Datenbank Spektrum (2020) 20:115–121
https://doi.org/10.1007/s13222-020-00347-7

SCHWERPUNKTBEITRAG

# ArgumenText: Argument Classification and Clustering in a Generalized Search Scenario

**Johannes Daxenberger[1] · Benjamin Schiller[1] · Chris Stahlhut[1] · Erik Kaiser[1] · Iryna Gurevych[1]**

Received: 14 February 2020 / Accepted: 2 June 2020 / Published online: 16 June 2020
© The Author(s) 2020

**Abstract**
The ArgumenText project creates argument mining technology for big and heterogeneous data and aims to evaluate its use in real-world applications. The technology mines and clusters arguments from a variety of textual sources for a large range of topics and in multiple languages. Its main strength is its generalization to very different textual sources including web crawls, news data, or customer reviews. We validated the technology with a focus on supporting decisions in innovation management as well as customer feedback analysis. Along with its public argument search engine and API, ArgumenText has released multiple datasets for argument classification and clustering. This contribution outlines the major technology-related challenges and proposed solutions for the tasks of argument extraction from heterogeneous sources and argument clustering. It also lays out exemplary industry applications and remaining challenges.

**Keywords** Argument Mining · Argument Clustering

## 1 Introduction

Argument mining (AM) has become an established field of research in Natural Language Processing (NLP) with numerous works published over the last years [8, 12, 16]. AM is used with growing success to automatically detect argumentative structures in textual discourse, including student essays [8] and web forums [11]. Argumentative structures which can be automatically resolved include claims [6] and premises, argument relations [8], or pro- and con-arguments [22]. As such, AM can be used to support decision making by retrieving the most important arguments for and against controversial matters.

The current contribution details how we addressed the challenging task of argument search in heterogeneous data in the ArgumenText project.[1] ArgumenText has pioneered the generalization of AM at sentence level and created important resources for both argument classification and argument clustering. To achieve this goal, we had to overcome several research challenges:

(1) **Generalizing AM to heterogeneous sources** (e.g. news as well as web content): to extract relevant arguments from all sources available, we need to ensure that the AM model is able to detect arguments from any type of text.
(2) **Scaling AM technology to big data** (e.g. millions of web pages): to be able to work on large datasets or data streams, the extraction must be fast as in other information retrieval (IR) scenarios.
(3) **Clustering similar arguments** (if they refer to the same reasoning): to better present long lists of arguments to users, it is necessary to detect similar and dissimilar arguments.

In the following, we explain those challenges in more depth and show how they were solved in the ArgumenText project.

## 2 Argument-based Search

AM offers the perfect ground to combine machine learning with human decision making, as it is supposed to detect viewpoints (in the form of argumentative structures) using machine intelligence. Given a controversial topic (e.g. "wind energy") and a large enough text collection to search

[1] www.argumentext.de.

✉ Johannes Daxenberger
daxenberger@ukp.informatik.tu-darmstadt.de

[1] Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technische Universität Darmstadt, Darmstadt, Germany

**Table 1** Argumentative search engines. *Sources*: document collections from which the arguments are extracted; *Argument Classification*: argument detection at query (online) or indexing (offline) time.

| Reference | Name | Sources | Arg. Class. | Prototype |
|---|---|---|---|---|
| [26] | args.me | debate portals | offline | www.args.me |
| [21] | ArgumenText | generic web crawl | online | www.argumentsearch.com |
| [4] | PerspectroScope | curated online sources | offline and online | www.perspectroscope.com |
| [10] | IBM Debater | news articles, Wikipedia | online | not available |

in (e.g. a web crawl), the ideal AM system should be able to extract all relevant reasoning from previous debates about the topic of interest. For example, AM-supported decision making has been investigated in the context of evidence-based reasoning [19], where AM is used to detect and distinguish kinds of evidence with applications in the medical domain [14].

