Abstract Argumentation Frameworks Extraction for Dispute Resolution in Scientific Peer Review

Ildar Baimuratov[1,2,*], Alexandr Karpovich[3]

[1]L3S Research Center, Leibniz University Hannover, Germany
[2]TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany
[3]ITMO University, St Petersburg, Russia

Abstract

Peer review is a fundamental component of the academic editorial decision-making process, but it faces significant challenges such as the rapid growth in submission volumes and the reinforcement of existing biases in academia. While artificial intelligence may help resolve these challenges, its reliability remains questionable. In contrast, formal argumentation-theoretic frameworks provide a structured way to analyze and resolve argumentative disputes in peer review, but representing them within such frameworks can be time-consuming. In this work, we apply argument mining to streamline the formalization of peer review. Specifically, we combine existing argument identification techniques with argumentative relation extraction, enabling the construction of comprehensive abstract argumentation frameworks from peer review, which can then be resolved using argumentation solvers. By leveraging BERT embeddings and LSTM architecture, we achieve an F1 score of 63.05 for argument identification and 86.2 for relation extraction. A key advantage of our method is its transparency and controllability. At each step, human oversight is possible, allowing manual correction of model outputs, while the final dispute resolution is produced deterministically based on formal semantics. In real-world peer review scenarios, our method can support meta-reviewers and editors in making final decisions on manuscript acceptance.

Keywords

Abstract argumentation frameworks, Argument mining, Dispute resolution, Peer review, OWL

1. Introduction

The peer review process for scientific publications is becoming increasingly complex due to the rapid growth in submission volumes. Since 2013, the annual increase in the number of manuscripts submitted to peer-reviewed journals has been an unprecedented 6.1%, with a significant increase in the number of rejections [1]. More than 15 million hours are spent each year reviewing manuscripts that are initially rejected and subsequently resubmitted to other journals [2]. The peer review process is further complicated by biases existing in academia, such as "first impression" bias, the Dr. Fox effect, ideological and theoretical biases, as well as language and social identity bias [3]. Additional challenges include the selfish or competitive rejection of high-quality papers to the acceptance of low-quality manuscripts without careful validation [4]. Several initiatives are exploring the use of artificial intelligence (AI) to enhance the peer review process. However, concerns remain about the reliability of AI systems and their potential to reinforce existing biases [5].

Our study addresses these issues by bridging the gap between argument mining on one hand and computational argumentation and the Semantic Web on the other, with an emphasis on explainability and unbiasedness in the evaluation process. Recently, Baimuratov et al. [6] demonstrated that scientific peer review can be framed as an argumentative dispute between manuscript authors and reviewers, modeled using abstract argumentation frameworks [7] and resolved through OWL DL reasoning. However, while manually formalizing peer reviews is time-consuming and argument mining offers a way to streamline this process, to the best of our knowledge, no prior studies have explored the extraction of

SymGenAI4Sci 2025: First International Workshop on Symbolic and Generative AI for Science co-located with Semantics-2025, September 3–5, 2025, Vienna, Austria.
*Corresponding author.
ildar.baimuratov@l3s.de (I. Baimuratov); karpovehalex@gmail.com (A. Karpovich)
0000-0002-6573-131X (I. Baimuratov); 0009-0001-2740-2136 (A. Karpovich)

argumentation frameworks from scientific peer reviews with the goal of enabling computational dispute resolution. In this research, we build on existing methods for argument identification in peer reviews and extend them with argumentative relation mining, enabling the construction of comprehensive abstract argumentation frameworks from review texts. A key advantage of our method is its transparency and controllability. At each step, human oversight is possible, allowing manual correction of model outputs, while the final dispute resolution is produced deterministically based on formal semantics. We envision that this approach can assist editors and meta-reviewers in making more informed final decisions.

The paper is structured as follows: in section 2, we review related work, section 3 provides background on abstract argumentation frameworks and their representation in OWL DL, section 4 describes our method for extracting abstract argumentation frameworks from peer review texts. We evaluate our argument mining techniques in section 5 and conclude in section 6.

2. Related Work

In this section, we review applications of artificial intelligence in the peer review process, with a particular focus on argument mining from peer review texts and on argument representation.

2.1. Peer Review and Artificial Intelligence

Experimental initiatives aimed at transforming the peer review process are currently under development. A variety of review systems have been explored by Tennant et al. [8], ranging from voting mechanisms similar to those on Reddit to innovative blockchain-based models. One approach to addressing challenges in peer review is the automation of manuscript selection using AI. Price and Flach [9] demonstrated that AI and machine learning can effectively automate and enhance several stages of the review process, including the assignment of articles (or grant applications) to appropriate reviewers. Ghosal et al. [10] investigated the impact of reviewer sentiment embedded in review texts as a predictor of review outcomes. The PEERAssist system [11] leverages a cross-attention mechanism between the full article text and the review text to predict reviewer decisions. Mrowinski et al. [12] showed that evolutionary algorithms can significantly optimize editorial strategies, accelerating the review process and reducing the burden on editors. Other notable examples include Statcheck [13], software that verifies the consistency of authors' statistics with a focus on p-values; Penelope.ai[1], a commercial platform that ensures citations and manuscript structures meet journal guidelines; and StatReviewer[2], which checks the validity of statistics and methods in manuscripts. The ethical challenges posed by such approaches are often related to the risk of reproducing biases within AI systems. Checco et al. [5] discussed the potential and limitations of employing AI to support human decision-making in the quality assurance and peer review of scientific research. Their findings suggest correlations between the decision-making process and other proxy measures of quality, raising concerns that AI could unintentionally reinforce existing biases in the peer review process.

