Overview of Touché 2026: Argumentation Systems

Extended Abstract

Johannes Kiesel[1], Marc Feger[2], Tim Hagen[3,7], Sebastian Heineking[4], Maximilian Heinrich[5], Maik Fröbe[6], Katarina Boland[2], Wilhelm Pertsch[6], Julia Romberg[1], Ines Zelch[4,6], Stefan Dietze[1,2], Matthias Hagen[6], Martin Potthast[3,7,8], and Benno Stein[5]

[1] GESIS - Leibniz Institute for the Social Sciences
[2] Heinrich Heine University Düsseldorf [3] University of Kassel [4] Leipzig University
[5] Bauhaus-Universität Weimar [6] Friedrich-Schiller-Universität Jena [7] hessian.AI
[8] ScaDS.AI

touche@webis.de touche.webis.de

Abstract What is an argument? Is an argument valid? Was a text manipulated to persuade? Since 2020, Touché fosters the development of support-technologies for decision-making and opinion-forming. To this end, the lab brings together researchers that develop systems to automatically answer questions like those above. At CLEF 2026 we do so in four tasks: (1) Fallacy Detection (new task), in which participants determine whether an argument follows a valid argument pattern; (2) Causality Extraction (new task), in which participants extract pro- and concausal claims from text; (3) Generalizability of Argument Identification in Context (new task), in which participants predict whether sentences would be annotated as an argument under different guidelines; and (4) Advertisement in Retrieval-Augmented Generation (2nd edition), in which participants detect and block advertisements in generated text. This paper details these tasks and summarizes the results of Touché 2025.

Keywords: Advertisement Detection · Argument Mining · Causality Detection · Fallacy Detection · Generalizability.

1 Introduction

Decision-making and opinion-forming are everyday task. With the ubiquity of web search and language models, it is easy to find arguments on any topic. However, while the systems of today are quick to produce an answer, they rarely provide transparent quality assurance of arguments and might even misuse their position as intermediary to manipulate information for commercial gains. In this context, the Touché lab series, running since 2020,[1] has organized several tasks to advance both argumentation systems and the evaluation thereof aimed to tackle current challenges. In 2026, we organize the following shared tasks:

[1] Previous tasks, data, and publications are available at https://touche.webis.de/

1. Fallacy Detection (new task) features three subtasks in the argumentation type detection, namely to classify (1) whether an argument is fallacious, (2) the argument scheme of an argument, and (3) the type of a fallacy.
2. Causality Extraction (new task) features three subtasks, in which participants (1) classify whether a text span contains causal information, (2) mark the respective entities in the span, and (3) classify the relation between two marked entities and procausal, concausal, or uncausal.
3. Generalizability of Argument Identification in Context (new task), in which participants classify whether a sentence, in its context and with provenance data, constitutes an argument or not.
4. Advertisement in Retrieval-Augmented Generation (2nd edition) features three subtasks in countering advertisements in the output of LLMs, namely (1) classify whether a response contains an advertisement or not, (2) locate the advertisement in a response, and (2) remove the advertisement from the response without affecting the coherence of the response.

After having organized six successful Touché labs on argument retrieval at CLEF 2020–2025 [2, 5, 4, 3, 18, 17], we now organize a seventh lab edition to bring together researchers from the fields of information retrieval, natural language processing, computational linguistics, and dialogue working on argumentation. During the previous labs, we received 384 runs from 106 teams. We manually labeled the relevance and quality of more than 35,000 argumentative texts, web documents, and images for 227 topics (topics and judgments are publicly available at the lab's web page, https://touche.webis.de). As in the previous Touché editions, we encourage participants to deploy their software in our cloud-based evaluation-as-a-service platform TIRA [12] for better reproducibility.

2 Task Definitions

Task 1: Fallacy Detection (new task) Argumentation is the process of presenting and evaluating reasons in support of, or in opposition to, a claim. Under ideal conditions, accepting an argument's premises should rationally warrant accepting its conclusion. In practice, however, this connection often breaks down: for example, the premises may fail to provide adequate support or the reasoning may be structurally defective. Detecting such fallacious reasoning is therefore essential for ensuring the reliability and trustworthiness of reasoning-based applications. In many cases, a fallacy can be understood as a defective, misleading, or misapplied instantiation of an otherwise legitimate argument scheme (a recurring pattern of reasoning [30, 20]). This task investigates whether integrating fallacy detection with argument scheme classification yields deeper theoretical insights and improves the performance of automated systems for both tasks.

