



Argument Identification for Neuro-Symbolic Dispute Resolution in Scientific Peer Review

Ildar Baimuratov

L3S Research Center, Leibniz University Hannover
TIB - Leibniz Information Centre for Science and
Technology
Hannover, Germany
ildar.baimuratov@l3s.de

Elena Lisanyuk

Institute of Philosophy, Russian Academy of Sciences
Moscow, Russia
elisanyuk@hse.ru

Alexandr Karpovich

ITMO University
St Petersburg, Russia
karpovehalex@gmail.com

Dmitry Prokudin

St Petersburg University
St Petersburg, Russia
d.prokudin@spbu.ru

Abstract

Peer review is a cornerstone of the academic editorial decision-making process, yet it faces significant challenges. Artificial intelligence can help address these challenges, but its use raises concerns about reliability and the potential for reproducing existing biases. In this research, we employ a formal argumentation-theoretic framework that allows for explicit analysis of arguments and their interrelations, combined with argument mining techniques to streamline the formalization of peer reviews, and resulting in a neuro-symbolic approach to dispute resolution. Our method involves identifying parties' arguments in peer reviews and representing them as abstract argumentation frameworks, which facilitate dispute resolution through logical inference. We annotate these frameworks within a corpus of scientific peer reviews, achieving a high Krippendorff's alpha of 0.81. Having the annotated corpus, we implement an argument mining pipeline that integrates BERT sentence embeddings with an LSTM model, classifying sentences into three categories: authors' arguments, reviewers' arguments, and non-arguments. We achieved an accuracy of 0.634 and an F1 score of 0.631, which are comparable to models trained on other datasets. However, our approach stands out by enabling the processing of the extracted argumentation with logical inference.

CCS Concepts

• **Information systems** → **Expert systems**; • **Computing methodologies** → **Information extraction**; **Discourse, dialogue and pragmatics**; *Language resources*; • **Applied computing** → **Publishing**; **Annotation**.

Keywords

Argumentation Mining, Neuro-Symbolic AI, Dispute Resolution, Scientific Peer Review, Abstract Argumentation Frameworks, Text Annotation

ACM Reference Format:

Ildar Baimuratov, Alexandr Karpovich, Elena Lisanyuk, and Dmitry Prokudin. 2024. Argument Identification for Neuro-Symbolic Dispute Resolution in Scientific Peer Review. In *The 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL '24)*, December 16–20, 2024, Hong Kong, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3677389.3702506>

1 Introduction

Peer review is a cornerstone of the academic editorial decision-making process across nearly all scientific disciplines. However, this process faces significant challenges due to the constant increase in submission volumes [12]. Additionally, the process is further complicated by the amplification of existing biases within the academic environment, manifesting in various forms [17]. Other notable issues include the selfish or competitive rejection of high-quality papers and the acceptance of low-quality manuscripts without thorough validation [5]. Several initiatives aim to leverage artificial intelligence (AI) to tackle the challenges in scientific peer review. However, the use of AI introduces concerns regarding its reliability and the potential for reproducing existing biases [3].

In this research, we rely on a formal argumentation-theoretic framework that enables explicit analysis of arguments and their interrelations, facilitating a more objective evaluation of the peer review process. Peer review can be viewed as an argumentative dispute, where authors attempt to persuade reviewers or editorial team to accept their manuscript. We address the challenge of resolving such disputes and take a step towards applying a neuro-symbolic approach that integrates formal frameworks for dispute resolution with argument mining techniques to extract these frameworks from peer review texts, see Fig 1. We envision that the proposed approach can assist editors and meta-reviewers in making the final decision.

The literature proposes various argumentation theories [26] and argumentation schemes of differing complexity [28]. Specifically, we consider peer review as a single mixed argumentative dispute between the authors of a manuscript submitted for publication in a scientific journal or conference and the reviewers who evaluate



This work is licensed under a Creative Commons Attribution International 4.0 License.
JCDL '24, December 16–20, 2024, Hong Kong, China
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1093-3/24/12
<https://doi.org/10.1145/3677389.3702506>

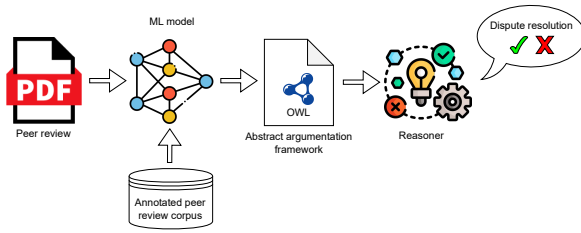


Figure 1: Our neuro-symbolic approach to dispute resolution in peer review

it. This dispute can be formalized using abstract argumentation frameworks [4] and represented in the OWL DL language to be resolved with reasoning engines [1]. However, manually formalizing peer reviews is a time-consuming and skill-intensive task. In this research, we explore how argument mining techniques can streamline the process of representing peer reviews as abstract argumentation frameworks.

Argument mining is a branch of natural language processing that focuses on automatically identifying arguments in natural language text and determining their structure [16]. In this work, we integrate argument mining with the formal approach to dispute resolution in peer review. Specifically, we aim to identify authors' and reviewers' arguments within scientific peer reviews and distinguish them from non-argumentative text. To achieve this, we developed an annotation scheme for peer review text and an algorithm to translate these annotations into abstract argumentation frameworks. We created a new dataset for argument mining by annotating the MDPI open peer review corpus [19] according to the proposed annotation scheme and investigated the performance of state-of-the-art argument mining techniques on this annotated corpus.

