# Simple Signals for Detecting Major Claims in Scientific Articles: An Exploratory Study

Francis Lareau[1,2,3]

[1]Université de Sherbrooke, 2500, boulevard de l'Université Sherbrooke (Québec), J1K 2R1, Canada
[2]Université du Québec à Montréal, 1430, rue Saint-Denis, Montréal (Québec), H2X 3J8, Canada
[3]Centre interuniversitaire de recherche sur la science et la technologie, 1430, rue Saint-Denis, Montréal (Québec), H2X 3J8, Canada

**Abstract**

Automatic detection of major claims in scientific texts is a fundamental task in argument mining and in computational modelling of scientific discourse. This article presents an exploratory study that investigates how far simple semantic, positional, and discourse-based signals can go in prioritizing major-claim statements in the Discussion sections of medical articles. Starting from the AbstRCT corpus of annotated abstracts, we construct AbstRCT+, a derived corpus of 110 full-text randomized controlled trials whose Discussions contain 123 projected major claims obtained by aligning abstract-level annotations to Discussion sentences using string similarity. On this fixed corpus, we evaluate four heuristic models based on Doc2Vec-style vector similarity to the title and Introduction, sentence position, and discourse connectives from the Penn Discourse TreeBank v.3, together with a simple voting-based ensemble. Performance is assessed using top-$k$ F-measure for $k = 1$ to 5 under strong class imbalance. The best model, the ensemble, reaches an F-measure of 0.267 at $k = 1$, about 9.2 times higher than a random baseline on the same corpus, yet remains modest in absolute terms. We therefore interpret these results as evidence about the strength of simple signals in this specific setting rather than as a robust, generalizable detection system. Our contributions are: (i) the transparent construction of an aligned full-text corpus (AbstRCT+); (ii) an empirical characterization of where major claims tend to occur within Discussions; and (iii) a set of simple, well-documented baselines intended to ground future work with more sophisticated models.

**Keywords**

argument mining, major claim, scientific discourse, IMRAD, exploratory baselines, AbstRCT

## 1. Introduction

Identifying the main proposition defended by a scientific article — the *major claim* — is a central problem for argument mining in scholarly communication and for computational models of natural argument[1]. In IMRAD-structured articles (Introduction, Methods, Results, and Discussion), the Discussion section is the natural locus for such claims, since it is where results are interpreted and an answer to the research question is articulated. The task is complicated by pronounced class imbalance: in our data, the average Discussion contains around 42 sentences but typically only one major claim, and occasionally two or three. From a computational perspective, it is therefore natural to cast major-claim identification as a ranking task in which models propose a small set of candidate sentences that can subsequently be inspected or refined. The present work asks: to what extent can very simple, interpretable signals — semantic similarity to peripheral structures, position in the Discussion, and the presence of discourse connectives — support such a ranking task in a realistic but small medical corpus? Rather than emphasizing architectural sophistication, we focus on providing a transparent, corpus-specific baseline and a descriptive analysis of these signals. This aligns with work on argumentative and rhetorical zoning in scientific articles, which models how argumentative roles correspond to section structure and surface cues [2], as well as with the CMNA tradition of studying natural argumentation in concrete genres. We study four hypotheses: **H1.** Certain textual structures, especially the title (H1a) and the Introduction (H1b), maintain a relation of semantic similarity with the major claim. **H2.** Major-claim statements tend to occupy specific positions in the Discussion section. **H3.** Certain discourse connectives are more

*The 26th International Workshop on Computational Models of Natural Argument (CMNA'26)*
lareau.francis@courrier.uqam.ca (F. Lareau)
0000-0002-0352-5246 (F. Lareau)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

likely to appear in major-claim statements than in other sentences. **H4.** Semantic (H1), positional (H2), and discourse-based (H3) information sources are complementary, and their combination yields better detection performance than any single source. These hypotheses are examined using five lightweight models: title similarity, Introduction similarity, positional heuristics, discourse connectives, and a voting-based ensemble. The similarity models rely on a Doc2Vec/DBOW representation trained on the corpus. We neither compare against modern contextual embeddings or supervised classifiers nor perform held-out evaluation, so we present this work as an exploratory baseline study and corpus analysis that provides a starting point for future research with richer representations, larger corpora, and more sophisticated learning architectures. Our contributions are: (1) The construction and documentation of AbstRCT+, an automatically aligned full-text extension of the AbstRCT corpus for exploratory work on scientific argumentation; (2) A descriptive analysis of where major claims tend to occur in Discussion sections and of the relative strength of simple semantic and positional signals; (3) A set of transparent baseline models and empirical results that can ground future work with modern transformer-based architectures and richer argument-mining pipelines.

