Argumentation Mining in Scientific Literature for Sustainable Development

Aris Fergadis[1,2](✉), **Dimitris Pappas**[2], **Antonia Karamolegkou**[2], and **Haris Papageorgiou**[2]

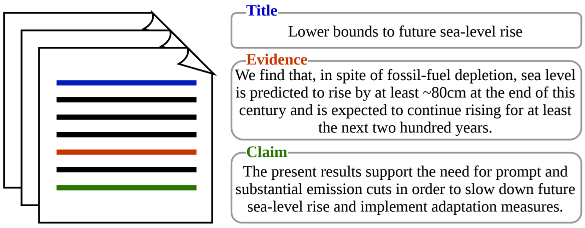
[1]School of Electrical and Computer Engineering, National Technical University of Athens
[2]Athena Research and Innovation Center
aris.fergadis@athenarc.gr

Abstract

Science, technology and innovation (STI) policies have evolved in the past decade. We are now progressing towards policies that are more aligned with sustainable development through integrating social, economic and environmental dimensions. In this new policy environment, the need to keep track of innovation from its conception in Science and Research has emerged. Argumentation mining, an interdisciplinary NLP field, gives rise to the required technologies. In this study, we present the first STI-driven multidisciplinary corpus of scientific abstracts annotated for argumentative units (AUs) on the sustainable development goals (SDGs) set by the United Nations (UN). AUs are the sentences conveying the Claim(s) reported in the author's original research and the Evidence provided for support. We also present a set of strong, BERT-based neural baselines achieving an f1-score of 70.0 for Claim and 62.4 for Evidence identification evaluated with 10-fold cross-validation. To demonstrate the effectiveness of our models, we experiment with different test sets showing comparable performance across various SDG policy domains. Our dataset and models are publicly available for research purposes[1].

1 Introduction

Arguments are the fundamental building blocks (i.e., groups of statements) in the reasoning path from assumptions to conclusions. Argumentation mining (AM[2]) is becoming increasingly a popular topic that addresses the issue of converting unstructured text into structured argument data (Green et al., 2014; Cardie et al., 2015; Reed, 2016; Habernal et al., 2017; Slonim and Aharonov, 2018; Stein and Wachsmuth, 2019).



Lower bounds to future sea-level rise

We find that, in spite of fossil-fuel depletion, sea level is predicted to rise by at least ~80cm at the end of this century and is expected to continue rising for at least the next two hundred years.

The present results support the need for prompt and substantial emission cuts in order to slow down future sea-level rise and implement adaptation measures.

Figure 1: Example of an annotated paper. The authors (Zecca and Chiari, 2012) claim that a substantial emission cut is needed to slow down sea-level rise. The evidence they provide is the result of their method that predicts a continued rising for at least the next two hundred years.

AM datasets have been developed in various domains such as legal collections (Mochales and Ieven, 2009; Savelka and Ashley, 2016; Yamada et al., 2017), political debate fora (Walker et al., 2012; Abbott et al., 2016), persuasive essays (Stab and Gurevych, 2014; Carlile et al., 2018), biomedical publications (Wilbur et al., 2006; Blake, 2010; Achakulvisut et al., 2019; Mayer et al., 2020), newspapers, blogs, and the social web (Goudas et al., 2015; Kiesel et al., 2015; Habernal and Gurevych, 2017).

These datasets facilitate practical applications of argumentation mining such as supporting sensemaking (Schneider, 2014), practical reasoning (Walton, 2015), argument retrieval (Rastegar-Mojarad et al., 2016), sentiment analysis (Liu et al., 2017), and claim verification (Thorne et al., 2018; Wadden et al., 2020). We present a novel AM framework streamlining evidence-informed policy making.

The *Resolution adopted by the UN General Assembly on 25th of September 2015*[3] set 17 interlinked SDGs and 169 targets to "...stimulate action over the next 15 years in areas of critical importance for humanity and the planet". In this context, policy makers and public administrators seek evidence-based scientific claims assisting the formulation of optimal STI policy for achieving

[1] https://github.com/afergadis/SciARK

[2] Both argument and argumentation mining terms are used in the literature.

[3] https://undocs.org/A/RES/70/1

Proceedings of The 8th Workshop on Argument Mining, pages 100–111
Punta Cana, Dominican Republic, November 10–11, 2021. ©2021 Association for Computational Linguistics

the sustainable path of development. To this end, we construct SciARK, a novel multidisciplinary dataset with abstracts of scientific publications related to six Sustainable Development Goals (SDGs) of the United Nations (UN). SciARK abstracts are annotated for their *Argument Units*, i.e., sentences in which the authors state their *Claims* and the *Evidence* to support them. We present such an example in Figure 1.

The term *Claim* in SciARK refers to the main findings reported in the authors' original research and usually coincides with the conclusions. In addition, we use the term *Evidence* for the sentences that refer to particular kinds of arguments, such as those based on observations, factual findings, statistics, experimental tests or other scientific findings. Implicitly, all the relations are *Supporting*, and we justify our decision in Section 3.1. We develop end-to-end neural baselines on SciARK, showing that all baselines can benefit from training on a diverse set of SDG policy domains[4]. The major contributions of this paper can be summarised as follows:

1. We introduce a novel STI-driven multidisciplinary dataset with argumentation annotation. The dataset covers six of the 17 Sustainable Development Goals of United Nations covering a wide range of sustainability aspects, from Climate to Clean Energy to Health and well-being to Gender equality and responsible consumption and production.