2.2. Argument Identification in Peer Review

Scientific peer review can be viewed as an argumentative dispute between manuscript authors and reviewers, enabling the use of argumentation solvers to resolve such disputes [6]. However, manually formalizing peer reviews is time-consuming, and argument mining offers a way to streamline this process. Argument mining involves several tasks, including opinion mining, controversy detection, argumentative zoning, argument/non-argument classification, and the automatic identification of relations between arguments [14]. It has been applied in fields such as law, medical informatics, robotics, the Semantic Web, and security [15]. Despite its wide applicability, argument mining in scientific peer review remains relatively underexplored.

[1] https://www.penelope.ai/

[2] http://blogs.biomedcentral.com/bmcblog/2016/05/23/peerless-review-automating-methodological-statistical-review

Argument identification is a subtask in argument mining, focused on detecting and extracting argumentative components from natural language text, such as claims, premises, rebuttals, etc. Among argument identification from peer review, Hua et al. [16] collected reviews from major machine learning and natural language processing venues and annotated them with five types of argumentative propositions: 1) evaluation, 2) request, 3) fact, 4) reference, and 5) quote. Fromm et al. [17] retrieved peer reviews from computer science conferences via the OpenReview platform and annotated them using an argumentation scheme from [18], which categorizes text into 1) non-arguments, 2) supporting arguments, and 3) attacking arguments. Baimuratov et al. [6] annotated a corpus of peer reviews from various domains with abstract argumentation frameworks [7], identifying both argumentative and non-argumentative components. A comparison of argument identification approaches is presented in Table 1. Since the corpus of Baimuratov et al. [6] is explicitly annotated with abstract argumentation frameworks, we adopt and further extend their approach.

Table 1
Summary of methods for argument identification in peer review

Study	Domain	N. of classes	Krippendorf's α	Model	F1
[16]	ML, NLP	5	0.61	BiLSTM	0.626
[17]	CS	3	0.568	BERT	0.789
[6]	Various	2	0.81	LSTM	0.631

2.3. Argumentative Relation Extraction

Argumentative relation extraction is the task of automatically identifying and classifying the logical or rhetorical relationships, such as support or attack, between argumentative components within natural language text. Although, to the best of our knowledge, no studies have specifically focused on extracting attack relations from scientific peer reviews, we review approaches from other domains to adopt best practices. As such, Ruiz-Dolz et al. [19] evaluated transformer-based models (BERT, XLNET, RoBERTa, DistilBERT and ALBERT) for extracting argument relations on the US2016 debate [20] and Moral Maze corpora[3]. Chakrabarty et al. [21] proposed an argument mining model for online persuasive discussion forums [22] based on Rhetorical Structure Theory with a modified BERT model. Mayer et al. [23] experimented with various neural architectures (LSTM, GRU and CRF) for argument mining from randomized controlled trials. Bao et al. [24] conducted experiments on two datasets: Persuasive Essays (PE) [25] and Consumer Debt Collection Practices [26], with a combination of BERT and LSTM models. Paul et al. [27] proposed an unsupervised graph-based ranking method integrated into a biLSTM-based model, evaluated on the PE and Debatepedia datasets. Jo et al. [28] classified argumentative relations based on four logical mechanisms: 1) factual consistency, 2) sentiment coherence, 3) causal relation and 4) normative relation. They annotated datasets from the contentious topic platform Kialo[4] and Debatepedia and developed a BERT-based model. Sun et al. [29] proposed a dual prior graph neural network (DPGNN) that jointly incorporates knowledge from pretrained language models (BERT) and syntactical information, with experiments on Debatepedia and PE datasets. Liu et al. [30] framed argument mining as a multi-hop reading comprehension task, leveraging BART-based models to learn argument structures as a "chain of thought". Gorur et al. [31] investigated the capabilities of Large Language Models (LLMs) with prompting for identifying argumentative relations. They experimented with two open-source LLMs (LLaMA-2 and Mistral) across ten datasets. Similarly, Cabessa et al. [32] framed argument mining, including argumentative relation extraction, as a text generation task using fine-tuned LLMs, achieving their best results with LLaMA-3-8B (4-bit). A summary of these approaches is provided in Table 2. While LLMs perform strongly, fine-tuned BERT- and LSTM-based models can still surpass them while requiring significantly fewer resources. Therefore, we adopt them in our study.

[3] https://siwells.github.io/dataset_moral.maze/
[4] https://www.kialo.com/

Table 2
Summary of argumentative relation extraction methods

Study	Models	Data	F1-score
[19]	BERT, XLNET, RoBERTa, DistilBERT, ALBERT	US2016 debate	70
		Moral Maze	61
[21]	BERT, RST	Online persuasive discussion forums	28.3
[23]	BERT*, GRU, CRF	Randomized controlled trials	67.8
[24]	BERT, LSTM	PE	82.5
		Consumer debt collection practices	67.8
[27]	biLSTM+, ELMO*	PE	59.43
		Debatepedia	63.69
[28]	LogBERT	Kialo	80.2
		Debatepedia	80.7
[29]	BERT, DPGNN	PE	63.8
		Debatepedia	84.1
[30]	BART	PE	82.7
[31]	LLaMa-2-70B-4bit	Average of 10 datasets	82
[32]	LLaMA-3-8B-4bit	PE	83.5

2.4. Argument Representation

To enable computational dispute resolution, argument representations must conform to specific argumentation frameworks. One of the foundational frameworks in argumentation theory are Dung's abstract argumentation frameworks (AAF) [7]. Computational models designed to solve such frameworks are evaluated through the International Competition on Computational Models of Argumentation[5] (ICCMA). While abstract argumentation solvers are an indispensable component of theories of argumentation, they do not address the nature of individual arguments or guide the modeling of real-world argumentation problems. For example, consider the argument representation used in ICCMA, as illustrated in 1. This format lacks mechanisms for representing the content or provenance of the arguments, limiting its ability to capture the full context of argumentative interactions.