## 2. Background and related work

### 2.1. Argumentation in scientific texts

Our analysis draws on the communicational approach to argumentation, according to which argumentation arises from the interaction between a sender, primarily interested in persuasion, and a receiver, primarily interested in acquiring true information [3, 4]. This interaction produces a structure of reasons that support (arguments) and reasons that oppose (counterarguments) a central proposition, the major claim. In the case of randomized clinical trial abstracts in AbstRCT, we follow Mayer et al.'s guidelines, where major claims are defined as general or concluding claims that are supported by more specific claims or evidence [5]. In structured scientific texts, this argumentative hierarchy is reflected in the IMRAD macro-structure [6]. The Introduction raises the research question; the Methods describe how it will be addressed; the Results present the data; and the Discussion interprets the results to show how they do or do not answer the research question. Within this framework, the major claim is expected to appear predominantly in the Discussion, in sentences that formulate an answer to the Introduction's question. The title and Introduction also play a role: the title announces the central topic and broader problem addressed by the article, while the Introduction formulates the research question as a refinement of this topic. Together, they motivate H1a (title–major-claim similarity) and H1b (Introduction–major-claim similarity). Discourse relations, often signalled by explicit connectives, provide additional cues. The Penn Discourse TreeBank v.3 (PDTB-3) inventories discourse connectives organized into temporal, contingent, comparison, and expansion relations [7]. Connectives associated with contrast, concession, or conclusion may be particularly likely to appear in major-claim sentences.

### 2.2. Argument mining and scientific text analysis

Argument mining research has explored surface and structural features — lexical cues, discourse markers, positional information, and document structure — to detect argumentative components and relations [2, 8, 9]. Early work on scientific articles and technical genres already leveraged positional and section-based features, recognising that conclusions and central claims often occupy privileged locations [2, 10]. Subsequent research introduced increasingly sophisticated models, from feature-engineered classifiers to neural architectures and transformer-based representations, including applications to biomedical abstracts and clinical trial reports [11, 10]. More broadly, recent work has combined contextual embeddings with sequence-labelling and graph-based approaches to jointly identify argumentative components and relations [12, 13, 14, 15]. Against this backdrop, the present study deliberately does not attempt to compete with state-of-the-art transformer-based systems. Instead, it complements them by providing a transparent, interpretable analysis of how far simple, theory-motivated signals go on a small, well-defined corpus. The baselines we report can serve as reference points for future work

that applies modern embeddings or hybrid approaches combining simple structural cues with richer representations to the same data.

# 3. Corpus and methodology

## 3.1. The AbstRCT corpus

Our starting point is the AbstRCT corpus developed by Mayer, Cabrio and Villata [5], which consists of 669 abstracts of randomized controlled trials collected via PubMed and covering five medical domains. Each abstract is segmented into argumentative components and annotated with premises (evidence), subordinate claims, and major claims, linked by support or attack relations; inter-annotator agreement is satisfactory (Fleiss' $\kappa = 0.72$). In AbstRCT, a major claim expresses the conclusion of an inferential process: it is supported by one or more subordinate claims or premises but does not support any other claim.

## 3.2. Constructing the AbstRCT+ corpus

To study major claims in their rhetorical context, we construct a derived corpus, AbstRCT+, through the following steps. (1) *Selection.* We select the 110 abstracts in AbstRCT that contain at least one major-claim statement, out of 113 identified; three are excluded because full texts were not accessible under our constraints. (2) *Retrieval.* For each selected abstract, we retrieve the full text via PubMed using its PMID. (3) *Preprocessing.* Full texts are segmented into sentences, tokenized, lemmatized, and annotated with part-of-speech information using the Stanza NLP toolkit. (4) *Projection.* For each annotated major claim in the abstract, we identify the Discussion sentence that is most similar in surface form, using Levenshtein distance, and project the major-claim label onto that sentence. (5) *Vector representation.* Sentence lemmas are restricted to adverbs, adjectives, auxiliaries, verbs, nouns, pronouns, and proper nouns, and used to train a Distributed Bag of Words (DBOW) Doc2Vec model with 300-dimensional vectors in `gensim`. AbstRCT+ thus contains 110 articles whose Discussions include 123 projected major claims. Ninety-nine texts contain one projected major claim, nine contain two, and two contain three. Discussion sections range from 6 to 93 sentences, with an average of about 42 sentences per Discussion.