2. We formalise and develop an annotation protocol assuring quality by accounting for different groups of annotators curating publications in different domains.

3. We interpret the magnitude of inter-annotator agreement following the Cumulative Probability approach (Gwet, 2014), instead of a simple scale matching.

4. We provide three strong modern deep learning baselines addressing the claim/evidence extraction as a classification task, showing their effectiveness in a diverse set of sustainability domains.

5. We demonstrate the capacity of our models and dataset in tackling unseen SDG domains. Compared to previous domain-specific approaches, our best model can handle snippets of claims and evidence in a broad spectrum of SDG domains.

2 Related Work

A large number of annotated datasets has been released for argumentation mining in various domains (Lawrence and Reed, 2019). Many researchers approach the claim annotation task under the prism of functional roles (Blake, 2010; Alamri and Stevenson, 2016) or ownership (Lauscher et al., 2018; Mayer et al., 2020). Evidence annotation has analogous categorisation, ranging from its strength (Wilbur et al., 2006; Shatkay et al., 2008) to reproduction and generalisation (Mayer et al., 2018). However, it is not always possible to apply these fine-grained differentiations to other domains.

Recently, researchers employed argumentative schemes that incorporate both argumentative units and their relationships. The most common argumentative relationships employed are the "support" and "attack" ones (Peldszus and Stede, 2013; Mayer et al., 2020). Researchers also adapt additional relation types from Rhetorical Structure Theory (Mann and Thompson, 1988) such as "semantically same" (Lauscher et al., 2018), "detail", "sequence" (Kirschner et al., 2015), and "additional" (Accuosto and Saggion, 2019). Although those fine-grained relation types are of great value, their presence in the context of a scientific abstract is limited.

The most recent methods for the AU classification task leverage neural networks architectures. A common architecture is the one comprising a BiLSTM layer followed by a CRF layer for sequence tagging (Achakulvisut et al., 2019; Accuosto and Saggion, 2019). Mayer et al. (2020) evaluate many different neural network models and shows that those based on Transformers have better performance on the AU extraction and the relation classification tasks. The before mentioned models classify sequences of words within a AU ignoring the context before and after the AU. We experiment with Transformers and BiLSTM layers taking advantage of the context of AUs.

Most of the datasets based on scientific literature are domain specific with strong emphasis on the biomedical domain (Blake, 2010; Alamri and Stevenson, 2016; Achakulvisut et al., 2019; Mayer et al., 2020). Other domains are educa-

[4] We also use the term "SDG domain" or "SDG" for short.

tion (Kirschner et al., 2015), computer graphics (Lauscher et al., 2018), and computational linguistics (Accuosto and Saggion, 2019). SciARK, as a multidisciplinary dataset, includes abstract from biomedical, social, environmental and other domains.

3 Corpus

3.1 Annotation Schema

A popular schema in argumentation mining application is Toulmin's model (Lytos et al., 2019). Toulmin (1958) defines the functional roles of *datum, claim, warrant, backing, qualifiers,* and *rebuttal*. We use a subset of functional roles, comprising datum and claim, as a scheme for annotating our corpus (Lauscher et al., 2018; Stede and Schneider, 2019).

Arguments due to their pragmatic nature can be expressed by a variety of forms and linguistic cues. This is why there have been efforts to classify the two basic argument components, claim and evidence, into further categories. Blake (2010) introduced a Claim Framework distinguishing five claim categories, but they found that the majority of the claims in a corpus of scientific publications were explicit claims. Similarly, Mayer et al. (2018) suggest fine-grained evidence categories, but they do not use those in their extended dataset (Mayer et al., 2020). Regarding argument relations, there have been also attempts to distinguish them into supporting and attacking. However, Accuosto and Saggion (2019) did not found any attack relation in 60 abstracts of the SciDTB corpus (Yang and Li, 2018). Also, Lauscher et al. (2018) and Mayer et al. (2020) found that the number of attacking relations in full papers and abstracts is relatively low.

Based on the reported studies and the context of our corpus, we choose to keep our annotation schema simple. We focus on the kernel of an argument that is the *Evidence* and the *Claim* without further categorisation. Our argumentative units are sentences, a decision following Kirschner et al. (2015), and supported by Accuosto and Saggion (2019) that report 93% of argumentative units in their corpus coincide with the boundaries of the sentences. Implicitly, a supportive relation holds between Evidence and Claim.

In SciARK, we define Argument, Claim and Evidence, in the context of scientific abstract, as follows:

Argument: a set of statements with two different categories: Claim and Evidence.

Claim: an argumentative statement that reports the study findings and derives from the author's original work.

Evidence: statement that reports observations, statistical findings, and experimental results used to support a claim.

SDG	Annotators	Abstracts
3	6	300
5	5	255
7	3[a]	61
10	3[a]	70
12	3[a]	52
13	6	262
Total	20	1000

[a] The same annotators in SDGs 7, 10 and 12.

Table 1: Number of annotators and annotated abstracts grouped by SDG policy domain.