Given the subjective nature of evidence evaluation [1], initial applications of AM-supported decision making quickly converged into the creation of argumentative search engines [26]. Inspired by manually curated online debating portals such as kialo.com, procon.org or idebate.org, this line of research frames automatic AM as a retrieval task, aiming to maximize the relevance of search results with respect to the input query [17]. As opposed to standard web search engines, argumentative search engines need to detect the most relevant arguments given query term(s) and document collections to search in. Most approaches divide arguments into statements supporting (pro) or attacking (con) the input query, motivated by the goal to avoid biased or one-sided retrieval [2]. Arguments themselves are typically defined as "text expressing evidence or reasoning that can be used to either support or oppose a given topic" [22] – where the topic may be equal or highly relevant to the input query. The first case, in which arguments are classified as such based on the query itself, can be referred to as *online* argument classification [2]. In the latter case, arguments are classified regardless of the query (i.e. *offline*). Table 1 lists recently proposed argumentative search engines. We only list search engines dividing arguments by (binary) stance (pro vs. con). Further AM-driven search engines such as MARGOT [13] and TARGER [5] provide online interfaces for *argument tagging*, i.e. argument component detection on token-level [8]. Fig. 1 shows the ArgumenText search engine results for the query "wind energy".

The ArgumenText search engine was created as part of our effort to demonstrate the applicability of AM-driven approaches to decision-making, in particular to unrestricted and unstructured text collections. To date, ArgumenText is the only publicly available argumentative search engine retrieving English and German arguments in real-time from completely uncurated web sources (see Table 1). Recent work in the context of the IBM Debater project [10] presents a similar system extracting arguments from news and Wikipedia articles – however, they only release the Wikipedia portion of the dataset and do not offer a public search engine. The methodological details of our proposed solution to this problem are described in Sect. 3. For the project goal of testing the usefulness of AM technologies for real-world applications (see Sect. 1), we needed to go beyond argument search and develop end-user applications. This resulted in two additional requirements: the technology needed to be able to work on dynamic data streams (e.g. social media) and the clustering of recurring arguments (to reveal and quantify reasoning strategies for given topics). These two challenges are detailed in Sects. 4 and 5.

## 3 Extracting Arguments from Heterogeneous Sources

Early work on AM in NLP research used highly structured argumentation schemes to parse argumentative discourse [16, 20]. These argumentation schemes make rather strong assumptions on the argumentative nature of the input documents they can be applied to; e.g. the claim-premise scheme proposed by [20] relates *premises* (evidence) to *claims* which in turn refer to *major claims*. While it has been shown that such discourse-level approaches to AM can also be applied to web data [11], it remains doubtful whether they can be reliably applied to certain kinds of user-generated web content such as customer reviews [15]. Furthermore, for the purpose of training deep learning models, it is also necessary to collect large amounts of training data, which is much more difficult for fine-grained hierarchical schemes as the one proposed by [20]. We also found that often the major claim or even the claims themselves are not given explicitly, but must be inferred from the context or by using world knowledge. For example, an argument explicitly attacking coal energy could also serve as a supporting argument for wind energy implicitly.

As a remedy to this, [22] suggest *information-seeking* AM, which is "general enough for use on heterogeneous data sources, and simple enough to be applied manually by untrained annotators at a reasonable cost" [22]. The work shows that reliable annotation via crowdsourcing and automatic inference across eight topics is possible, when using a given controversial topic (e.g. "minimum wage") to