Example 1. *An AAF with a set of arguments* $A = \{a, b, c, d, e\}$ *and a set of attacks* $R = \{(a, b), (b, d), (d, e), (e, d), (e, e)\}$*, assuming the indexing* $a = 1$*,* $b = 2$*,* $c = 3$*,* $d = 4$*,* $e = 5$*, in ICCMA is represented as follows:*

```
p af 5
1 2
2 4
4 5
5 4
5 5
```

Alternatively, ASPIC+ [33] is a structured argumentation framework that enables modeling conflicts between arguments and assumes three ways of attacking: 1) by challenging their uncertain premises, 2) by attacking their defeasible inferences, and 3) by disputing the conclusions drawn from defeasible inferences. Another notable format is Argdown[6], an argument markup language inspired by Markdown, implemented using a context-free grammar and parser. However, ASPIC+ and Argdown do not offer implementations based on Semantic Web standards, such as RDF and OWL, which would significantly enhance the interoperability and machine interpretability of argument representations.

In contrast, the Argument Interchange Format (AIF) [34], which is based on the concept of argumentation schemes from Walton et al. [35], is designed to facilitate data exchange between different argumentation tools and applications. Various implementations of AIF exist, for example, Rahwan and Simari [36] proposed an AIF implementation using OWL. The online database AIFdb [37] was

[5] https://argumentationcompetition.org/index.html
[6] https://argdown.org/

created to store annotated argumentative texts. The AIF format is also used in the OVA tool [38] for analyzing and annotating natural language argumentation. However, OWL implementation of AIF lacks tools for computational dispute resolution. This highlights a notable gap between advanced argument representation methods and argumentation solvers. Only a few studies have sought to bridge this gap. For example, Moguillansky and Simari [39] explored the encoding of abstract argumentation within ALC description logic to enable reasoning over inconsistent ontologies. Recently, Baimuratov et al. [40] proposed a promising method for representing AAF in OWL DL, enabling dispute resolution via OWL reasoning while preserving the advantages of OWL-based knowledge modeling. The present research builds on this approach to model disputes within the context of peer review.

3. Background

In this section, we provide the necessary background to define our method, including the concept of abstract argumentation frameworks and their representation in the OWL DL language, which enables computational dispute resolution.

3.1. Abstract Argumentation Frameworks

Instead of evaluating individual arguments based on their internal structure, as explored in works such as [35] and [41], abstract argumentation frameworks [7] abstract away from the internal structure of arguments and the formal logical reasoning used to validate conclusions from premises, focusing exclusively on the relationships between arguments. In abstract argumentation frameworks, an unstructured argument serves as the atomic unit in an argumentative dispute. The framework is represented as a graph, in which arguments are connected by a binary, asymmetric attack relation that represents criticism or counterargumentation.

Definition 1. *An **argumentation framework** AF is a pair*

$$AF =< A, R >,$$

where A is a set of arguments and $R \subseteq A \times A$ is the attack relation.

Thus, any argumentation framework can be represented as a directed graph.

We say that an argument $\alpha \in A$ attacks an argument $\beta \in A$, or that β is attacked by an argument α if $(\alpha, \beta) \in R$. Additionally, we say that a set of arguments $S \subseteq A$ attacks α, or that α is attacked by S if some argument $\beta \in S$ attacks α:

$$attacks(S, \alpha) \equiv_{def} \exists \beta \in S \; (\beta, \alpha) \in R.$$

The minimal criterion for persuading a rational agent is the notion of an acceptable argument.

Definition 2. *An argument $\alpha \in A$ is **acceptable** on a set of arguments S only if, whenever it is attacked by an argument β, S attacks β.*

$$acceptable(\alpha, S) \equiv_{def} \forall \beta \in A \; (\beta, \alpha) \in R \implies attacks(S, \beta).$$

In order to define outcomes to disputes, we first utilize the notion of a conflict-free set of arguments.

Definition 3. *A set of arguments S is called **conflict-free** if there are no arguments α and β in S such that $(\alpha, \beta) \in R$.*

$$cf(S) \equiv_{def} \forall \beta \in S \; \neg attacks(S, \beta).$$

Now we can define an admissible set of arguments.

Definition 4. *A conflict-free set of arguments S is **admissible** only if every argument in S is acceptable with respect to S.*

$$adm(S) \equiv_{def} cf(S) \wedge \forall \alpha \in S \; acc(\alpha, S).$$

In [42], it was shown that argumentation frameworks in peer review are well-founded.

Definition 5. *An argumentation framework is **well-founded** only if there exists no infinite sequence $\alpha_0, \alpha_1, ..., \alpha_n, ...$ such that $\forall i$, $(\alpha_{i+1}, \alpha_i) \in R$.*

To formulate dispute resolutions, various acceptance semantics are introduced. These semantics allow for the computation of sets of arguments, known as extensions, including preferred, stable, complete and grounded extensions. However, it is known from [7], that if an argumentation framework is well-founded, it has exactly one complete extension, which coincides with the grounded, preferred, and stable extensions.