## 3.3. Limitations of the projection-based gold standard

A central limitation of this work is the automatic nature of the projection step. The assumption that each abstract-level major claim has a close lexical counterpart in the Discussion and that this counterpart can be captured by string similarity alone is strong. In practice, abstracts may paraphrase or compress the main conclusion, while Discussions may elaborate it across several sentences or express related claims that share large parts of their wording. To date, this projection has not been independently validated by human annotators. As a result, the gold-standard labels used in all subsequent analyses may contain mismatches. We therefore interpret all results as conditional on the quality of this projection and explicitly as approximate rather than definitive, and we view systematic human validation of the projected labels as a high-priority direction for future work. A second limitation is corpus size and domain specificity: AbstRCT+ contains 110 texts and 123 projected major claims, all in English and drawn from a small number of medical subfields. Our findings may not generalize to other scientific domains, article structures, or languages.

## 3.4. Evaluation as an in-sample ranking analysis

Major-claim detection in Discussions is a highly imbalanced task: Discussions contain many non-major sentences and at most a few major ones. We therefore emphasize its ranking aspect: given a Discussion, a model outputs a ranked list of candidate sentences, and success is measured by whether the projected major claim appears among the top $k$ candidates. We evaluate models using top-$k$ precision, recall, and

F-measure for $k = 1$ to 5. Let $N$ be the number of Discussions, and let $k$ be the number of candidates retained per Discussion. Top-$k$ precision is

$$\text{Precision}_{\text{top-}k} = \frac{\text{True positives}}{k \times N} \tag{1}$$

and top-$k$ recall is

$$\text{Recall}_{\text{top-}k} = \frac{\text{True positives}}{\text{Total major claims}}. \tag{2}$$

The top-$k$ F-measure is the harmonic mean of these two quantities. Because Discussions contain on average about 42 sentences but typically only one projected major claim, even an ideal model can achieve at most one true positive per Discussion at $k = 1$, so F-measure values are necessarily small in absolute terms. For comparison, we implement a random baseline that selects $k$ sentences uniformly at random from each Discussion (1000 iterations); in our corpus, this baseline reaches an F-measure of 0.029 for $k = 1$ and 0.046 for $k = 5$. Our models are purely heuristic and are analysed on the same AbstRCT+ corpus from which the positional patterns and discourse connectives are derived. Consequently, the reported numbers should be interpreted as corpus-internal descriptive indicators of the strength of these signals in this dataset, rather than as estimates of out-of-sample generalization performance.

## 4. Detection models

We now describe the five models examined in this study. All four base models are deliberately simple and heuristic, chosen to be interpretable and closely tied to the research hypotheses. The fifth model combines them via a voting-based ensemble.

### 4.1. Title-based model (M-Title)

Hypothesis H1a posits that the article title is semantically similar to the major claim, since both articulate the core contribution. M-Title encodes this idea by computing the cosine similarity between the DBOW vector of the title and the DBOW vector of each Discussion sentence. Formally, for each sentence $p_i$ in Discussion $D$,

$$\text{sim}(T, p_i) = \frac{\vec{T} \cdot \vec{p_i}}{\|\vec{T}\| \cdot \|\vec{p_i}\|}, \tag{3}$$

where $\vec{T}$ is the DBOW vector of the title and $\vec{p_i}$ is that of sentence $p_i$. The top $k$ sentences by similarity are retained as candidates.