3.2 Data Collection

Our dataset is the first one that connects *Policy Targets* and *Scientific Literature*. In SDGs, we find specific targets defined by policymakers to deliver a more sustainable, prosperous and peaceful global future. We collect (a small part of) the scientific literature that provides scientific arguments related to real policy targets. To formulate the policy targets, we leverage the definition of the SDG targets and their indicators as keyterms (Duran-Silva et al., 2019) for searching in publisher's portals and scientific literature search engines (e.g., PubMed, Semantic Scholar etc.). The keyterms we use are from the SDGs 3 (good health and well-being), 5 (gender equality), 7 (affordable and clean energy), 10 (reduce inequalities), 12 (responsible consumption and production), and 13 (climate action). Examples of such keywords are: (Neonatal OR Maternal) Mortality, Female Genital Mutilation, Clean (Fuels OR Fossil-Fuel Technology), (Climate OR Natural Disaster) Resilience, etc.

We use the term *domain* to group abstracts under policy targets as described in SDGs targets. So, all the abstracts we gather using keyterms from the SDG5 targets, form the SDG5 policy domain, and so on. In Table 1, we present the number of abstracts collected per policy domain.

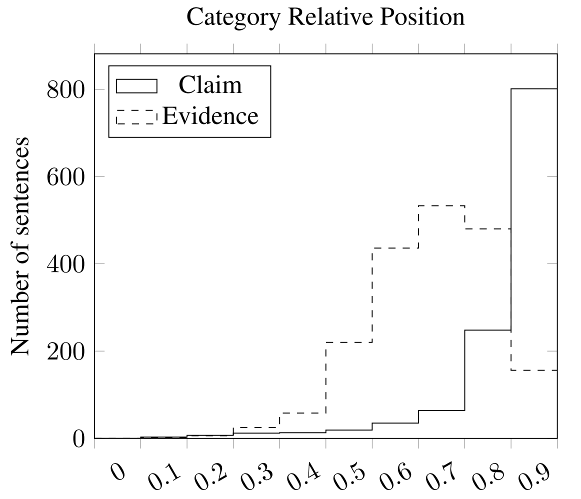


Figure 2: Histogram with the relative positions of Claim and Evidence categories within an abstract.

3.3 Exploratory Data Analysis

SciARK consists of 1,000 abstracts with 12,374 sentences, 1,202 sentences annotated as Claim and 1,915 annotated as Evidence. On average, per abstract, 1.2 sentences are annotated as Claim and 1.92 sentences as Evidence.

To investigate the position of the Claim and Evidence categories within the abstracts, we calculate the relative position for each sentence. The first sentence is in relative position 0 and the last in position 1. Figure 2 depicts the positions of the categories displaying the histograms of the two argumentative units. A Claim is located mainly in the last sentences. Specifically, about 40% of the total claims are in the last sentence and about 80% in the $[0.9 - 1]$ range. That is expected for scientific abstracts because usually, the claim coincides with the conclusions, which are the last sentences of an abstract.

The Evidence category lies in the middle to the end of the abstracts. We find the Evidence category in the range $[0.8 - 1]$ with a percentage of about 33% and about 62% in the range $[0.5 - 0.8)$ where, usually, we find the results and their analysis. These findings serve as a baseline for SciARK evaluation in Section 5.2.

Figure 2 shows that the most common pattern of the argument structure is the one that Claim follows the Evidence. Table 2 shows the trend on each domain. The main reason that this pattern is almost a rule on SDG3 and SDG5 is that many abstracts are from the biomedical domain. Most of those abstracts follow the IMRAD (Sollaci and Pereira, 2004) or the CONSORT format (Hopewell et al., 2008) that instruct the discussion/conclusions to be at the end of the abstract.

SDG	3	5	7	10	12	13
Structure						
Evidence / Claim (%)	99.67	95.29	88.52	90.00	86.54	88.17
Claim / Evidence (%)	00.34	04.71	11.48	10.00	13.46	11.83

Table 2: The argument structure on the corpus, i.e., Evidence followed by Claim or Claim followed by Evidence.

SDG	3	5	7	10	12	13
κ_{a_1}	.671	.595	.596	.665	.628	.557
κ_{a_2}	.621	.610	.641	.657	.653	.518
κ_{a_3}	.721	.740	.567	.577	.503	.525
κ_{a_4}	.599	.725				.542
κ_{a_5}	.622	.748				.582
κ_{a_6}	.705					.553

Table 3: Annotator assessment coefficients (κ_{α_i}). Each cell represents an annotator and the allocation represents the annotator distribution to each SDG.

4 Assessing Agreement

4.1 Annotators

The task of annotating such abstract constructs as the argumentative units is quite challenging (Lippi and Torroni, 2016). Also, our multidisciplinary corpus requires experts from a variety of scientific domains, which was beyond our scope. Twenty postgraduate students with background in engineering, economics, and applied sciences, provided annotations (Table 1) in a distributed fashion. The annotators selected the SDGs that were more familiar with and felt more comfortable to work on. Each annotator worked independently on an average of 150 abstracts, and three annotators annotated each abstract. Annotators should categorise all the sentences in an abstract into one of the mutually exclusive categories: *Claim*, *Evidence*, or *Neither*. We used MACE (Hovy et al., 2013) and majority vote between all annotator triplets to get category predictions. The predictions of MACE and the majority vote output were identical.