Theorem 1. *Every well-founded argumentation framework has exactly one complete extension which is grounded, preferred and stable.*

Thus, we consider here only the notion of complete extension.

Definition 6. *A set S is a **complete extension** only if it is admissible and every acceptable argument with respect to S belongs to it.*

$$complete(S) \equiv_{def} adm(S) \wedge (\forall \alpha \in A \; acc(\alpha, S) \implies (\alpha \in S)).$$

Thus, to resolve a dispute in peer review, it is required to identify all acceptable arguments and the unique complete extension.

3.2. Representation of Abstract Argumentation Frameworks in OWL DL

A general approach for representing abstract argumentation frameworks in the OWL DL language, designed to automatically classify arguments into admissible sets using reasoning, was presented in [40]. In [6], the authors demonstrated how this general approach is applied to peer review.

In this approach, each review party is modeled as an `owl:Class`. Each argument of the review parties is represented as `owl:NamedIndividual`, with the argument text captured using a custom `owl:AnnotationProperty` named *text*. The association of each argument with its respective review party is indicated by the `rdf:type` relation. Attack relations between arguments are asserted using the `owl:ObjectProperty` *attacks* and its inverse *isAttackedBy*. Additionally, specific properties introduced for the peer review scenario, *round* and *number*, are also represented as `owl:AnnotationProperty`. To ensure logical inference under the open world assumption, each individual is "closed" with respect to the *attacks* and *isAttackedBy* relations. Listing 1 provides an example of a peer review argument represented in OWL. Listing 2 presents the declarations of conflict-free and admissible argument subsets for the *Author* party.

By representing abstract argumentation frameworks in OWL, reasoners such as Pellet [43] can be used to classify each party's arguments as acceptable. Figure 1 shows a visualization of a peer review argumentation framework represented in OWL using the OntoGraf[7] tool after the argument classification.

[7] https://protegewiki.stanford.edu/wiki/OntoGraf

```
Individual: <onto#reviewer_11>

  Annotations:
    <onto#number> "1"^^xsd:string,
    <onto#round> "1"^^xsd:string,
    <onto#text> "However, being experts in their field the authors might not be aware that for
         readers less familiar with the metabolism/physiology of archaea, the examples are not
         always easy to follow..."^^xsd:string

  Types:
    <onto#Reviewer_1>
    <onto#attacks> only({onto#author_3>}),
    < onto#isAttackedBy> only({<onto#author_1>})

  Facts:
    <onto#attacks><onto#author_3>
    <onto#isAttackedBy> <onto#author_1>
```

Listing 1: Representation of a peer review argument in OWL

```
Class: <onto#AuthorConflictFree>

  EquivalentTo:
    <onto#Author>
    and (<onto#attacks> only (<onto#Reviewer_1> or <onto#Reviewer_2>))

Class: <onto#AuthorAdmissible>

  EquivalentTo:
    <onto#AuthorConflictFree>
    and (<onto#isAttackedBy> only(<onto#isAttackedBy> some <onto#AuthorConflictFree >) )
```

Listing 2: Conflict-free and admissible subsets of authors' arguments in OWL

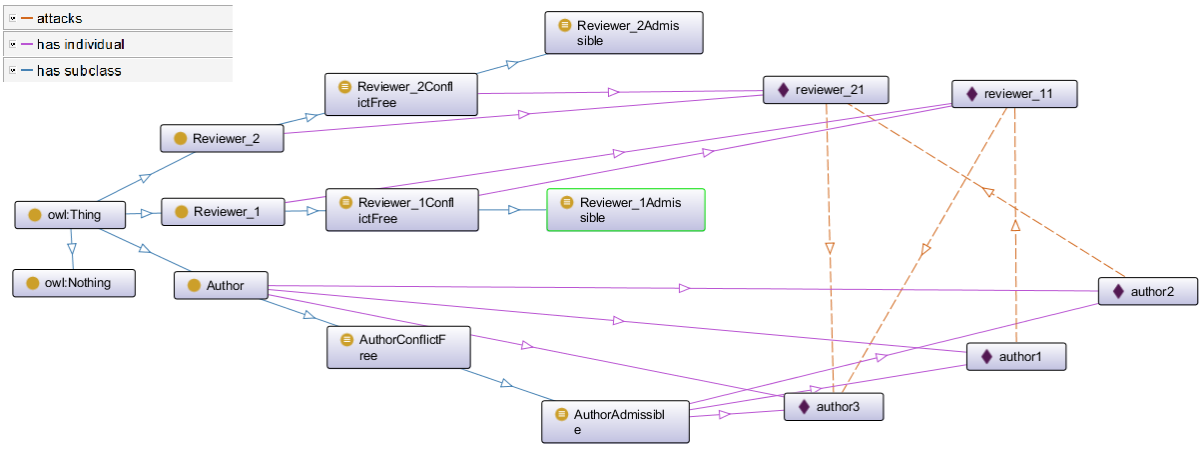

Figure 1: Visualization of a peer review abstract argumentation framework in OWL with classified arguments

4. Method

Modeling argumentative disputes in peer review using argumentation frameworks enables their resolution with argumentation solvers. However, manually constructing these abstract frameworks from peer review texts is labor-intensive. In this section, we present a method for automatically extracting complete abstract argumentation frameworks from peer reviews.