### 4.2. Introduction-based model (M-Intro)

Hypothesis H1b suggests that the Introduction, by formulating the research question that the major claim answers, should also be semantically related to the major claim. M-Intro represents the Introduction by averaging the DBOW vectors of its sentences:

$$\vec{I} = \frac{1}{n} \sum_{j=1}^{n} \vec{p_j^I}, \tag{4}$$

where $\vec{p_j^I}$ is the vector of the $j$-th Introduction sentence and $n$ is the number of such sentences. Cosine similarity between $\vec{I}$ and each Discussion sentence is then computed as above, and the top $k$ sentences are retained.

### 4.3. Position-based model (M-Position)

Hypothesis H2 states that major claims occupy specific positions in the Discussion. An analysis of projected major-claim positions in AbstRCT+ shows that they tend to cluster near the beginning and even more at the end of Discussions. M-Position abstracts from all lexical content and relies only on sentence indices. Let $D = \{s_1, s_2, \dots, s_n\}$ be the ordered set of Discussion sentences. For a given $k$, the model returns

$$C_k = \{s_n, s_{n-1}, \dots, s_{n-\lceil k/2 \rceil+1}\} \cup \{s_1, s_2, \dots, s_{\lfloor k/2 \rfloor}\}. \tag{5}$$

For $k = 1$, the last sentence is selected; for $k = 2$, the first and last; for $k = 3$, the last two and the first, and so on. This model embodies a simple but rhetorically motivated positional heuristic.

### 4.4. Discourse-connective model (M-PDTB)

Hypothesis H3 holds that certain discourse connectives occur more frequently in major-claim sentences. Using the PDTB-3 inventory, we compute, for each connective $c_i$, an empirical precision on AbstRCT+:

$$\pi(c_i) = \frac{|\{s \in S^+ \,:\, c_i \in s\}|}{|\{s \in S \,:\, c_i \in s\}|}, \tag{6}$$

where $S^+$ is the set of Discussion sentences projected as major claims and $S$ the set of all Discussion sentences. For each $n$ between 1 and 25, we take $\mathscr{C}_n$ to be the $n$ connectives with highest $\pi(c)$ and define the candidate set

$$C_n = \{s \in D \,:\, \exists c \in \mathscr{C}_n,\, c \in s\}. \tag{7}$$

Unlike the previous models, M-PDTB does not produce a fixed number of candidates per Discussion; the number of candidates depends on how often the selected connectives appear. Consequently, M-PDTB is evaluated using standard F-measure rather than top-$k$ metrics. On AbstRCT+, the best M-PDTB variant achieves an F-measure of 0.113 when using nine connectives. Because connective selection and evaluation occur on the same corpus, these numbers should be interpreted as descriptive of this dataset rather than as an estimate of general performance.

### 4.5. Ensemble model (M-Ens)

Hypothesis H4 maintains that the semantic, positional, and discourse cues exploited by the base models are complementary. The ensemble model M-Ens assigns each Discussion sentence a vote count equal to the number of base models that selected it. Let $\mathscr{M} = \{M_{\text{Title}}, M_{\text{Intro}}, M_{\text{Position}}, M_{\text{PDTB}}\}$ be the set of base models, and $C^{(j)}$ the candidate set produced by model $M_j$. For each sentence $s_i$,

$$N(s_i) = \sum_{M_j \in \mathscr{M}} \mathbb{1}[s_i \in C^{(j)}], \tag{8}$$

where $\mathbb{1}[\cdot]$ is the indicator function. To reflect the strong performance of M-Intro, ties are broken in favour of sentences selected by it:

$$I(s_i) = \mathbb{1}[s_i \in C^{(\text{Intro})}]. \tag{9}$$

Sentences are then ranked lexicographically by the pair $(N(s_i), I(s_i))$, and the top $k$ sentences are returned:

$$C_k^{\text{Ens}} = \text{top-k}_{s_i \in D}(N(s_i), I(s_i)). \tag{10}$$

## 5. Results

Table 1 presents the main quantitative results on AbstRCT+. Several observations emerge from these results. First, all simple models outperform the random baseline, suggesting that the first three research hypotheses capture non-trivial signals for major-claim detection. The random model achieves a