The paper is structured as follows: In the next section, we review existing studies on argument mining in peer reviews and scientific papers. Section 3 provides background on disputes and abstract argumentation frameworks. Section 4 describes our schema for annotating abstract argumentation frameworks in peer reviews, details the annotated corpus, and introduces a model for identifying arguments in peer review texts. We evaluate our annotation and argument mining techniques in Section 5 and conclude in Section 6.

2 Related work

Argument mining encompasses several tasks, such as opinion mining, controversy detection, argumentative zoning, argument/non-argument classification, and automatic identification of relations between arguments [16]. It has found applications in fields like law, medical informatics, robotics, the Semantic Web, and security [27]. Despite its broad applicability, argument mining for scientific peer-reviews remains relatively underexplored, which is why we also include works on argument mining in scientific papers.

Among them, Teufel et al. [25] analyzed the argumentative and rhetorical structure of a scientific paper using an argumentation

scheme with 15 categories, including aim, common ground, others' work, etc. They annotated papers from the fields of chemistry and computational linguistics, achieving the inter-annotator agreement with Fleiss' $\kappa = 0.71$ for chemistry and with $\kappa = 0.65$ for computational linguistics.

In [14], the authors introduced ArguminSci, a tool designed to analyze argumentation and rhetorical aspects of scientific writing. ArguminSci employs separate LSTM models pre-trained for five specific tasks in scientific publication mining: 1) argumentative component identification (background claim, own claim, and data); 2) discourse role classification (background, unspecified, challenge, future work, approach, outcome); 3) subjective aspect classification (none, limitation, advantage, disadvantage-advantage, disadvantage, common practice, novelty, advantage-disadvantage); 4) summary relevance classification (very relevant, relevant, may appear, should not appear, totally irrelevant); and 5) citation context identification (begin citation context, inside citation context, and outside). For training, they used an extended version of the Dr. Inventor Corpus [6], annotated with argumentation structures [15]. In the task of argumentative component identification, the model achieved a macro F1 score of 0.438 at the token level.

There is a work [9] that utilized a neuro-symbolic approach by applying Logic Tensor Networks to the task of argument mining from a corpus of scientific abstracts in the medical domain. The authors introduce two neural networks: NNCOMP for classifying argumentative component and NNLINK for predicting links between component pairs. They utilized the AbstRCT corpus [18], annotated with two classes, 1) EVIDENCE and 2) CLAIM. For the sub-symbolic part, they use a pre-trained GloVe model for sentence embeddings and a neural network composed of three stacked fully-connected layers with ReLU as activation function, followed by a softmax classification layer. For the symbolic part, two axioms are used: no symmetric link can exist, and claims can only be linked to other claims. Additionally, the authors trained an ensemble of 20 networks for both NNCOMP and NNLINK, evaluating the output of the networks by majority voting and averaging. They achieved an F1 score of 0.85 on a validation sample with Krippendorff's $\alpha = 0.81$.

Regarding peer reviews, Hua et al. [11] proposed an argument mining framework to automatically detect argumentative propositions and their types in peer reviews. They 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. The inter-annotator agreement for proposition segmentation was measured with Cohen's $\kappa = 0.93$, and the consensus level for type annotation was measured with Krippendorff's $\alpha = 0.61$. The authors then trained proposition segmentation and classification models on this annotated data. The best performance was achieved using a BiLSTM model enhanced with a conditional random field and word embeddings, yielding an overall F1 score of 0.626.

In [7], the authors retrieved peer reviews from computer science conferences via the OpenReview platform and annotated them using an argumentation scheme from [23], which categorizes text into 1) non-arguments, 2) supporting arguments and 3) attacking arguments. For the annotation, they achieved Krippendorff's $\alpha = 0.568$.

For the argument mining, they evaluated the following models: ArgBERT – zero-shot learning performance of a BERT model fine-tuned on another dataset annotated with the same scheme; PeerBERT-ArgInit – a model with the weights of ArgBERT and additionally fine-tuned on the collected dataset; PeerBERT – a smaller BERT model with 110M parameters fine-tuned on the collected dataset; and PeerBERT-L – a larger BERT model with 340M parameters fine-tuned on the collected dataset. The PeerBERT model achieved a macro F1 score of 0.789 at the sentence level, while PeerBERT-L achieved 0.9 at the token level.

As a separate category, there are work on computational argumentation, but they do not apply argument mining from text. The work [10] introduced an approach for mining argumentation schemes from biomedical research articles, implemented as logic programs in Prolog. Six argumentation schemes were considered: 1) agreement, 2) difference, 3) failed agreement, 4) eliminate difference, 5) analogy, and 6) consistent explanation. To mine individual arguments matching these schemes, the authors proposed initially applying NLP tools to extract entities and relations from the source text, followed by manual annotation of the arguments. Another study [8] proposed an approach to learning comparison criteria between arguments from defeasible logic programs.

Thus, although there are studies on neuro-symbolic argument mining, argument mining in peer reviews, and dispute resolution with structured argumentation, no work has addressed neuro-symbolic dispute resolution in peer reviews or specifically utilized abstract argumentation frameworks. Nonetheless, the best results in argument mining from peer review have been achieved using BERT embeddings and LSTM-based models, see Table 1. Therefore, we will adopt this approach in our study.