4.2 Annotator Assessment

In tasks which incorporate a large number of annotators or use crowd-sourcing, one needs to have methods to filter out biased annotators. To assess

Strength of Agreement	Substantial [0.6 - 0.8)	Moderate [0.4 - 0.6)
Corpus ($\kappa = .669$)	88%	100%
Claim ($\kappa_C = .730$)	100%	
Evidence ($\kappa_E = .637$)	13%	100%
Neither ($\kappa_N = .664$)	82%	100%

Table 4: Estimated agreement coefficient interpretation using the cumulative probability approach (Gwet, 2014) on the Landis-Koch scale (Landis and Koch, 1977). Using the simple match to the benchmark, all kappa values are interpreted as *Substantial*.

the reliability of the annotators we use the *Annotator* κ (Toledo et al., 2019) as an annotator quality control coefficient. This metric is calculated as the average of all pairwise agreement for each annotator. Toledo et al. (2019) use Equation 1 to calculate the Annotator κ of an annotator i

$$\kappa_{\alpha_i} = \frac{\sum_{j=1, j \neq i}^{n} F(i,j)}{N_i} \tag{1}$$

where n is the total number of annotators, $F(i,j)$ is the pairwise Fleiss' κ between annotators i, j, and N_i the number of annotator pairs that the annotator i is a member.

The rationale of this metric is that an annotator who systematically disagrees will have all pairwise scores low and consequently a low κ_α value. The opposite holds for a reliable annotator.

Table 3 shows the values we calculate on our corpus. Each row $\kappa_{\alpha_i}, i \in [1, 6]$ corresponds to an annotator who worked on an SDG. Annotator κ_{α_1} in SDG3, is a different person than annotator κ_{α_1} in SDG5. Some rows have empty values because there were fewer than 6 annotators. The table follows the allocation we present in Table 1.

We expect to capture biased annotators as outliers in the columns of Table 3. Toledo et al. (2019) use the value .35 as a threshold to discard annotators with lower Annotator κ values. In our case, all annotators achieve a satisfactory average pair-wise agreement, despite the difficulty of the task and the lack of expertise in every domain.

4.3 Inter-Annotator Agreement

Fleiss' κ (Fleiss, 1971) was run to estimate the magnitude of agreement between annotators resulting[5] in κ_F=.669 (95% Confidence Interval (CI), .658 to .681), Standard Error (SE) .006, p < .001.

To further investigate the difficulties of the annotation task, we calculate the level of agreement on each of the three categories computing the individual kappa values. The individual kappas are simply Fleiss' κ calculated for each of the categories separately against all other categories combined. The Claim category has an estimated level of agreement equal to κ_C=.730 (95% CI, .713 to .746), SE .009, p < .001. Agreement for the Evidence category is equal to κ_E=.637 (95% CI, .622 to .652), SE .008, p < .001, and for the Neither category κ_N=.664 (95% CI, .652 to .676), SE .006, p < .001. The results are statistically significant and show an agreement above the agreement expected by chance.

Many benchmark scales (Landis and Koch, 1977; Cicchetti and Sparrow, 1981; Altman, 1990; Regier et al., 2013) aim to interpret the magnitude of agreement using the inter-annotator agreement coefficients. Usually, this is done by simple matching the calculated coefficient value within a benchmark range and report the corresponding interpretation.

However, Gwet (2014) demonstrated that this approach is highly optimistic for the characterisation of the agreement. Thus, Gwet recommends the *"Cumulative Probability"* approach, a probabilistic process that takes into account the standard error to calculate the likelihood that a coefficient falls into the benchmark range of values. Using the Landis-Koch scale (Landis and Koch, 1977) and the Cumulative Probability approach (Gwet, 2014), we report[6] our interpretation in Table 4. The results indicate that the likelihood of our corpus to fall into the substantial range of values is 88%. Also, the results show that the agreement can be characterised as moderate with a likelihood of 100%. Claim agreement is substantial with a likelihood of 100%, while the Evidence category is harder for the annotators as we have a moderate agreement (only 13% likelihood of substantial agreement).

The Claim category is usually found in the concluding sentences of the abstract. Also, we find strong discourse markers introducing the claim, such as "overall, this study reveals that", "in conclusion, these findings confirm", "the data suggest that" etc. The above observations mainly explain the substantial agreement on the Claim category. A source of disagreement is found in sentences in

[5] For the calculations we use the library irrCAC v1.0 for R (https://cran.r-project.org/web/packages/irrCAC/)

[6] One can easily confirm the results using the beforementioned library, the coefficient values with their corresponding standard error values reported in this section.

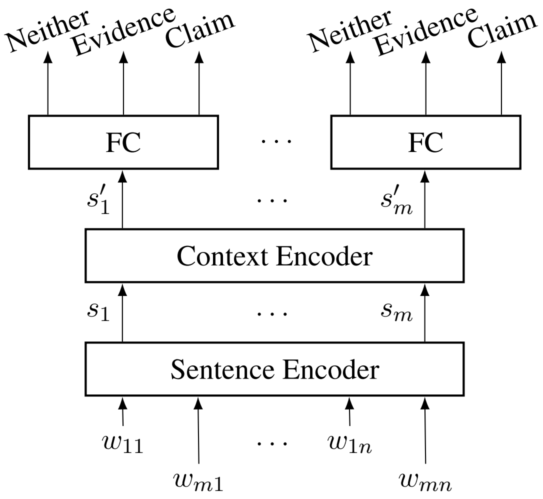


Figure 3: Diagram of our neural network architecture for the Sentence Sequence Tagging task. The *Sentence Encoder* layer outputs m sentence vectors for the m sentences of an abstract with n words each, and the *Context Encoder* updates the sentence vectors utilising their context. A *Fully Connected* layer gives a label for each sentence.

which the authors express possibility (may) or opinion ("we propose/suggest/believe"). Our annotation protocol is to avoid annotating such sentences as Claim *unless* there are no other declarative sentences baring the claim *and* there are sentences that provide evidence to support the claim.