**Table 1**
Top-$k$ F-measure (F1) and true positives (TP) for all models on the AbstRCT+ corpus (123 major claims).

| | $k$=1 | | $k$=2 | | $k$=3 | | $k$=4 | | $k$=5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | F1 | TP | F1 | TP | F1 | TP | F1 | TP | F1 | TP |
| Random | 0.029 | 3 | 0.039 | 7 | 0.042 | 10 | 0.044 | 12 | 0.046 | 15 |
| M-Position | 0.138 | 16 | 0.164 | 28 | 0.168 | 38 | 0.142 | 40 | 0.137 | 46 |
| M-Title | 0.181 | 21 | 0.216 | 37 | 0.199 | 45 | 0.181 | 51 | 0.164 | 55 |
| M-Intro | 0.241 | 28 | 0.263 | 45 | 0.226 | 51 | 0.203 | 57 | 0.190 | 64 |
| M-Ens | 0.267 | 31 | 0.269 | 46 | 0.248 | 56 | 0.221 | 62 | 0.199 | 67 |

maximum F-measure of 0.046, whereas the least effective simple model, M-PDTB, achieves an F-measure of 0.113, a gain of about a factor of 2.5. Second, M-Intro is the best-performing simple model, with an F-measure of 0.263 at $k = 2$, about 6.7 times the performance of the random baseline at the same $k$. This result is consistent with hypothesis H1b: the Introduction, as the structure that states the research question to which the major claim is the answer, appears to constitute the closest semantic anchor to the major claim. Third, M-Title obtains an F-measure of 0.216 at $k = 2$, slightly below M-Intro. This pattern is in line with H1a while also qualifying its relative importance: the title does provide a semantic signal, but this signal is more general than that of the Introduction and therefore less precise for identifying the major claim in the Discussion. Fourth, M-Position reaches its best performance at $k = 3$, with $F = 0.168$, which supports H2: major claims tend to cluster at the beginning and the end of Discussions. This result is noteworthy because it uses no semantic information at all, only position in the text, which suggests that rhetorical structure itself is a useful, albeit coarse, signal. Fifth, M-PDTB is the least effective simple model, with $F = 0.113$, although it still performs above chance. This suggests that some discourse connectives do act as major-claim signals, but that this approach has important limitations, including the fact that PDTB-3 connectives were developed on a journalistic corpus and may be suboptimal for medical scientific texts. Sixth, the ensemble model M-Ens, with an F-measure of 0.269 at $k = 2$, indicates that the four information sources are at least partly complementary. M-Ens correctly identifies 31 major claims out of 123 in its first prediction ($k = 1$), or 25%, and 67 out of 123 among its 5 best predictions ($k = 5$), or 55%. At $k = 2$, the ensemble recovers 46 major claims, compared with 45 for M-Intro, so the improvement over the best single model is modest and may partly reflect noise in the projection-based gold standard. We therefore interpret the ensemble's gains as suggestive rather than conclusive evidence of complementarity among the signals.

## 6. Conclusion

This paper reported an exploratory study of simple signals for detecting major claims in the Discussion sections of medical scientific articles, using a projected full-text extension of the AbstRCT corpus. We showed that semantic similarity to the Introduction and title, position within the Discussion, and the presence of certain discourse connectives all carry non-trivial information about the location of projected major claims, and that a simple voting-based ensemble can modestly improve performance over individual models on this corpus. At the same time, the absolute performance levels are low in practical terms, the models are heuristic, and both the gold standard and evaluation protocol are subject to significant limitations. We therefore present our results as corpus-specific baselines and descriptive evidence about the strength of simple signals, rather than as a competitive detection system. We see validation of the projection-based gold standard, benchmarking of transformer-based and generative models on the same task, and integration of major-claim detection with supporting-evidence and counterargument identification as natural next steps building on this work.

## Acknowledgments

This study was funded by the Canadian Social Sciences and Humanities Research Council (grant number 756-2024-0557). Parts of this work were developed within the framework of the author's PhD thesis.

## Declaration on Generative AI

During the preparation of this work, the author used Grammarly to check grammar and spelling and to improve prose. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

[1] M. Lippi, P. Torroni, Argumentation Mining: State of the Art and Emerging Trends, ACM Trans. Internet Technol. 16 (2016) 10:1–10:25. URL: https://doi.org/10.1145/2850417. doi:10.1145/2850417.

[2] S. Teufel, J. Carletta, M. Moens, An annotation scheme for discourse-level argumentation in research articles, in: Ninth Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Bergen, Norway, 1999, pp. 110–117. URL: https://aclanthology.org/E99-1015.