3 Background

Due to recent advances in machine learning and natural language processing, particularly with large language models, there is potential for these technologies to aid in resolving disputes in scientific peer review. However, concerns remain regarding the explainability, reliability, and fairness of these techniques. In this research, we rely on a formal argumentation-theoretic framework that enables explicit analysis of arguments and their interrelations, facilitating a more objective evaluation of the peer review process.

3.1 Peer review as argumentative dispute

In argumentation theory, disputes are classified based on the number of statements constituting the propositional content of the parties' viewpoints, distinguishing them as either single or multiple. Additionally, disputes are categorized into unmixed or mixed based on the parties' roles and intentions. In unmixed disputes, parties either present their own viewpoints or critique others' viewpoints, while in mixed disputes, parties both defend their own views and challenge those of others [26].

We conceptualize peer review as a single mixed argumentative dispute between the authors of a manuscript submitted for publication in a scientific journal or conference and the reviewers who evaluate it based on specific criteria. In a single mixed dispute, one party argues in favor of the view that P (is true), while the opposing party disagrees and critiques both P and the supporting arguments,

but does not propose or defend an alternative viewpoint regarding $\neg P$ (is true), as would be the case in a mixed multiple dispute. Ultimately, the resolution of a single mixed dispute will determine whether P is accepted or rejected, without necessitating a defense of $\neg P$.

Similarly, when authors submit their manuscript, they initiate an argumentative dispute by asserting that the manuscript deserves acceptance, i.e., they present the opinion P . Reviewers then evaluate the manuscript and can either agree with P , leading to the manuscript's acceptance as is, or reject it. In these cases, the dialogue between authors and reviewers concludes after the first round. Alternatively, reviewers may object to the manuscript, highlighting the need for revisions. We interpret these objections, comments, and recommendations as argumentative reasoning that challenges P and criticizes the authors' arguments. When reviewers raise such issues, the dialogue continues for at least one additional round. During this round, authors respond with new arguments and report the corrections made to the manuscript, after which reviewers decide whether to accept the revised manuscript.

3.2 Abstract argumentation frameworks

Interpreting scientific peer review as a single mixed dispute enables us to formalize it using abstract argumentation frameworks [4]. In these frameworks, arguments are represented as nodes in a graph, and binary asymmetric attack relations denote criticisms or counterarguments as the primary connections between these nodes.

Definition 3.1. An **argumentation framework** AF is a pair

$$AF = \langle AR, attacks \rangle,$$

where AR is a set of arguments and $attacks \subseteq AR \times AR$.

We say that an argument α attacks an argument β , or that β is attacked by an argument α if the relation $attacks(\alpha, \beta)$ is satisfied. Similarly, we say that a set of arguments S attacks α , or that α is attacked by S if some argument of S attacks α .

In order to define outcomes and then solutions to disputes, we first introduce the notion of a conflict-free set of arguments.

Definition 3.2. A set of arguments S is called **conflict-free** if there are no arguments α and β such that $attacks(\alpha, \beta)$ in S .

However, a conflict-free set of arguments alone is not enough to resolve a single mixed dispute. To determine a solution, we need a more robust criterion: an admissible subset. This subset consists of acceptable arguments that meet the minimum standard of reasonableness required to persuade a rational agent. An argument is called acceptable on some set of arguments if, when it is attacked, there is an argument in that set that attacks the argument that attacked it.

Definition 3.3. An argument $\alpha \in AR$ is **acceptable** with respect to a set of arguments S if and only if, for each argument $\beta \in AR$, if $attacks(\beta, \alpha)$ then $attacks(S, \beta)$.

Finally, let us define an admissible subset of arguments.

Definition 3.4. A conflict-free set of arguments S is **admissible** if and only if every argument in S is acceptable with respect to S .

Table 1: Comparison of related work

	Type	Field	N. of classes	Annotation	Model	F1
[25]	Papers	Chemistry, computational linguistics	15	Fleiss’ $\kappa = 0.71$ for chemistry, $\kappa = 0.65$ for computational linguistics	—	—
[14]	Papers	Computer graphics	24	—	LSTM	0.438
[9]	Papers	Medicine	2	Krippendorff’s $\alpha = 0.81$	LTN	0.85
[11]	Peer reviews	ML, NLP	5	Cohen’s $\kappa = 0.93$, Krippendorff’s $\alpha = 0.61$	BiLSTM	0.626
[7]	Peer reviews	CS	3	Krippendorff’s $\alpha = 0.568$	BERT	0.789 for sentences, 0.9 for tokens

In order to define a solution to a dispute when there are multiple subsets of admissible arguments, the notion of preferred extension is introduced.

Definition 3.5. A **preferred extension** E of the argumentation framework AF is the maximal (with respect to the set-theoretic inclusion operation) admissible set in AF .

Therefore, by employing abstract argumentation frameworks, we avoid evaluating individual arguments based on their internal structure, as discussed in [29] and [20]. Instead, our focus is on determining the preferred extension of the argument set to discern the outcome of the scientific peer review. This involves identifying whether the preferred extension belongs to the authors’ party (indicating the manuscript should be accepted), the reviewers’ party (suggesting the authors’ opinion is not justified), or a combination of both.

3.3 Representation of abstract argumentation frameworks in OWL DL

In [1], an implementation of Dung’s frameworks in OWL DL was proposed, designed to automatically classify arguments into admissible sets using reasoning. We illustrate this implementation with listings in the Manchester syntax¹.