Sentences annotated as Evidence do not have such strong discourse markers as those found for the Claim. For the Evidence category, annotators had to choose mostly between sentences that report experimental results or observations. The instructions to the annotators were to select the minimum number of sentences that provide enough support to the Claim. We pose this restriction to avoid the case to categorise as Evidence every sentence reporting results. The increased number of candidate sentences and the restriction mentioned are the main reasons that explain the level of agreement for the Evidence category.

Finally, the agreement on the Neither category is affected by the Evidence category and is characterised as moderate but with a likelihood of 82% to be substantial. Overall, we show that the strength of agreement in our corpus is moderate above chance and with a very high probability can be characterised as substantial.

5 Argumentation Units Classification

5.1 Setup

We formulate the argumentation units classification as a Sentence Sequence Tagging task within the context of the abstract. In Figure 3, our SciARK architecture is depicted encompassing three layers: a) the *Sentence Encoder*, b) *Context Encoder*, and c) a *Fully Connected* layer.

The input of the Sentence Encoder is a matrix $\mathcal{A} \in \mathbb{R}^{m \times n}$ that represents an abstract $\mathcal{A}$ with m sentences of n words each. We set $m = 20$ and $n = 40$, values representing 90% of the corpus in the number of sentences and number of words in a sentence, respectively. In cases where the sentences of an abstract exceed the limit, we truncate the sentences in the beginning. The truncated sentences get the Neither label.

The output is a sentence vector $\boldsymbol{s} \in \mathbb{R}^d$, where d is the output dimensionality of the layer. A Context Encoder layer updates each sentence vector $\boldsymbol{s}$ utilising the *context*, before and after, each sentence to a new vector $\boldsymbol{s}' \in \mathbb{R}^{d'}$. Finally, a Fully Connected layer classifies each sentence as *Evidence*, *Claim*, or *Neither*.

We implement the architecture in Figure 3 as a SciBERT - BiLSTM model. The uncased version of the SciBERT (Beltagy et al., 2019) model serves as a Sentence Encoder. We use the `[CLS]` token of each sentence from the SciBERT, a sentence vector $\boldsymbol{s} \in \mathbb{R}^{728}$, as input to the Context Encoder. To implement the Context Encoder layer, we use a bidirectional LSTM. We set the LSTM layer size to 64 with 0.3 forward and backward dropout. The values were selected by testing all the combinations of the following hyper-parameters: LSTM units 32, 64 and 128, forward and backward dropout 0.1, 0.3 and 0.5. Thus, the output of the Context Encoder is a vector $\boldsymbol{s}' \in \mathbb{R}^{128}$. Also, we experiment by replacing the SciBERT with an uncased BERT-base model (Devlin et al., 2018) keeping the same hyper-parameters.

As a last variation in our architecture, we replace SciBERT with a BiLSTM. The Sentence Encoder BiLSTM layer encodes the word embeddings of a sentence to a sentence vector using attention. As word embeddings, we use 200-dimensional pre-trained Glove embeddings[7] (Pennington et al., 2014) with Gaussian noise sampled from a zero-mean distribution with $\sigma = 0.1$. The LSTM layer

[7] https://nlp.stanford.edu/data/glove.6B.zip

	Claim			Evidence		
	P	R	F	P	R	F
Baseline	42.2	**72.8**	53.3	38.5	**77.5**	51.4
Mayer et al. (2020)	65.2	56.8	60.5	56.7	56.7	56.7
SciBERT	69.7	47.8	55.9	55.9	57.0	56.0
BiLSTM - BiLSTM	62.6	56.2	58.3	57.0	51.5	53.8
BERT - BiLSTM	69.6	65.8	67.3	**62.6**	57.5	59.2
SciBERT - BiLSTM	**73.5**	67.3	**70.0**	62.4	62.8	**62.4**

Table 5: The baseline and the performance of neural network models on SciARK calculated with 10-fold cross-validation.

size is 64 with 0.3 forward and backward dropout. Follows a dropout layer with 0.4 and an attention layer which outputs a 128 dimensional sentence vector, $s \in \mathbb{R}^{128}$. We also tested all combinations with 200 and 300-dimensional vectors, $\sigma \in [0.1, 0.2]$ and dropout 0.2, 0.3, 0.4 and 0.5.

To validate the contribution of the Context Encoder layer of our architecture, we present results of a) an intuitive baseline, b) a SciBERT model as Sentence Encoder only, and c) the SciBERT-GRU-CRF model of Mayer et al. (2020).

The baseline utilises the observations presented in Section 3.3. For the baseline all sentences that have a relative position within an abstract in the range $[0.5 - 0.8)$ are categorised as Evidence, those in the range $[0.8 - 1]$ as Claim and the rest as Neither. Finally, the SciBERT-GRU-CRF model of Mayer et al. (2020) gave the best results in the multi-class sequence tagging task on their dataset (AbstRCT). Since the model classifies spans of text using the BIO format, instead of sentence, we do the following modifications. If a sentence has a Claim or Evidence category, the first token of the sentence gets the B-Claim/Evidence label and all the rest the I-Clam/Evidence label. All other sentences get an O label on every token. To get a category for a sentence, we take the most common label of all tokens of the sentence.