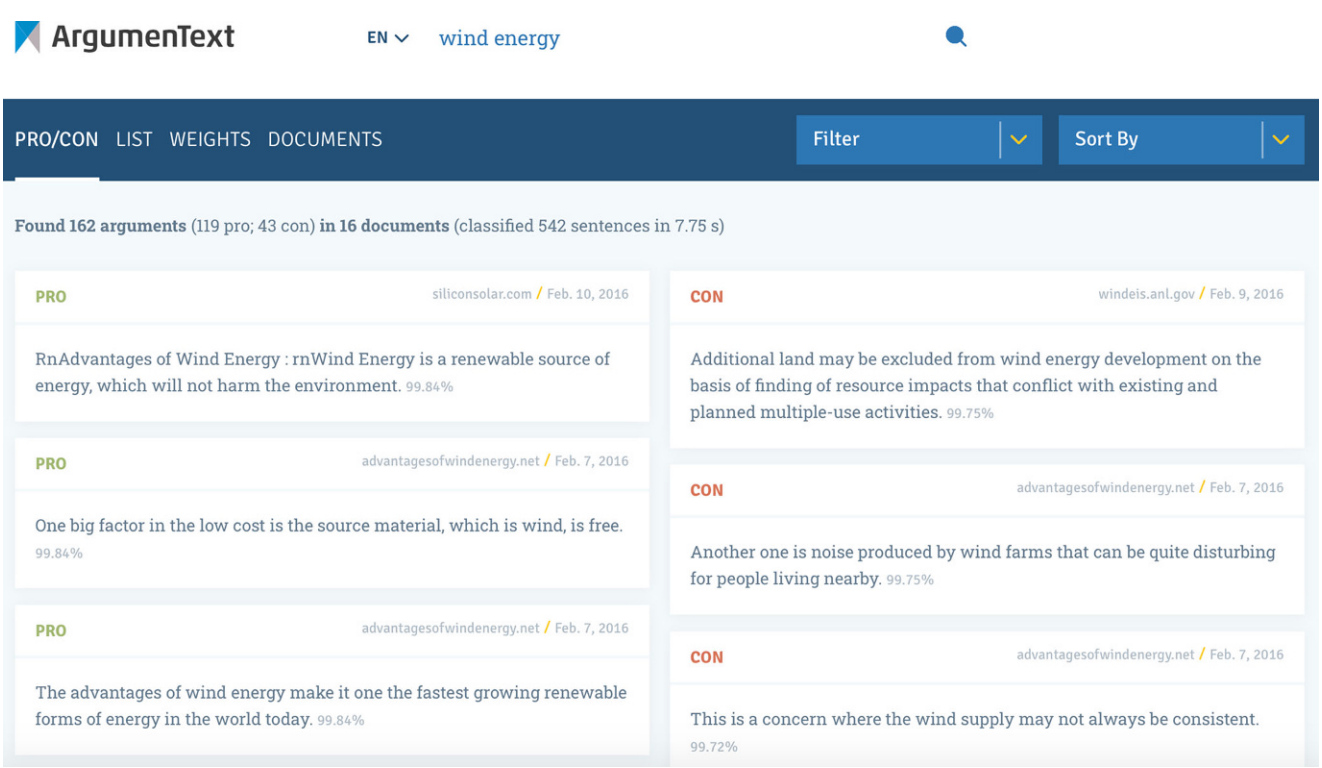


**Fig. 1** The first few hits for the search query "wind energy" as displayed by the argument search engine ArgumenText. ArgumenText ranks arguments by the confidence score of its argument extraction algorithm [21]

classify isolated sentences into either non-, pro-, or con-argument. The resulting dataset is released as part of the ArgumenText project.[2] Training and inference is performed by a Contextual BiLSTM architecture ("biclstm") which integrates the information about the topic into some of the LSTM gates, such that a sentence and topic can be processed jointly. Another advantage of the simpler annotation scheme is that the training data which was originally created on English sources can be translated into other languages using state-of-the-art machine translation (as exemplary shown for German by [23]). The translated data can then be used to directly train a model in the target language, which has been recognized as a very efficient way to create cross-lingual models for AM [9].