Each argument set is represented as an `owl:Class` and each argument is considered to be an `owl:NamedIndividual`. The membership of an argument in an argument set is represented with the `rdf:type` relation. To represent the attack relation, an object property `attacks` is introduced. Additionally, the property `isAttackedBy` is defined as `owl:inverseOf attacks`, which is necessary for defining admissible sets of arguments. To provide reasoning under Open World Assumption (OWA), each individual argument is “closed” with respect to the list of arguments it attacks, constructed with `owl:oneOf` operator under `owl:allValuesFrom` restriction. If an argument attacks no argument, it attacks only `owl:Nothing`. For the same reason, each individual argument is closed regarding the `isAttackedBy` relation.

To define a conflict-free set in a manner that supports reasoning under the OWA, the `owl:complementOf` operator is used. For each argument set, a `owl:unionOf` all other argument sets is formed. Then, its conflict-free subset is defined as a class that has the `attacks` relation only to the union of the other argument sets. Listing 1

¹<https://www.w3.org/TR/owl2-manchester-syntax/>

Listing 1: Definition of a conflict-free set of arguments

```
Class: <onto.owl#AConflictFree>

EquivalentTo:
  <onto.owl#A>
  and (<onto.owl#attacks> only (<onto.owl#I>))

SubClassOf:
  <onto.owl#A>
```

Listing 2: Definition of an admissible set of arguments

```
Class: <onto.owl#AAdmissible>

EquivalentTo:
  <onto.owl#AConflictFree>
  and (<onto.owl#isAttackedBy> only (<onto.owl#
    isAttackedBy> some <onto.owl#AConflictFree
  >))

SubClassOf:
  <onto.owl#AConflictFree>
```

demonstrates the definition of a conflict free subset `AConflictFree` for an argument set A in an argumentation framework that also includes an argument set I .

For each conflict-free set, an admissible subset is defined as a class whose `isAttackedBy` relation is only to arguments that themselves have the `isAttackedBy` relation to some (`owl:someValuesFrom`) arguments from the initial conflict-free set. Thus, if an argument is not attacked by any other arguments, it also belongs to the admissible set. Listing 2 shows the definition of the admissible set `AAdmissible`.

4 Method

Manual modeling of peer reviews as abstract argumentation frameworks is both time-consuming and skill-intensive. We propose to apply argument mining techniques to facilitate the extraction of abstract argumentation frameworks from peer review texts.

4.1 Annotation schema

Since no existing corpus of peer reviews is annotated with abstract argumentation frameworks, our first step is to create one. We examined various text annotation tools, such as INCEpTION [13], but

found them unsuitable for our task. Arguments in peer reviews can be extremely long, and attack relations between arguments can span the entire text. Such annotations are not displayed properly in standard text annotation environments, making them difficult for annotators to process. Thus, we developed a table-based markup scheme to annotate peer reviews. In this scheme, each row of the table represents a single argument, while the columns capture various characteristics of the arguments:

- *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.

Thus, the *Side*, *Round*, and *Number* columns together form a unique complex argument identifier. The *Opponent* and *Attacks* columns are necessary for identifying the arguments being attacked:

$$attacks(a_{i,j,k,l,m}) = a_{j,i,k-1,m,n}$$

i.e., if the current argument a is characterized by the tuple (i, j, k, l, m) , where i is *Side*, j is *Opponent*, k is *Round*, l is *Number* and m is *Attacks*, then the attacked argument is characterized by the tuple $(j, i, k - 1, m, n)$, where j becomes *Side*, i becomes *Opponent*, $k - 1$ corresponds to the previous round, m becomes the *Number* of the attacked argument, and n is the next argument in the attack chain.

When constructing an abstract argumentation framework for a peer review, we create a placeholder argument for the authors' party, which represents the entire paper and serves as the root node. This argument is assigned the value 0 in the *Number* column.

The text of a peer review may include various comments from the parties that are neither arguments nor attack the comments of other parties. For instance, the first paragraph of a peer review often contains a brief summary of the article without posing questions or making comments. These types of fragments are not marked up.

The proposed annotation scheme enables the automatic conversion of annotated peer reviews into the JSON format utilized in [1] for generating OWL representations of abstract argumentation frameworks and resolving disputes with reasoning. The process for this conversion is detailed in Algorithm 1, where (S, R, N, O, A) is an array of annotation tuples with sides S , rounds R , numbers N , opponents O and attacks A .

4.2 Annotated corpus

To train argument mining models, we annotated an open peer review corpus from the MDPI publishing [19] with the proposed schema. This corpus includes 123 peer-reviewed articles from various research fields, as of June 16, 2022. The peer reviews available there are PDF, TXT, and DOCX files uploaded by reviewers through the MDPI editorial system. Additionally, the corpus contains metadata on specific peer reviews, author responses, and article details

Algorithm 1 Translating peer review annotation into an abstract argumentation framework

```

Require: Annotation  $(S, R, N, O, A)$ 
initialize an empty list of argument sets  $AS$ 
for side  $S_i \in S$  do
  for argument  $ar \in S_i$  do
     $ar_{i,r,n} = t$ 
    add  $ar_{i,r,n}$  to  $AS$ 
  end for
end for
add  $ar_{a,0,0}$  to  $AS_a$ 
initialize an empty list of attack pairs  $AP$ 
for  $(s, r, n, o, a) \in (S, R, N, O, A)$  do
  if  $a$  is 0 then:
    add  $attacks(ar_{s,r,n}, ar_{a,0,0})$  to  $AP$ 
  else
    add  $attacks(ar_{s,r,n}, ar_{o,r-1,a})$  to  $AP$ 
  end if
end for

```

in JSON format. The corpus is distributed under the Creative Commons Attribution 4.0 (CC BY) license.

