Fine-grained argument unit recognition and classification. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
17. Wang, Y., Wang, L., Li, Y., He, D., Liu, T.Y., Chen, W.: A theoretical analysis of NDCG type ranking measures. In: Proceedings of the Conference on Learning Theory (2013)