Our later work on argument classification [18] shows that the biclstm approach of [22] is largely outperformed by a transformer-based architecture using contextualized BERT-large embeddings [7]. In [21], we showed that when training on a larger set of topics, the performance of the sentence classification into non-, pro-, or con-argument can be further improved. We further showed that this kind of argument classification can also be performed on word level, allowing to decompose sentence-level arguments into more fine-grained units [24]. This approach requires token-level annotations for training a sequence labeling method, which we also release as part of the ArgumenText project.[3]

For the public version of the ArgumenText search engine, we indexed more than 400 million English and German web pages from the CommonCrawl project and segmented all documents into sentences [21]. For English and German queries, the system first retrieves a limited number of relevant documents ranked by a BM25 score, and second classifies all sentences from these documents with the above described classifier. Only arguments which have been identified as pro- or con-arguments are displayed and ranked by classifier confidence. Using this two-stage approach for argument search in heterogeneous sources, the ArgumenText system yields a coverage as high as 89% when comparing top-ranked search results to expert-curated lists [21].

## 4 Scaling AM to Big Data

The ArgumenText search engine described in Sect. 3 extracts arguments from a static web crawl. To be able to validate the technology beyond generic argument search, we built a service-oriented infrastructure around the core components. In particular, we wanted to be able to extract arguments from any given source, including arbitrary document

[2] https://www.ukp.tu-darmstadt.de/sent_am.

[3] https://github.com/trtm/AURC.

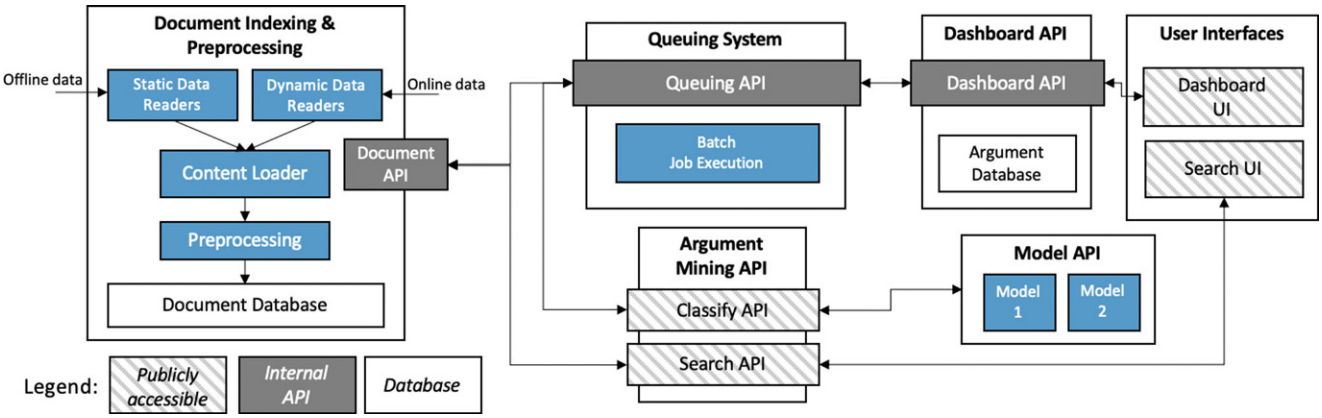


**Fig. 2** Overview of the ArgumenText service infrastructure. The document storage (*left*) can process and store content from static or dynamically growing document collections. The core components (middle) are responsible for argument processing and storage. Two graphical interfaces allow to interact with the system (*right*)