It appeared that some reviews in the corpus were incomplete, making it impossible to reconstruct the attack relationships between arguments. As a result, these reviews were excluded from annotation. In total, 88 peer reviews were annotated twice by different annotators. The annotations were stored in CSV format. The resulting dataset comprised a total of 37,285 sentences. On average, sentences contain 98 characters and 16 words. The distribution of sentence lengths in characters is notably skewed to the left, as shown in Fig. 2. There are also numerous outliers, particularly among non-argument sentences, with the longest argument stretching to 706 characters. Similar patterns are observed in the distribution of word counts, detailed in Fig. 3. The longest argument comprised 110 words, while non-arguments extended up to 312 words. Additionally, the class distribution was analyzed, revealing a slight imbalance: non-argument sentences outnumber argument sentences by 4,773, as illustrated in Fig. 4.

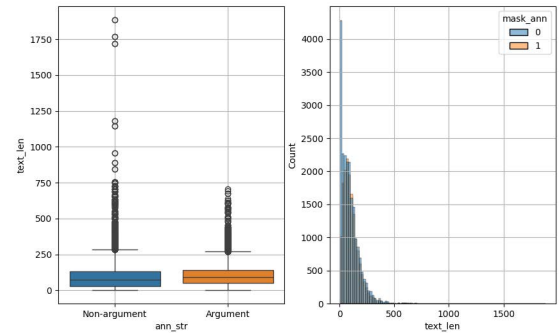


Figure 2: Distribution of sentence length in characters

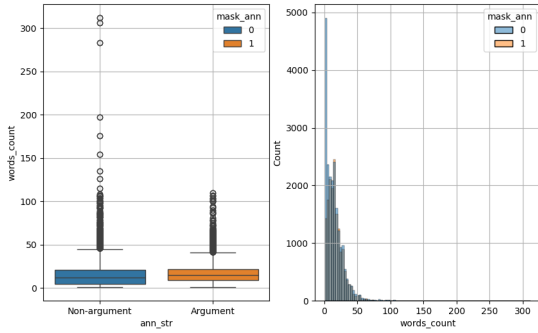


Figure 3: Distribution of sentence length in words

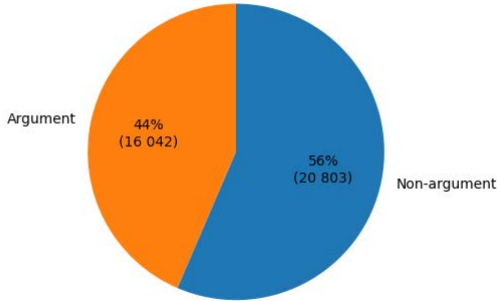


Figure 4: Distribution of classes

Example 4.1. An example of an annotated peer review is provided in Table 2. The corresponding abstract argumentation framework consists of five arguments, including the paper itself, three argument sets representing the authors and two reviewers, and four attack pairs. By applying Algorithm 1 and the workflow proposed in [1], which involves representing an abstract argumentation framework in the OWL DL language and using reasoning to classify arguments into conflict-free and admissible sets. Logs of running the Pellet reasoner [22] for the representation of this example in OWL are provided in Listing 3. These logs indicate that all three authors’ arguments were classified as acceptable, while both reviewers’ admissible argument sets were empty. Consequently, the authors have a preferred extension consisting of three arguments. The resulting graph is presented in Fig. 5.

4.3 Argument mining

In this subsection, we describe our approach to mining abstract argumentation frameworks from the annotated corpus. Since peer reviews are typically stored in formats such as PDF or DOCX, it

Listing 3: Classification of peer review arguments by Pellet reasoner

```
* Owlready2 * Pellet took 1.5788683891296387 seconds
* Owlready2 * Pellet output:

http://www.w3.org/2002/07/owl#Thing
  onto#Author
    onto#AuthorConflictFree
      onto#AuthorAdmissible - (onto#author3, onto#
        author2, onto#author1)
  onto#Reviewer_1
    onto#Reviewer_1ConflictFree - (onto#reviewer_11)
    onto#Reviewer_1Admissible
  onto#Reviewer_2
    onto#Reviewer_2ConflictFree - (onto#reviewer_21)
    onto#Reviewer_2Admissible
```

is necessary to extract their text, clean it, and align it with annotated arguments. This process involves a pipeline that includes removing stop words and uninformative characters using regular expressions. Next, we frame the argument identification task as a sentence classification problem, where sentences are classified into three categories: 1 – authors’ argument, 2 – reviewers’ argument, and 0 – non-argument. To achieve this, we segment a peer review text into sentences and apply a text-to-annotation matching algorithm. The segmentation was implemented with the NLTK library. An example of the matching result is shown in Figure 6.

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 [21] as input, the Adam optimizer and the cross-entropy loss function, 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, a dropout layer with a rate of 0.75 was added. Additionally, a simple model consisting of two fully connected layers with ReLU activation was used as a baseline for comparison. The hyperparameter values for both the baseline model and the LSTM model are provided in Table 3. The models were implemented using the PyTorch Lightning library².