Our approach consists of four main steps: 1) argument identification, 2) relation extraction, 3) framework construction from the identified arguments and attack relations, and 4) resolving the constructed argumentation frameworks with OWL-reasoning, as illustrated on Figure 2. A key advantage of our method is its transparency and controllability. At each step, human oversight is possible, allowing manual correction of model outputs, while the final dispute resolution is produced deterministically based on formal semantics.

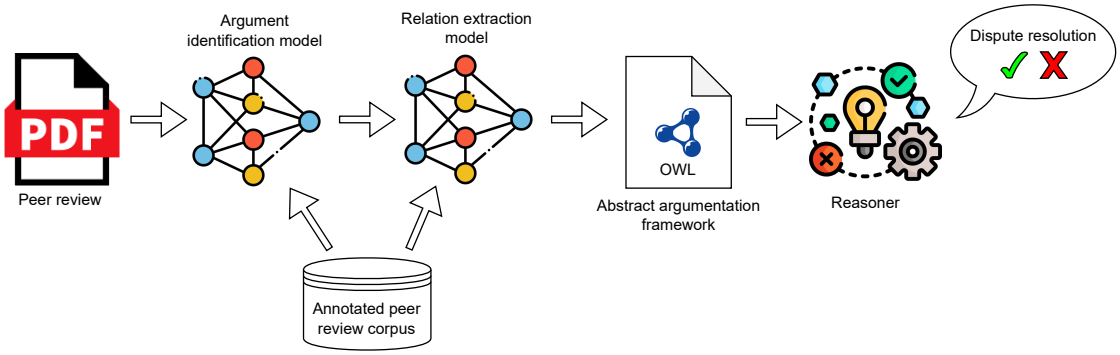

Figure 2: Our approach to dispute resolution in peer review

4.1. Data

To the best of our knowledge, only one corpus of scientific peer reviews explicitly includes annotated abstract argumentation frameworks [6]. The authors annotated the open peer review corpus from MDPI [44], which comprises 123 peer-reviewed articles published in various MDPI journals accessible as of June 16, 2022.

In this annotation, each row corresponds to a single argument and the columns represent various characteristics of the arguments in peer reviews:

- *Text:* text of the argument.
- *Side:* party of the peer review to whom the argument belongs (authors or one of the reviewers).
- *Opponent:* the *Side* that owns the argument being attacked by the current one.
- *Round:* review phase, starts at 1 and increases by 1 with each attack between the same *Side* and *Opponent* pair.
- *Number:* the unique number of the current argument within the same *Side* and *Round*, starting from 1.
- *Attacks: Number* of the argument attacked by the current one, 0 - if the author's whole article is criticized.

Not all reviews in the original corpus were sufficient to construct complete abstract argumentation frameworks. As a result, the annotated corpus includes 88 peer reviews comprising 37,285 sentences. Inter-annotator agreement, measured using Krippendorff's α, is 0.81.

4.2. Argument identification

We frame the argument identification task as sentence classification, where sentences are classified into two categories: arguments and non-arguments. To achieve this, we segment a peer review text into sentences and apply a text-to-annotation matching algorithm. The resulting class distribution is slightly imbalanced, with 44% argumentative sentences and 56% non-argumentative sentences. The segmentation was implemented with the NLTK library. An example of the matching result is shown in Figure 3.