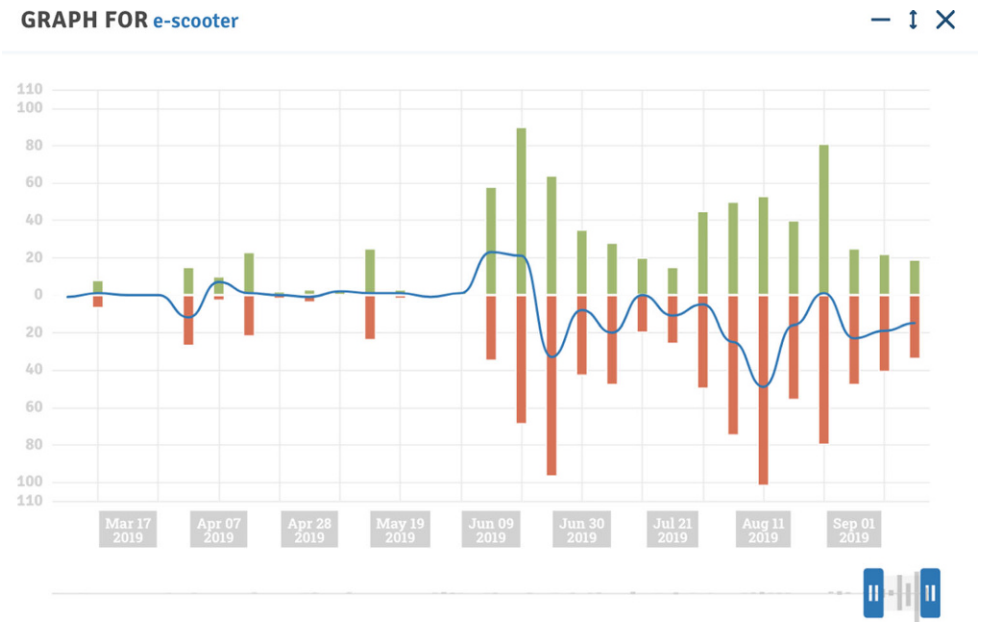


**Fig. 3** Excerpt from the ArgumenText dashboard. The argument graph for the topic "e-scooter" reveals an initial positive trend in June 2019, which turned negative in later months. Green and red bars indicate the number of pro and con arguments on the time axis

collections specified by end users. For that purpose, we decoupled argument classification from document retrieval and wrapped it as service available via REST APIs.[4] This service accepts arbitrary textual input and – given a topic which is used to decide on the argumentativeness of the sentences – returns sentence-level arguments from that input.

As direct queries to the REST APIs can only process a limited number of documents in order to prevent timeouts, we connected the argument classification API with a queuing functionality which handles query monitoring and execution in the background. The queuing component is connected to a graphical frontend which records search queries by registered users and pulls novel arguments periodically from the queue. The overall infrastructure is illustrated in Fig. 2. Fig. 3 shows the output of the graphical frontend for the query "e-scooter", as extracted from a web crawl.[5]

## 5 Argument Clustering

Arguments retrieved from multiple sources as in the above described scenarios often repeat similar reasoning. For example, on the topic of "nuclear energy", arguments referring to the problem of *radioactive waste* (an argumentative *aspect*) can be phrased in many ways. While it can be insightful to compare multiple instances of arguments from

[4] api.argumentsearch.com.

[5] Based on an in-house web crawl with timestamped web documents.


**Fig. 4** Word clouds and example arguments for three exemplary clusters for the topic "abortion". **a** "Fetuses are incapable of feeling pain when most abortions are performed." **b** "Abortion is the killing of a human being, which defies the word of God." **c** "Allowing abortion conflicts with the unalienable right to life recognized by the Founding Fathers of the United States."

the same argumentative aspect, smart AM decision-supporting systems should provide end-users with argument clusters rather than unsorted lists of arguments. Multiple lines of research have addressed this problem, including unsupervised learning of semantic similarities of arguments [3, 27].

However, as we have shown in [18], unsupervised methods are outperformed by supervised methods for the task of argument similarity assessment. Unsupervised learning methods rely on semantic overlap between pairs of arguments, which is not ideal for arguments that already discuss the same topic. Instead, we propose to train dedicated argument similarity models to provide similarity scores for the clustering approach. For this purpose, we released a corpus of sentence-level argument pairs extracted from heterogeneous web sources across 28 topics (ASPECT corpus).[6] The pairs were annotated on a range of three degrees of similarity, according to their overlap with regard to the argumentative aspect they address. Following the experiments described in [18], we only distinguish between related and unrelated arguments which enables to evaluate similarity prediction methods with F1 scores. The best supervised model (fine-tuned BERT-base) performs almost 10pp better than an unsupervised model based on BERT embeddings. Using agglomerative hierarchical clustering with stopping threshold, we are able to aggregate all arguments retrieved for a topic into clusters of aspects. Fig. 4 visualizes three example clusters that were produced using the above procedure.