Overview Given an argument, the task is structured as follows: (1) Determine whether the argument is fallacious. (2) If the argument is non-fallacious, identify its underlying argumentation scheme. (3) If the argument is fallacious, identify the specific type of fallacy it exhibits.

Data We use a dataset comprising over 1,000 argumentative examples drawn from multiple sources. Following the approach of Jin et al. [16], we include fallacious arguments collected from student quiz websites. In addition, we incorporate both valid arguments and fallacies from several online debate platforms. To further increase the diversity of argument forms and topics, we also include synthetically generated arguments covering a broad range of argument types. All data sources are used in accordance with their respective licensing terms, and the final dataset will be made freely available. The dataset focuses on the five most frequently occurring types of fallacies. For argument scheme detection, we follow the framework proposed by Macagno [20], which models argumentation along two complementary dimensions.

Example "One study found that a new diet helped 20 people lose weight. Therefore, this diet works for everyone." This is an example of a "Faulty Generalization," as it draws a broad conclusion from a small and unrepresentative sample.

Evaluation We evaluate each subtask using a held-out test set and report standard classification metrics, namely precision, recall, and F_1-score.

Task 2: Causality Extraction (new task) Many important questions in various domains are causal: diagnostics in medicine wonders about the root cause of symptoms or about the effect elicited by interacting drugs, and brokers are interested in how current events impact the market. Answering such questions requires a database of causal knowledge, which can be extracted from natural language text through causality extraction.

However, so far prior work has almost exclusively extracted causal claims and ignored their counterclaims, i.e., statements asserting *A does not cause B*. This is critical as it means that discourse on causal knowledge is not properly captured by current causality extraction systems, which may lead to faulty downstream application of extracted causal knowledge. To improve this, Hagen et al. [15] introduce the Countercausal News Corpus, the first resource to train and evaluate the extraction of both, causal statements and their counterclaims (*countercausal* statements). This task is the next step towards more complete causality extraction which can capture contradictory opinions about causality.

Overview This task is about the extraction of (counter-)causal claims from natural language text. Participating teams can submit software in any combination of the following three sub-tasks that build on each other: (1) Given a natural language text, classify whether it contains causal information or not; (2) given a natural language text, mark entities, events, or concepts for which the text claims or refutes a causal relationship; and (3) given a natural language text and two marked text spans, e_0 and e_1, classify whether the text supports that e_0 causes e_1 (causal), refutes its (countercausal) or does not make a statement about causality from e_0 to e_1 (uncausal).

Data For the dataset, we use the Countercausal News Corpus (CCNC) [15], a modified version of the Causal News Corpus v2 [28] in which some of the causal claims were manually rewritten to be countercausal. The published training and validation splits contain 3415 labeled sentences in total, out of which 1028 are causal and 952 are countercausal. The Causal News Corpus v2 and Countercausal News Corpus are licensed under CC BY and CC0 respectively.

Evaluation Submissions are evaluated on unpublished test data. Sub-tasks 1 and 3 are evaluated as binary and ternary classification problems, respectively, using F_1-score. Sub-task 2 is evaluated using F_1-score.

Task 3: Generalizability of Argument Identification in Context (new task) Argument identification is a fundamental prerequisite for discourse analysis across domains such as political debate, online discussion, and scientific reasoning. Pre-trained language models such as BERT [7], designed for contextualized language representation, have demonstrated state-of-the-art performance on established benchmarks. However, recent research suggests that state-of-the-art performance often stems from exploiting spurious correlations [29] and shortcut learning [13], as benchmarks rely on specialized datasets that encourage models to capture dataset-specific patterns shaped by topic bias, argument definitions, and labeling schemes rather than abstractions that generalize across contexts [10]. Yet arguments are defined not only by form or content, but also by their pragmatic function and contextualized use [9]. Just as humans rely on context to identify and interpret arguments in discourse, so must machines. This task therefore examines how contextual cues can support automated argument identification, focusing on the impact of different types and amounts of context on building more generalizable and task-aligned systems.