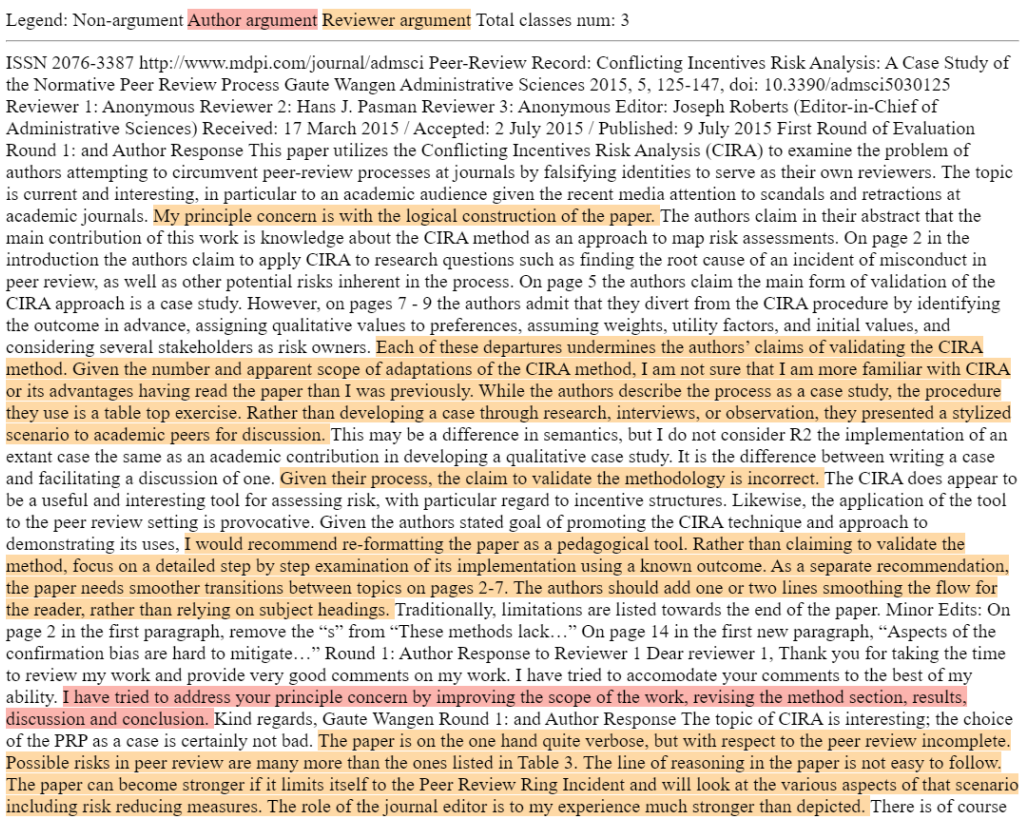

Figure 3: Example of matching peer review text with the annotated arguments

Based on the literature review, we selected a model consisting of BERT embeddings and an LSTM network for this classification task. The model utilizes `distilbert-base-uncased` embeddings [45] as input and includes an LSTM layer with a Sigmoid activation function, followed by a fully connected layer with a SoftMax activation function. The model outputs the probabilities of a sentence belonging to each class, with the final classification determined by selecting the class with the highest probability. To prevent overfitting, the model includes a dropout layer. Additionally, a simple model consisting of two fully connected layers with ReLU activation was used as a baseline for comparison.

4.3. Relation extraction

The extraction of attack relations between arguments is framed as a binary classification problem, classifying argument pairs as either having a relation or not. To achieve this, the dataset presented in [6] was transformed to explicitly represent attack relations between arguments. The resulting dataset consists of three columns: the text of the first argument, the text of the second argument, and a binary label indicating the presence of an attack relation between them. Moreover, the original dataset only contain argument pairs with established relations. However, machine learning models require both positive and negative examples for training. To generate negative examples, we randomly selected argument pairs from the same review that do not have an annotated attack relation.

We employed an LSTM architecture with BERT-based embeddings for relation extraction, exploring two approaches to text embedding and two strategies for model output. For text embeddings, we aimed to investigate whether general-purpose embeddings would outperform domain-specific pretrained embeddings. To this end, we explored two options: distilled BERT and SciBERT [46] — a domain-specific version of BERT, tailored for handling scientific text. Regarding model output, the first approach utilized a Sigmoid activation function to generate a one-dimensional output. A 0.5 threshold was applied to classify the output: values above the threshold indicated the presence of an attack relation (class 1), while values below it indicated its absence (class 0). In the second approach, we used a Softmax

activation function to produce a two-dimensional output, where each value represented the probability of the argument pair belonging to either class 0 or class 1. The final classification was determined by selecting the class with the highest probability. The resulting approaches are denoted as follows: 1) BERT+LSTM1, 2) SciBERT+LSTM1 and 3) BERT+LSTM2.

5. Results

In this section, we provide an empirical evaluation of our argument identification and relation extraction models, as well as an overall evaluation of the resulting argumentation frameworks.

5.1. Argument Identification

We trained the both LSTM and baseline argument identification models on the preprocessed corpus. The dataset was partitioned into training, validation, and test samples with proportions of 70%, 10%, and 20%, respectively. To address the class imbalance, stratification by argument type was employed during the data split. For both the baseline model and the LSTM model, we used the Adam optimizer and the cross-entropy loss function, other hyperparameter values are provided in Table 3.

Table 3
Hyperparameters of the argument identification models

Hyperparameter	Baseline	LSTM
batch_size	32	64
embedding_dim	350	350
max_length	350	110
vocab_size	30522	30522
optimizer/lr	0.001	0.0006
hidden_dim	350	18
hidden_dim2	256	12
dropout	0	0.75

The performance metrics for both the baseline and LSTM models are listed in Table 4. The LSTM model correctly classifies sentences in approximately 68% of cases, which is 12% higher than the baseline accuracy and comparable to models trained on other datasets.

Table 4
Performance of the argument identification models

	Precision	Recall	Accuracy	F1
Baseline	57.69	51.11	56.06	54.20
LSTM	**63.44**	**62.67**	**68.04**	**63.05**

5.2. Relation Extraction

For relation extraction, the preprocessed dataset was split into training, test, and validation sets in an 80%/10%/10% ratio. Hyperparameters for the models were selected experimentally. For all models, we used a batch size of 128, a maximum input sequence length of 110, the Adam optimizer, and a cross-entropy loss function. Other hyperparameters are listed in Table 5.

As a result, the BERT+LSTM1 model, which uses fine-tuned general-purpose BERT embeddings and one-dimensional output, achieved the highest performance with an F1 score of 86.2, see Table 6. Its two-dimensional-output variant reached an F1 score of 80.43. BERT+LSTM1 also slightly outperformed the SciBERT+LSTM1 model, which achieved an F1 score of 85.27.

Table 5
Hyperparameters of the relation extraction models

Hyperparameter	BERT+LSTM1	BERT+LSTM2	SciBERT+LSTM1
embedding_dim	-	-	350
vocab_size	-	-	30522
optimizer/lr	5e-4	5e-4	0.001
hidden_dim	400, 50	400, 50	800
dropout_fc	0.3	0.3	0.7

Table 6
Performance of the relation extraction models

Model	Accuracy	F1
BERT+LSTM1	**93.12**	**86.2**
BERT+LSTM2	87.3	80.43
SciBERT+LSTM1	90.74	85.27

5.3. Framework Construction

The combination of argument identification and the extraction of attack relations enables the construction of abstract argumentation frameworks and their resolution using OWL reasoning. Each extracted framework was converted into JSON and then transformed into an OWL DL representation, following the approach of [40]. These OWL representations were processed using the Pellet reasoner to classify the arguments. As a result, all generated OWL representations were successfully processed and classified into admissible sets. The implementation and results are available on GitHub[8].