## 6 Applications

We identified two promising applications for AM in supporting decisions: innovation assessment and advanced customer feedback analysis.

**Technology and Innovation Assessment:** Innovative technology often goes along with overly positive reasoning ("hype") at an early stage, such that it is difficult to identify potential risks. AM-based decision support can help this dilemma as it seeks to retrieve a balanced representation of supporting and attacking arguments on early or more mature innovative technologies. When applied to real-time news collections reporting about innovation and technology (e.g. online magazines), AM can help taking smarter investment decisions. Furthermore, novel trending aspects can be detected and quantified early on, using a combination of the technologies described in Sects. 4 and 5.

**Advanced Customer Feedback Analysis:** Companies with a broad product range in the consumer sector are often unable to accurately evaluate the large amount of customer feedback on different products and from multiple channels. Existing automatic methods to analyze the customer feedback rely on sentiment mining or unsupervised methods (clustering). While sentiment analysis might be able to separate positive from negative feedback or to distinguish degrees of criticality, it cannot reveal reasons behind the feedback which would be helpful for product development. Thus, the AM technologies as explained in Sects. 4 and 5 can be used to discover and quantify problematic aspects of existing products, to increase product-market-fit and decrease time-to-market.

[6] Available at https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/1998.

## 7 Future Directions

We presented challenges and solutions for AM-based decision support in the context of the ArgumenText project. Some remaining open challenges include:

(a) **Sorting arguments by quality**: Current argument search engines rank arguments by classifier confidence or by IR-based ranking functions. However, end users might prefer arguments of high quality [25] over arguments with high relevance to search query.
(b) **End-to-end argument clustering evaluation**: A large-scale benchmark dataset which contains sentence-level arguments for multiple topics and further groups them into subtopics is urgently required.
(c) **Labeling argument clusters**: Interpreting clusters is a difficult task which can be approximated by specifying predominant word lists (e.g. using LDA) or word frequency clouds. However, to clearly identify and label argument clusters, dedicated methodologies to extract aspect identifiers are required.