[3] H. Mercier, D. Sperber, Why do humans reason? Arguments for an argumentative theory, Behav Brain Sci 34 (2011) 57–74. URL: https://www.cambridge.org/core/product/identifier/S0140525X10000968/type/journal_article. doi:10.1017/S0140525X10000968.

[4] H. Mercier, D. Sperber, Argumentation: Its adaptiveness and efficacy, Behavioral and Brain Sciences 34 (2011) 94–111. URL: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/argumentation-its-adaptiveness-and-efficacy/EE3667EE6B26310F308B9AA840648835. doi:10.1017/S0140525X10003031.

[5] T. Mayer, E. Cabrio, M. Lippi, P. Torroni, S. Villata, Argument Mining on Clinical Trials, in: Computational Models of Argument, IOS Press, 2018, pp. 137–148. URL: https://ebooks.iospress.nl/doi/10.3233/978-1-61499-906-5-137. doi:10.3233/978-1-61499-906-5-137.

[6] L. B. Sollaci, M. G. Pereira, The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey, J Med Libr Assoc 92 (2004) 364–371. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC442179/.

[7] R. Prasad, B. Webber, A. Lee, A. Joshi, Penn Discourse Treebank Version 3.0, 2019. URL: https://catalog.ldc.upenn.edu/LDC2019T05. doi:10.35111/QEBF-GK47, artwork Size: 43056 KB Pages: 43056 KB.

[8] R. M. Palau, M.-F. Moens, Argumentation mining: the detection, classification and structure of arguments in text, in: Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL '09, Association for Computing Machinery, New York, NY, USA, 2009, pp. 98–107. URL: https://doi.org/10.1145/1568234.1568246. doi:10.1145/1568234.1568246.

[9] J. Lawrence, C. Reed, Argument Mining: A Survey, Computational Linguistics 45 (2019) 765–818. URL: https://aclanthology.org/J19-4006/. doi:10.1162/coli_a_00364.

[10] P. Accuosto, H. Saggion, Transferring Knowledge from Discourse to Arguments: A Case Study with Scientific Abstracts, in: B. Stein, H. Wachsmuth (Eds.), Proceedings of the 6th Workshop on Argument Mining, Association for Computational Linguistics, Florence, Italy, 2019, pp. 41–51. URL: https://aclanthology.org/W19-4505/. doi:10.18653/v1/W19-4505.

[11] T. Mayer, E. Cabrio, S. Villata, Transformer-based Argument Mining for Healthcare Applications, in: ECAI 2020 - 24th European Conference on Artificial Intelligence, Santiago de Compostela / Online, Spain, 2020. URL: https://hal.science/hal-02879293.

[12] S. Eger, J. Daxenberger, I. Gurevych, Neural End-to-End Learning for Computational Argumentation Mining, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the

Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 11–22. URL: https://aclanthology.org/P17-1002/. doi:10.18653/v1/P17-1002.

[13] P. Accuosto, H. Saggion, Mining arguments in scientific abstracts with discourse-level embeddings, Data & Knowledge Engineering 129 (2020) 101840. URL: https://www.sciencedirect.com/science/article/pii/S0169023X20300446. doi:10.1016/j.datak.2020.101840.

[14] M. Y. Abkenar, M. Stede, S. Oepen, Neural Argumentation Mining on Essays and Microtexts with Contextualized Word Embeddings (short paper), in: F. Benites, D. Tuggener, M. Hürlimann, M. Cieliebak, M. Vogel (Eds.), Proceedings of the Swiss Text Analytics Conference 2021, volume 2957 of *CEUR Workshop Proceedings*, CEUR, Winterthur, Switzerland, 2021. URL: https://ceur-ws.org/Vol-2957/#paper6.

[15] M. T. Sazid, R. E. Mercer, A Unified Representation and Deep Learning Architecture for Argumentation Mining of Students' Persuasive Essays, in: F. Grasso, N. L. Green, J. Schneider, S. Wells (Eds.), Proceedings of the 22nd Workshop on Computational Models of Natural Argument, volume 3205 of *CEUR Workshop Proceedings*, CEUR, Cardiff, Wales, 2022, pp. 86–97. URL: https://ceur-ws.org/Vol-3205/#paper10.