5.2 Results

We present the results of the models described in Table 5. All results, expect the ones of the Baseline, are calculated using 10-fold cross-validation on our dataset using 20% of the training set for validation.

The Baseline has the best recall in both categories which is expected since for each abstract 30% of the sentences are categorised as Evidence and 20% as Claim. On the other hand, this method has the lowest precision.

	Prediction		
	Claim	Evidence	Neither
Claim	807 (67.1%)	84 (7%)	311 (25.9%)
Evidence	58 (3%)	1203 (62.8%)	654 (34.2%)
Neither	238 (3%)	647 (7%)	8372 (90%)

Table 6: Confusion matrix of the predictions by the SciBERT - BiLSTM model.

From the neural network models, the SciBERT - BiLSTM model has the most balanced performance with the highest f1-score in both categories and the best precision on the Claim category. Also, the results highlight the contribution of the Context Encoder comparing to the plain Sentence Encoder (SciBERT). The model of Mayer et al. (2020) does not seem to benefit from the CRF layer because the classification spans are sentences and their model does not utilise the context before and after the classified sentence.

In Table 6 we present the confusion matrix with the predictions of the SciBERT - BiLSTM model. The results show that the model has a good discrimination between the two argumentative categories. Most misclassified Claim and Evidence sentences get the Neither category. The results overall are very promising taking into account the conceptual difficulty of the task and the variety of the policy domains in the dataset.

5.3 Generalising to New Policy Domains

We utilise the multidisciplinary characteristic of our dataset to experiment on a cross-domain task, holding successively one SDG policy domain exclusively as a test set, similarly to Stab et al. (2018). With this experiment, we expect to get results showing that models benefit from the diversity of the dataset and generalise to new policy domains, comparing with a fixed one. We use AbstRCT (Mayer et al., 2020), a corpus of randomised control trials, as a fixed dataset that is comparable to the size of our dataset, is on scientific abstracts and has Claim and Evidence annotations.

The experiment has two phases. On the first one, the SciBERT-BiLSTM model is trained on the five of the SDG domains and is tested on the sixth. On the second one, the model is trained on AbstRCT and successively tested on each SDG domain.

In Figure 4 we present our results on the Claim and Evidence identification. The results clearly show that the model has better scores with the

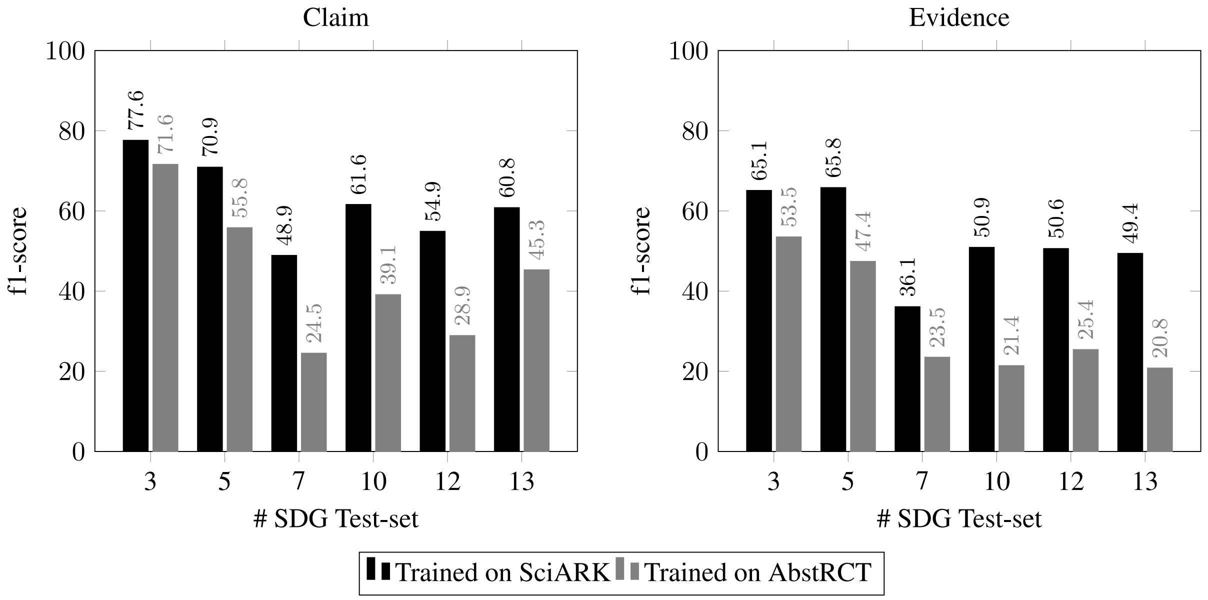


Figure 4: F1-scores for the Claim and Evidence using as a test set the SDG policy domains 3, 5, 7, 10, 12, and 13 successively. The results are from the SciBERT-BiLSTM model once trained on the remaining domains of SciARK dataset and once trained on AbstRCT dataset.

SciARK dataset. The imbalance of the dataset (Table 1), the different argument structure (Table 2), and the domains distinct characteristics are the main reasons that explain the variability of the results.