**Acknowledgements** This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 03VP02540 (ArgumenText).

**Funding** Open Access funding provided by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. Aharoni E, Polnarov A, Lavee T, Hershcovich D, Levy R, Rinott R, Gutfreund D, Slonim N (2014) A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In: ArgMining@ACL'14, pp 64–68. https://doi.org/10.3115/v1/W14-2109
2. Ajjour Y, Wachsmuth H, Kiesel J, Potthast M, Hagen M, Stein B (2019) Data acquisition for argument search: The args.me corpus. In: Benzmüller C, Stuckenschmidt H (eds) KI 2019: advances in artificial intelligence. Springer, Heidelberg, Berlin, New York, pp 48–59
3. Boltužić F, Šnajder J (2015) Identifying prominent arguments in online debates using semantic textual similarity. In: ArgMining@NAACL-HLT'15, pp 110–115. https://doi.org/10.3115/v1/W15-0514
4. Chen S, Khashabi D, Callison-Burch C, Roth D (2019) PerspectroScope: a window to the world of diverse perspectives. In: ACL'19: System Demonstrations, pp 129–134. https://doi.org/10.18653/v1/P19-3022
5. Chernodub A, Oliynyk O, Heidenreich P, Bondarenko A, Hagen M, Biemann C, Panchenko A (2019) TARGER: neural argument mining at your fingertips. In: ACL'19: system demonstrations, pp 195–200. https://doi.org/10.18653/v1/P19-3031
6. Daxenberger J, Eger S, Habernal I, Stab C, Gurevych I (2017) What is the essence of a claim? Cross-domain claim identification. In: EMNLP'17, pp 2055–2066
7. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL'19, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423
8. Eger S, Daxenberger J, Gurevych I (2017) Neural end-to-end learning for computational argumentation mining. In: ACL'17, pp 11–22
9. Eger S, Daxenberger J, Stab C, Gurevych I (2018) Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In: COLING'18, pp 831–844
10. Ein-Dor L, Shnarch E, Dankin L, Halfon A, Sznajder B, Gera A, Alzate C, Gleize M, Choshen L, Hou Y, Bilu Y, Aharonov R, Slonim N (2020) Corpus wide argument mining – a working solution. In: AAAI'20 (to appear)
11. Habernal I, Gurevych I (2017) Argumentation mining in user-generated web discourse. Comput Linguist 43(1):125–179. https://doi.org/10.1162/COLI_a_00276
12. Levy R, Bilu Y, Hershcovich D, Aharoni E, Slonim N (2014) Context dependent claim detection. In: COLING'14, pp 1489–1500
13. Lippi M, Torroni P (2016) MARGOT: a web server for argumentation mining. Expert Syst Appl 65:292–303. https://doi.org/10.1016/j.eswa.2016.08.050
14. Mayer T, Cabrio E, Villata S (2018) Evidence type classification in randomized controlled trials. In: ArgMining@EMNLP'18, pp 29–34. https://doi.org/10.18653/v1/W18-5204
15. Miller T, Sukhareva M, Gurevych I (2019) A streamlined method for sourcing discourse-level argumentation annotations from the crowd. In: NAACL'19, pp 1790–1796. https://doi.org/10.18653/v1/N19-1177
16. Mochales-Palau R, Moens MF (2009) Argumentation mining: the detection, classification and structure of arguments in text. In: ICAIL'09, pp 98–107
17. Potthast M, Gienapp L, Euchner F, Heilenkötter N, Weidmann N, Wachsmuth H, Stein B, Hagen M (2019) Argument search: assessing argument relevance. In: SIGIR'19, pp 1117–1120. https://doi.org/10.1145/3331184.3331327
18. Reimers N, Schiller B, Beck T, Daxenberger J, Stab C, Gurevych I (2019) Classification and clustering of arguments with contextualized word embeddings. In: ACL'19, pp 567–578. https://doi.org/10.18653/v1/P19-1054
19. Rinott R, Dankin L, Alzate Perez C, Khapra MM, Aharoni E, Slonim N (2015) Show me your evidence – an automatic method for context dependent evidence detection. In: EMNLP'15, pp 440–450. https://doi.org/10.18653/v1/D15-1050
20. Stab C, Gurevych I (2014) Annotating argument components and relations in persuasive essays. In: COLING'14, pp 1501–1510
21. Stab C, Daxenberger J, Stahlhut C, Miller T, Schiller B, Tauchmann C, Eger S, Gurevych I (2018) Argumentext: searching for arguments in heterogeneous sources. In: NAACL'18: system demonstrations, pp 21–25. https://doi.org/10.18653/v1/N18-5005
22. Stab C, Miller T, Schiller B, Rai P, Gurevych I (2018) Cross-topic argument mining from heterogeneous sources. In: EMNLP'18, pp 3664–3674
23. Stahlhut C (2018) Searching arguments in german with argumentext. In: DESIRES'18, vol 2167, p 104
24. Trautmann D, Daxenberger J, Stab C, Schütze H, Gurevych I (2020) Fine-grained argument unit recognition and classification. In: AAAI'20 (to appear)

25. Wachsmuth H, Naderi N, Hou Y, Bilu Y, Prabhakaran V, Thijm TA, Hirst G, Stein B (2017) Computational argumentation quality assessment in natural language. In: EACL'17, pp 176–187
26. Wachsmuth H, Potthast M, Al-Khatib K, Ajjour Y, Puschmann J, Qu J, Dorsch J, Morari V, Bevendorff J, Stein B (2017) Building an argument search engine for the web. In: ArgMining@EMNLP'17, pp 49–59. https://doi.org/10.18653/v1/W17-5106
27. Wachsmuth H, Syed S, Stein B (2018) Retrieval of the best counter-argument without prior topic knowledge. In: ACL'18, pp 241–251. https://doi.org/10.18653/v1/P18-1023