However, the accuracy of the final decisions, i.e. whether the reviewed paper should be accepted or not, compared to decisions based on the original annotated frameworks, was only 42%. This indicates that, despite acceptable performance on argument identification and relation extraction individually, error propagation occurs when both steps are combined. To address this, we recommend a human-in-the-loop approach by introducing an intermediate validation step for the identified arguments before extracting relations. We assume that if all arguments are correctly identified, the overall accuracy of the resulting frameworks will approach that of the relation extraction component, which achieves 93.12%.

6. Conclusion

In this research, we facilitated computational dispute resolution in scientific peer review by extracting complete abstract argumentation frameworks from peer review text. Specifically, we addressed the tasks of argument identification and the extraction of attack relations between arguments. Once extracted from peer review texts, these frameworks enable automated dispute resolution through argumentation solvers. A key advantage of our method is its transparency and controllability. At each step, human oversight is possible, allowing manual correction of model outputs, while the final dispute resolution is produced deterministically based on formal semantics. In real-world peer review scenarios, our method can support meta-reviewers and editors in making final decisions on manuscript acceptance.

To evaluate our approach, we tested several models, achieving a maximum F1 score of 63.05 for argument identification and 86.2 for relation extraction. Since no previous studies have evaluated argument and argumentative relation identification tasks on the specific dataset we used, our pipeline cannot be directly compared against other approaches. However, for both tasks we achieved performance comparable to results reported on other datasets. All extracted argumentation frameworks were represented in OWL and successfully resolved. Nevertheless, the overall accuracy compared to the

[8] https://github.com/Karpovich-alex/mdpi_argumentations/tree/add-relation-files

original annotations was only 42%, indicating that an intermediate validation step is needed to verify the identified arguments and prevent error propagation.

Limitations

The main limitation of our approach stems from the relatively small dataset, which raises concerns about the generalizability of our results. Improvements can be pursued in two directions: 1) expanding the training data by annotating peer reviews from additional sources, such as OpenReview, and 2) leveraging more advanced models, particularly LLMs. Additional limitations arise from the formal framework we employ. The binary attack setup omits other peer review discourse relations, such as support, partial agreement and comments. The argument representation would also benefit from incorporating the internal logical content of arguments and from weighting the attack relations between them. In this work, however, we use only basic abstract argumentation frameworks, where arguments are treated as atomic entities and dispute solutions are derived solely from attack relations. Nevertheless, our approach can be extended without loss to integrate more advanced argumentation frameworks. We plan to address these limitations in future research, particularly by analyzing the internal logical structure of identified arguments and interpreting the probabilities from relation extraction models as attack weights.

Acknowledgments

We would like to acknowledge the funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2163/1 - Sustainable and Energy Efficient Aviation – Project-ID 390881007 and the German Ministry of Education and Research (BmBF) for the project KISSKI AI Service Center (01IS22093C).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Fire, C. Guestrin, Over-optimization of academic publishing metrics: observing goodhart's law in action, GigaScience 8 (2019) giz053. URL: https://doi.org/10.1093/gigascience/giz053. doi:10.1093/gigascience/giz053.

[2] J. Huisman, J. Smits, Duration and quality of the peer review process: the author's perspective, Scientometrics 113 (2017) 633–650. URL: https://ideas.repec.org/a/spr/scient/v113y2017i1d10.1007_s11192-017-2310-5.html. doi:10.1007/s11192-017-2310-5.

[3] C. J. Lee, C. R. Sugimoto, G. Zhang, B. Cronin, Bias in peer review, Journal of the American Society for Information Science and Technology 64 (2013) 2–17. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22784. doi:https://doi.org/10.1002/asi.22784. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22784.

[4] R. D'Andrea, J. P. O'Dwyer, Can editors save peer review from peer reviewers?, PloS one 12 (2017) e0186111.

[5] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, G. Bianchi, Ai-assisted peer review, Humanities and Social Sciences Communications 8 (2021). doi:10.1057/s41599-020-00703-8.