6 Discussion

In this section, we discuss some of the erroneous predictions of our model. All possible misclassifications are: Neither to Claim/Evidence, Claim to Evidence/Neither, and Evidence to Claim/Neither. In Table 7 we present at least one misclassification, from each combination, in their context in order to better understand the type of errors.

A common source of errors, for both argumentation categories, are the discourse markers discussed in Section 4.3. The example sentences 2, 6 and 10 have common claim discourse markers but in the context of those abstracts, other sentences bare the author's claim. Another source of errors are when the authors express possibility (may) or opinion ("we propose/suggest/beleive"). Annotators should categorise those sentences as Claim if there is no other declarative one that has a claim. Two examples of such sentences that the model predicted as Claim, are 12 and 15.

The model selects sentences for the Evidence category mainly from those that report results. However, as we mentioned in Section 4.3, the instruction to the annotators was to select the minimum sentences that support the claim. Following the instruction, annotators find that sentences 1, 4 and 8 are surplus. However, the model predicted those as Evidence because of their position on the abstract and the fact that they report results.

Finally, to highlight the difficulty of the task we discuss abstract #3. Sentence 11 has a claim that "under-five mortality rate is a serious problem". Although from the general knowledge of our world we recognise the statement as a claim, in the abstract there is no sentence to support it. On the other hand, the claim that "the hazard of under-five mortality has a decreasing pattern in years" is supported by sentence 10.

7 Conclusions

Argumentation mining is a young and gradually maturing research area within computational linguistics. To develop practical applications of importance, we need reliable datasets. In this paper, we describe SciARK, a novel STI-driven multidisciplinary dataset with argumentation annotation on six of the 17 SDGs of the United Nations. Our annotation protocol resulted in a reliable dataset with a significant inter-annotation agreement. We evaluated the dataset on the claim/evidence extraction task using modern deep learning models getting promising results. We also demonstrate the need for multidisciplinary datasets since they could enable models to generalise better on unseen data

	Annotators	Prediction	Sentence
#1			*Climate-driven polar motion: 2003-2015* (Adhikari and Ivins, 2016)
1	Neither	Evidence	The changes in terrestrial water storage (TWS) and global cryosphere together explain nearly the entire amplitude ...
2	Evidence	Claim	We also find that the TWS variability fully explains the decadal-like changes in polar motion observed during the ...
3	Claim	Neither	This newly discovered link between polar motion and global-scale TWS variability has broad implications for the study of ...
#2			*The impact of large scale biomass production on ozone air pollution in Europe* (Beltman et al., 2013)
4	Neither	Evidence	Although this scenario is rather conservative, our simulations project that isoprene emissions are substantially increased ...
5	Evidence	Evidence	As a consequence, ozone peak values are expected to increase by up to 6%, and ozone indicators for damage to human health ...
6	Evidence	Claim	Finally, we show that after the change in land use NOx emission reductions of 15-20 % in Europe would be required to ...
7	Claim	Claim	Because biomass production is expected to increase throughout Europe in the coming decades, we conclude that...
#3			*Comparison of under-five mortality for 2000, 2005 and 2011 surveys in Ethiopia* (Ayele and Zewotir, 2016)
8	Neither	Evidence	The effect of respondent's current age, age at first birth and educational level on the under-five mortality rate ...
9	Evidence	Neither	Regarding total children ever born, child death is more for the year 2000 followed by 2005 and 2011.
10	Neither	Claim	Conclusion: Based on the study, our findings confirmed that under-five mortality is a serious problem in the country.
11	Claim	Claim	The analysis displayed that the hazard of under-five mortality has a decreasing pattern in years.
12	Neither	Claim	Our study suggests that the impact of demographic characteristics and socio-economic factors on child mortality should ...
#4			*Renewable energy consumption-economic growth nexus in emerging countries* (Ozcan and Ozturk, 2019)
13	Evidence	Evidence	The results indicated that the neutrality hypothesis does hold for all of the markets studied except for Poland, ...
14	Claim	Evidence	As such, because of the nonexistence of causality running from renewable energy demand to economic growth, energy saving ...
15	Neither	Claim	For Poland; however, energy conservation policies may have detrimental effects on the country's economic performance level.

Table 7: Excerpts from abstracts with the categories by the annotators and the model prediction. All other sentences have the Neither category.

and outperform systems trained on domain-specific data.

SciARK facilitates the development of "Policy Intelligence" by streamlining a big data, STI-driven policy modelling approach, improving human judgement for evidence-informed policy making. Our SciARK framework will be further exploited in adapting policies in the continuously evolving STI landscape, addressing sustainable development.

Acknowledgement

We thank all the annotators who have contributed to the annotations of our training data. This research work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 101004870 (the IntelComp Project).

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A. Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Pablo Accuosto and Horacio Saggion. 2019. Discourse-driven argument mining in scientific abstracts. In *Natural Language Processing and Information Systems - 24th International Conference on Applications of Natural Language to Information Systems, NLDB 2019, Salford, UK, June 26-28, 2019, Proceedings*, volume 11608 of *Lecture Notes in Computer Science*, pages 182–194. Springer.

Titipat Achakulvisut, Chandra Bhagavatula, Daniel E. Acuña, and Konrad P. Körding. 2019. Claim extraction in biomedical publications using deep discourse model and transfer learning. *CoRR*, abs/1907.00962.