5 Evaluation

To evaluate the annotation of the peer review corpus, we first measured the inter-annotation agreement between two different annotators for each peer review, resulting in Krippendorff’s $\alpha = 0.81$, which is higher than that of existing corpora. Next, we converted each annotation into JSON format using Algorithm 1 and applied the dispute resolution technique presented in [1]. Each OWL representation of the abstract argumentation framework was then processed using Pellet to classify the arguments. All annotated peer reviews were successfully processed and resolved.

Finally, we trained the argument mining models on the annotated 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

²<https://lightning.ai/docs/pytorch/stable/>

Table 2: Example of an annotated peer review

Side	Opponent	Round	Number	Attacks	Text
Reviewer1	Author	1	1	0	"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..."
Reviewer2	Author	1	1	0	"There is now available a great resource for the gene discovery... This should be mentioned and discussed in the text..."
Author	Reviewer1	2	1	1	"We have rephrased the four paragraphs where the referee found that the described examples are not always easy to follow."
Author	Reviewer2	2	2	1	"We have added a short "outlook-type" paragraph towards the end of the conclusions..."

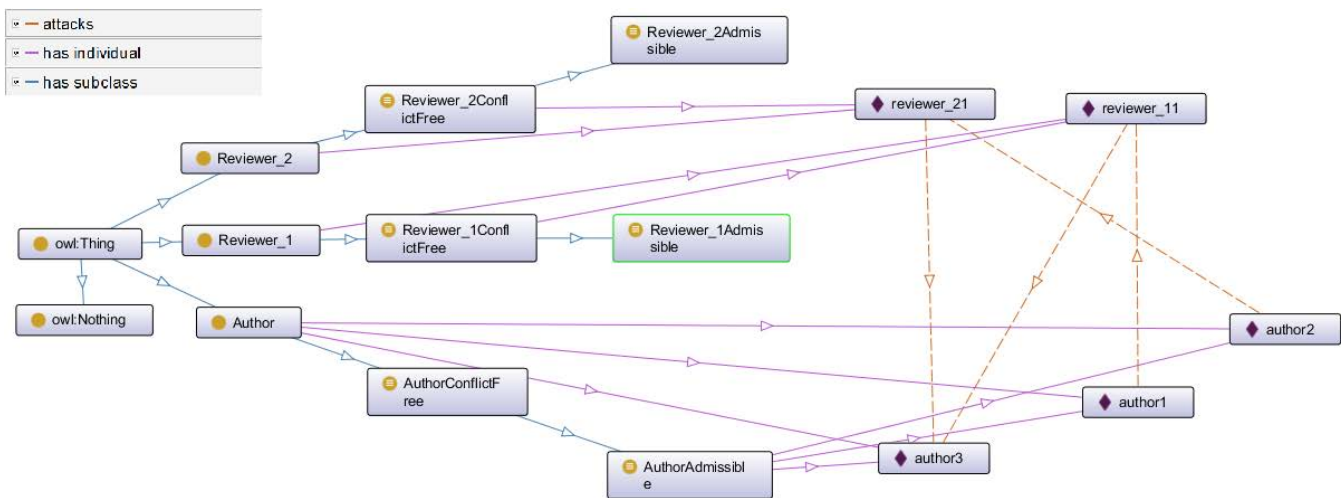


Figure 5: Resulting abstract argumentation framework in OWL for Example 4.1

Legend: Non-argument Author argument Reviewer argument Total classes num: 3

ISSN 2075-1729 www.mdpi.com/journal/life Peer-Review Record: A Manual Curation Strategy to Improve Genome Annotation: Application to a Set of Haloarchael Genomes Friedhelm Pfeiffer * and Dieter Oesterhelt Life 2015, 5, 1427-1444, doi:10.3390/life5021427 Reviewer 1: Gerald Schäfer, Inke Knecht Reviewer 2: Anonymous Editors: Hans-Peter Klenk, Michael W. Adams and Roger A. Garrett (Guest Editor of Special Issue "Archaea: Evolution, Physiology, and Molecular Biology") Received: 2 April 2015 / Accepted: 25 May 2015 / Published: 2 June 2015 First Round of Evaluation Round 1: and Author Response The manuscript by Pfeiffer and Oesterhelt describes the problem of "over-annotation" of genome/protein databases and its propagation by automated annotation pipelines. The authors describe an alternative, manual curation strategy aimed at avoiding over-annotation. While not especially flashy or exciting, the authors tackle a serious problem, the reliability of annotations in publically available databases. The authors are to be complimented for making the effort to provide well curated annotations for a set of haloarchaeal genomes, and contribute to improving public databases (EMBL/Genbank, UniProt). The manuscript describes in detail the approach used by the HaloLex genome annotation system to obtain reliable, consistent annotations, based on "Gold Standard Proteins". The manuscript is well written and good to read, and the only point of critique this reviewer would like to raise is that some of the examples given are not easy to follow. Giving specific examples is certainly a good idea. 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, such as examples given on p. 7 starting line 196 (The main difference here is between a ferredoxin-dependent and/or a NAD-dependent decarboxylation of pyruvate into acetyl-CoA? A few lines later when I read that "halophilic archaea do not contain the coenzyme methanopterin", I had to go back and reread the preceding sentences to get an idea of the argument.), p. 10 starting Line 299, and p. 12 starting Line 354. Response: We have rephrased the four paragraphs where the referee found that the described examples are not always easy to follow. R2 Round 1: and Author Response I received this manuscript and looked

Figure 6: Example of matching peer review text with the annotated arguments

Table 3: Hyperparameters of the models

Hyperparameter	Baseline	LSTM
batch_size	32	64
embedding_dim	350	350
max_length	350	110
vocab_size	30522	30522
optimizer/name	Adam	Adam
optimizer/lr	0.001	0.0006
hidden_dim	350	18
hidden_dim2	256	12
n_layers	1	1
dropout	0	0.75
loss	cross_entropy	cross_entropy

employed during the data split. The dynamics of loss and accuracy for the trained LSTM model are shown in Figure 7. The performance metrics for both the baseline and LSTM models are listed in Table 4.