[6] I. Baimuratov, A. Karpovich, E. Lisanyuk, D. Prokudin, Argument identification for neuro-symbolic dispute resolution in scientific peer review, in: Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries, 2024, pp. 1–9.
[7] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial Intelligence 77 (1995) 321–357. doi:10.1016/0004-3702(94)00041-X.
[8] J. P. Tennant, J. M. Dugan, D. Graziotin, D. C. Jacques, F. Waldner, D. Mietchen, Y. Elkhatib, L. B. Collister, C. K. Pikas, T. Crick, et al., A multi-disciplinary perspective on emergent and future innovations in peer review, F1000Research 6 (2017).
[9] S. Price, P. A. Flach, Computational support for academic peer review: A perspective from artificial intelligence, Commun. ACM 60 (2017) 70–79. URL: https://doi.org/10.1145/2979672. doi:10.1145/2979672.
[10] T. Ghosal, R. Verma, A. Ekbal, P. Bhattacharyya, DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1120–1130. URL: https://aclanthology.org/P19-1106. doi:10.18653/v1/P19-1106.
[11] P. K. Bharti, S. Ranjan, T. Ghosal, M. Agrawal, A. Ekbal, Peerassist: Leveraging on paper-review interactions to predict peer review decisions, in: H.-R. Ke, C. S. Lee, K. Sugiyama (Eds.), Towards Open and Trustworthy Digital Societies, Springer International Publishing, Cham, 2021, pp. 421–435.
[12] M. Mrowinski, P. Fronczak, A. Fronczak, M. Ausloos, O. Nedić, Artificial intelligence in peer review: How can evolutionary computation support journal editors?, PLOS ONE 12 (2017) e0184711. doi:10.1371/journal.pone.0184711.
[13] M. B. Nuijten, J. R. Polanin, "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses, Research synthesis methods 11 (2020) 574–579.
[14] J. Lawrence, C. Reed, Argument mining: A survey, Computational Linguistics 45 (2020) 765–818.
[15] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021) e5.
[16] X. Hua, M. Nikolov, N. Badugu, L. Wang, Argument mining for understanding peer reviews, arXiv preprint arXiv:1903.10104 (2019).
[17] M. Fromm, E. Faerman, M. Berrendorf, S. Bhargava, R. Qi, Y. Zhang, L. Dennert, S. Selle, Y. Mao, T. Seidl, Argument mining driven analysis of peer-reviews, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 4758–4766.
[18] C. Stab, T. Miller, I. Gurevych, Cross-topic argument mining from heterogeneous sources using attention-based neural networks, arXiv preprint arXiv:1802.05758 (2018).
[19] R. Ruiz-Dolz, J. Alemany, S. M. H. Barberá, A. García-Fornes, Transformer-based models for automatic identification of argument relations: A cross-domain evaluation, IEEE Intelligent Systems 36 (2021) 62–70.
[20] J. Visser, B. Konat, R. Duthie, M. Koszowy, K. Budzynska, C. Reed, Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction, Language Resources and Evaluation 54 (2020) 123–154.
[21] T. Chakrabarty, C. Hidey, S. Muresan, K. McKeown, A. Hwang, Ampersand: Argument mining for persuasive online discussions, arXiv preprint arXiv:2004.14677 (2020).
[22] C. Hidey, E. Musi, A. Hwang, S. Muresan, K. McKeown, Analyzing the semantic types of claims and premises in an online persuasive forum, in: Proceedings of the 4th Workshop on Argument Mining, Columbia Univ., New York, NY (United States), 2017.
[23] T. Mayer, E. Cabrio, S. Villata, Transformer-based argument mining for healthcare applications, in: ECAI 2020, IOS Press, 2020, pp. 2108–2115.
[24] J. Bao, C. Fan, J. Wu, Y. Dang, J. Du, R. Xu, A neural transition-based model for argumentation mining, 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), 2021, pp. 6354–6364.
[25] C. Stab, I. Gurevych, Parsing argumentation structures in persuasive essays, Computational Linguistics 43 (2017) 619–659.
[26] J. Park, C. Cardie, A corpus of erulemaking user comments for measuring evaluability of arguments, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[27] D. Paul, J. Opitz, M. Becker, J. Kobbe, G. Hirst, A. Frank, Argumentative relation classification with background knowledge, in: Computational Models of Argument, IOS Press, 2020, pp. 319–330.
[28] Y. Jo, S. Bang, C. Reed, E. Hovy, Classifying argumentative relations using logical mechanisms and argumentation schemes, Transactions of the Association for Computational Linguistics 9 (2021) 721–739.
[29] Y. Sun, B. Liang, J. Bao, M. Yang, R. Xu, Probing structural knowledge from pre-trained language model for argumentation relation classification, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 3605–3615.
[30] B. Liu, V. Schlegel, R. T. Batista-Navarro, S. Ananiadou, Argument mining as a multi-hop generative machine reading comprehension task, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 10846–10858.
[31] D. Gorur, A. Rago, F. Toni, Can large language models perform relation-based argument mining?, arXiv preprint arXiv:2402.11243 (2024).
[32] J. Cabessa, H. Hernault, U. Mushtaq, Argument mining with fine-tuned large language models, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 6624–6635.
[33] S. Modgil, H. Prakken, The aspic+ framework for structured argumentation: a tutorial, Argument & Computation 5 (2014) 31–62.
[34] C. Chesnevar, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, S. Willmott, et al., Towards an argument interchange format, The knowledge engineering review 21 (2006) 293–316.
[35] D. Walton, C. Reed, F. Macagno, Argumentation schemes, Cambridge University Press, 2008.
[36] I. Rahwan, G. R. Simari, Argumentation in artificial intelligence, volume 47, Springer, 2009.
[37] J. Lawrence, F. Bex, C. Reed, M. Snaith, Aifdb: Infrastructure for the argument web, in: Computational Models of Argument, IOS Press, 2012, pp. 515–516.
[38] M. Reed, M. Janier, J. Lawrence, Ova+: An argument analysis interface, in: Computational Models of Argument: Proceedings of COMMA, volume 266, 2014, p. 463.
[39] M. O. Moguillansky, G. R. Simari, A generalized abstract argumentation framework for inconsistency-tolerant ontology reasoning, Expert Systems with Applications 64 (2016) 141–168.
[40] I. Baimuratov, E. Lisanyuk, D. Prokudin, Dispute resolution with owl dl and reasoning., in: Proceedings of the 36th International Workshop on Description Logics (DL 2023), 2023.
[41] H. Prakken, An abstract framework for argumentation with structured arguments, Argument & Computation 1 (2010) 93–124.
[42] I. Baimuratov, E. Lisanyuk, D. Prokudin, Dispute resolution in peer review with abstract argumentation and owl dl, arXiv preprint (????).
[43] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, Y. Katz, Pellet: A practical owl-dl reasoner, Journal of Web Semantics 5 (2007) 51–53.
[44] M. Miłkowski, K. Jasieński, MDPI Open Peer Review Corpus, 2022. URL: https://doi.org/10.18150/D5L2EK. doi:10.18150/D5L2EK.
[45] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
[46] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).