S. Adhikari and E. Ivins. 2016. Climate-driven polar motion: 2003–2015. *Science Advances*, 2.

Abdulaziz Alamri and Mark Stevenson. 2016. A corpus of potentially contradictory research claims from cardiovascular research abstracts. *J. Biomed. Semant.*, 7:36.

D. Altman. 1990. *Practical statistics for medical research*. Chapman and Hall/CRC.

D. Ayele and T. Zewotir. 2016. Comparison of under-five mortality for 2000, 2005 and 2011 surveys in ethiopia. *BMC Public Health*, 16.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

J. Beltman, C. Hendriks, M. Tum, and M. Schaap. 2013. The impact of large scale biomass production on ozone air pollution in europe. *Atmospheric Environment*, 71:352–363.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *J. Biomed. Informatics*, 43(2):173–189.

Claire Cardie, Nancy Green, Iryna Gurevych, Graeme Hirst, Diane Litman, Smaranda Muresan, Georgios Petasis, Manfred Stede, Marilyn Walker, and Janyce Wiebe, editors. 2015. *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA*. The Association for Computational Linguistics.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 621–631. Association for Computational Linguistics.

Domenic V Cicchetti and Sara A Sparrow. 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American journal of mental deficiency*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Nicolau Duran-Silva, Enric Fuster, Francesco Alessandro Massucci, and Arnau Quinquillà. 2019. A controlled vocabulary defining the semantic perimeter of sustainable development goals.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Theodosios Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news, blogs, and the social web. *Int. J. Artif. Intell. Tools*, 24(5):1540024:1–1540024:22.

Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker, editors. 2014. *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Comput. Linguistics*, 43(1):125–179.

Ivan Habernal, Iryna Gurevych, Kevin D. Ashley, Claire Cardie, Nancy Green, Diane J. Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern R. Walker, editors. 2017. *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*. Association for Computational Linguistics.

S. Hopewell, M. Clarke, D. Moher, E. Wager, P. Middleton, D. Altman, and K. Schulz. 2008. Consort for reporting randomized controlled trials in journal and conference abstracts: Explanation and elaboration. *PLoS Medicine*, 5.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with MACE. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 1120–1130. The Association for Computational Linguistics.

Johannes Kiesel, Khalid Al Khatib, Matthias Hagen, and Benno Stein. 2015. A shared task on argumentation mining in newspaper editorials. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA*, pages 35–38. The Association for Computational Linguistics.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA*, pages 1–11. The Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Anne Lauscher, Goran Glavas, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 40–46. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2).

Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using argument-based features to predict and analyse review helpfulness. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1358–1363. Association for Computational Linguistics.

Anastasios Lytos, Thomas Lagkas, Panagiotis G. Sarigiannidis, and Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *CoRR*, abs/1907.02258.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2018. Evidence type classification in randomized controlled trials. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 29–34. Association for Computational Linguistics.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.

Raquel Mochales and Aagje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments?: a case study on the legal argumentation of the ECHR. In *The 12th International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 8-12, 2009, Barcelona, Spain*, pages 21–30. ACM.

B. Ozcan and I. Ozturk. 2019. Renewable energy consumption-economic growth nexus in emerging countries: A bootstrap panel causality test. *Renewable & Sustainable Energy Reviews*, 104:30–37.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Informatics Nat. Intell.*, 7(1):1–31.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, and Hongfang Liu. 2016. Beltracker: evidence sentence retrieval for BEL statements. *Database J. Biol. Databases Curation*, 2016.

Chris Reed, editor. 2016. *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Darrel A Regier, William E Narrow, Diana E Clarke, Helena C Kraemer, S Janet Kuramoto, Emily A Kuhl, and David J Kupfer. 2013. Dsm-5 field trials in the united states and canada, part ii: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, 170(1):59–70.

Jaromír Savelka and Kevin D. Ashley. 2016. Extracting case law sentences for argumentation about the meaning of statutory terms. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Jodi Schneider. 2014. Automated argumentation mining to the rescue? envisioning argumentation and decision-making support for debates in open online collaboration communities. In *Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2014, June 26, 2014, Baltimore, Maryland, USA*, pages 59–63. The Association for Computer Linguistics.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinform.*, 24(18):2086–2093.

Noam Slonim and Ranit Aharonov, editors. 2018. *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*. Association for Computational Linguistics.

Luciana B Sollaci and M. G. Pereira. 2004. The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. *Journal of the Medical Library Association : JMLA*, 92 3:364–367.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1501–1510. ACL.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2019. *Argumentation mining*. Morgan & Claypool Publishers.

Benno Stein and Henning Wachsmuth, editors. 2019. *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 809–819. Association for Computational Linguistics.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5624–5634. Association for Computational Linguistics.

Stephen E Toulmin. 1958. *The uses of argument*. Cambridge university press.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 812–817. European Language Resources Association (ELRA).

Douglas Walton. 2015. *Goal-based Reasoning for Argumentation*. Cambridge University Press.

W. John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinform.*, 7:356.

Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2017. Annotation of argument structure in japanese legal documents. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 22–31. Association for Computational Linguistics.

An Yang and Sujian Li. 2018. Scidtb: Discourse dependency treebank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 444–449. Association for Computational Linguistics.

Antonio Zecca and Luca Chiari. 2012. Lower bounds to future sea-level rise. *Global and Planetary Change*, 98-99:1–5.