Table 4: Performance of the models

	Precision	Recall	Accuracy	F1
Baseline	0.5769	0.5111	0.5606	0.5420
LSTM	0.6344	0.6267	0.6804	0.6305

Thus, the LSTM model correctly classifies sentences in approximately 68% of cases. Additionally, the Precision value is slightly higher than the Recall value, indicating that the model is more likely to avoid Type II errors. It is also worth noting that the performance of the LSTM model is significantly higher than that of the baseline and is comparable to models trained on the sentence level of other datasets.

6 Conclusion

Thus, we have advanced the automation of dispute resolution in scientific peer reviewing using a neuro-symbolic approach. Specifically, we addressed the task of identifying parties' arguments in peer reviews to represent them as abstract argumentation frameworks, which facilitate dispute resolution through logical inference. We annotated abstract argumentation frameworks within a corpus of scientific peer reviews and assessed the inter-annotation agreement, achieving a high Krippendorff's alpha of 0.81. Each annotated framework was validated by ensuring it could resolve the dispute.

Having the annotated corpus, we approached the identification of arguments in review texts by classifying sentences into three categories: authors' argument, reviewers' argument, and non-argument. A literature review indicated that using BERT embeddings with an LSTM model performs best for argument identification. We implemented this pipeline and compared it with a simpler model consisting of two fully connected layers. Our implementation achieved an accuracy of 0.634 and an F1 score of 0.631, which are comparable to models trained on other datasets. However, our approach stands out by enabling the processing of the extracted argumentation with logical inference.

Discussion and future work

In our approach, we focus only on static argumentation frameworks, where all parties' arguments are already present. In a real-world peer review scenario, the proposed method can assist metareviewers and editors who make a final decision after the interaction between authors and reviewers has concluded. For the interaction process itself, dynamic argumentation frameworks can be applied [2]. Alternatively, structuring the peer review process by input could improve efficiency, removing the need for argument mining techniques. In the future, we plan to develop user interfaces for peer review that support structured argument input.

Given the structure of the peer review process, the argumentation frameworks resulting from it are well-founded, meaning they consist only of finite sequences of attacks. According to [4], well-founded frameworks have a unique complete extension that is grounded, preferred, and stable. Therefore, for any peer review structured by abstract argumentation frameworks, it is possible to identify its acceptable arguments.

Although our approach does not consider the weight of individual arguments, it still can help metareviewers and editors by providing a more rigorous and systematic method for evaluating peer review results. Notably, given all papers in the annotated corpus were ultimately accepted by MDPI, our analysis shows that in 46.6% of cases, editors accepted papers without fully addressing all reviewers' concerns. While some of these unaddressed points may have been minor, incorporating our approach could help improve the overall quality of published work.

While our performance is comparable to the state of the art on other datasets, it remains unsatisfactory in absolute terms. Improvements can be made in two directions: 1) increasing the amount of training data by annotating peer reviews from additional sources such as OpenReview, which would also enhance the robustness and generalizability of the results, and 2) employing more advanced models. In particular, exploring the potential of large language models for extracting argumentation frameworks from peer reviews presents a promising direction for future research.

In this paper, we take the first step towards mining argumentation frameworks from text by extracting individual arguments. As the next step, we plan to explore the effectiveness of relation mining techniques to identify attack relations between arguments, aiming to fully automate the extraction of abstract argumentation frameworks from scientific peer reviews.

Data availability

Data and code are accessible on GitHub³ and, as a reborn article [24], in the Open Research Knowledge Graph at <https://doi.org/10.48366/R746064>. The supplementary data is licensed under CC0 and accessible at <https://doi.org/10.57702/elkep9bh>.

Acknowledgments

We would like to acknowledge the funding by the German Ministry of Education and Research (BmBF) for the project KISSKI AI Service Center (01IS22093C) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence

³https://github.com/Karpovich-alex/mdpi_argumentations

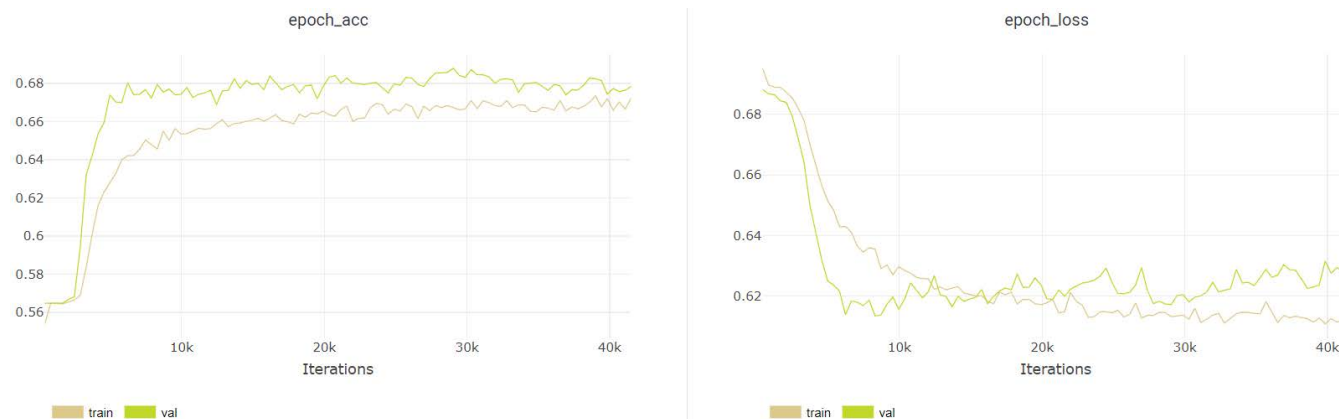


Figure 7: Accuracy and loss for the LSTM model

Strategy – EXC 2163/1 - Sustainable and Energy Efficient Aviation – Project-ID 390881007. The support from the Russian Science Foundation, project No. 20-18-00158, realised at St. Petersburg State University, is kindly recognized.

References

- [1] Ildar Baimuratov, Elena Lisanyuk, and Dmitry Prokudin. 2023. Dispute Resolution with OWL DL and Reasoning. In *Proceedings of the 36th International Workshop on Description Logics (DL 2023)*.
- [2] Stefano Bistarelli, Lars Kotthoff, Francesco Santini, Carlo Taticchi, et al. 2018. Containerisation and Dynamic Frameworks in ICCMA'19. In *SAFA@COMMA*. 4–9.
- [3] Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. 2021. AI-assisted peer review. *Humanities and Social Sciences Communications* 8 (01 2021). <https://doi.org/10.1057/s41599-020-00703-8>
- [4] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77, 2 (Sept. 1995), 321–357. [https://doi.org/10.1016/0004-3702\(94\)00041-X](https://doi.org/10.1016/0004-3702(94)00041-X)
- [5] Rafael D'Andrea and James P O'Dwyer. 2017. Can editors save peer review from peer reviewers? *PloS one* 12, 10 (2017), e0186111.
- [6] Beatriz Fisas Elizalde, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, editors. LREC 2016. Tenth International Conference on Language Resources and Evaluation; 2016 May 23-28; Portoroz, Slovenia.[Paris]: ELRA; 2016. p. 3081-8. ELRA (European Language Resources Association)*.
- [7] Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021. Argument mining driven analysis of peer-reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4758–4766.
- [8] Damián Ariel Furman, Stephanie Anneris Malvicini, Maria Vanina Martinez, Paulo Shakarian, Gerardo Ignacio Simari, and Yamil Osvaldo Soto. 2023. A Neuro-symbolic Approach to Argument Comparison in Structured Argumentation. In *AI²@AI^{*} IA*.
- [9] Andrea Galassi, Marco Lippi, and Paolo Torroni. 2021. Investigating logic tensor networks for neural-symbolic argument mining. In *Proc. 1st Int. Joint Conf. Learn., Reasoning*. 1–7.
- [10] Nancy L Green. 2018. Towards mining scientific discourse using argumentation schemes. *Argument & Computation* 9, 2 (2018), 121–135.
- [11] Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. *arXiv preprint arXiv:1903.10104* (2019).
- [12] Janine Huisman and Jeroen Smits. 2017. Duration and quality of the peer review process: the author's perspective. *Scientometrics* 113, 1 (October 2017), 633–650. <https://doi.org/10.1007/s11192-017-2310-5>
- [13] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: System demonstrations*. 5–9.
- [14] Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018. ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*. 22–28.
- [15] Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. Association for Computational Linguistics.
- [16] John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics* 45, 4 (2020), 765–818.
- [17] Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology* 64, 1 (2013), 2–17. <https://doi.org/10.1002/asi.22784> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22784>
- [18] Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020*. IOS Press, 2108–2115.
- [19] Marcin Milkowski and Ksawery Jasiński. 2022. MDPI Open Peer Review Corpus. <https://doi.org/10.18150/D5L2EK>
- [20] Henry Prakken. 2010. An abstract framework for argumentation with structured arguments. *Argument & Computation* 1, 2 (2010), 93–124.
- [21] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
- [22] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. 2007. Pellet: A practical owl-dl reasoner. *Journal of Web Semantics* 5, 2 (2007), 51–53.
- [23] Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *arXiv preprint arXiv:1802.05758* (2018).
- [24] Markus Stocker, Lauren Snyder, Matthew Anfuso, Oliver Ludwig, Freya Thießen, Kheir Eddine Farfar, Muhammad Haris, Allard Oelen, and Mohamad Yaser Jaradeh. 2024. Rethinking the production and publication of machine-reusable expressions of research findings. (2024). <https://doi.org/10.48550/ARXIV.2405.13129>
- [25] Simone Teufel, Advait Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*. 1493–1502.
- [26] Frans H Van Eemeren and Rob Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.
- [27] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36 (2021), e5.
- [28] Douglas Walton. 2012. Argument mining by applying argumentation schemes. *Studies in Logic* 4, 1 (2012), 2011.
- [29] Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Received 31 July 2024; revised 30 October 2024; accepted 11 December 2024