Overview Given a sentence from a dataset along with metadata about its provenance, such as the source text and the dataset's annotation guidelines, predict whether the sentence is annotated as an argument or not. In this cross-dataset setting, participants must develop robust systems that generalize beyond lexical shortcuts to unseen datasets and exploit rich context information.

Data A subset from 10 established, publicly available benchmark datasets [1, 6, 11, 14, 19, 21, 22, 23, 26, 27], identified as most relevant for argument identification [10], will be used. Each consists of 1.7k labeled sentences, partitioned with a 60/20/20 ratio into training, development, and test splits. Additionally, a new evaluation-only dataset will be released. Overall, the data includes sentences labeled as *argument* or *no-argument*, according to the respective dataset annotations. Accompanying metadata includes sentence IDs, generated splits, and (where available) context-relevant information via the original data sources, as well as annotation guidelines and corresponding papers.

Evaluation Systems will primarily be evaluated on the newly created, evaluation-only dataset. For further insights, evaluation results on the established test splits

from the held-out benchmark data will also be provided but not used for ranking. This setup addresses the risk of data contamination in LLMs and for participants' potential use of additional datasets during training. To evaluate the systems for their generalizability, the macro F_1-score will be measured for each test split, along with the overall average of all these values.

Task 4: Advertisement in Retrieval-Augmented Generation Opinion forming is a central part of argumentation and a process for which many people rely on search engines and increasingly also LLMs. Hence, it is important that the responses generated by LLMs, with or without retrieval-augmented generation (RAG), are not biased to influence their users. One way to introduce such bias would be through advertising that prompts an LLM to portray a given product, service, or brand in a favorable light [8, 25]. This type of advertising differs from existing approaches in that it enables highly contextual ads that blend into the surrounding text, thus requiring new types of ad blockers.

Overview This task aims for the detection and blocking of advertisements in generated text. Participants submit software in any combination of the following three subtasks: (1) Given a response and a query, classify whether the response contains an advertisement or not. (2) Given a response with advertisements and a query, predict the character spans of the ads. A response can contain multiple, interrupted spans of advertisements. (3) Given a response, a query, and a list of character spans that mark the advertisements in the response, remove the ads. After the removal, the response should still be fluent, factually correct (using the input response as reference), and relevant to the query.

Data For the development of submissions, we provide the Webis Generated Native Ads 2025 dataset.[2] The dataset contains 44,727 responses, with and without ads, that were generated by Brave Search, Microsoft Copilot, Perplexity, and YouChat for a total of 9,062 queries. To increase diversity, the advertisements were inserted by different LLMs: `GPT-4o` and `-mini` via OpenAI, as well as `deepseek-r1`, the 70B parameter versions of `llama-3` and `llama-3.3`, and `qwen-2.5-32b` via groq.[3] Each ad insertion was verified by a set of filters that check if the ad was correctly inserted and the response is otherwise identical to the input. For each response with an ad, the dataset contains the character spans of the advertisements, information on the item that was advertised, and the ID of the original response without advertisements.

Evaluation For the evaluation, we use an unpublished test split of the Webis Generated Native Ads 2025 dataset. Evaluation in subtask 1 uses standard F_1-score. Evaluation in subtask 2 uses an adapted F_1-score based on the overlap of

[2] https://zenodo.org/records/17830870
[3] https://groq.com/

predicted spans with ground-truth spans, drawing inspiration from the PlagDet-score developed for a similar purpose [24]. Evaluation in subtask 3 uses a combination of manual evaluation with LLM-as-a-judge through deepeval:[4] Human annotators evaluate fluency, correctness, and query relevance on a three-point scale from 0 to 2, while LLM-as-a-judge evaluation uses a scale from 0 to 1.

3 Touché at CLEF 2025: Brief Overview

In 2025, Touché at CLEF included these shared tasks [17]: (1) Retrieval-Augmented Debating, on simulating and evaluating deliberative debates; (2) Ideology and Power Identification in Parliamentary Debates (2nd iteration), including a new sub-task on populism identification; (3) Image Retrieval/Generation for Arguments (4th iteration), aiming to provide images that convey some claim; and (4) Advertisement in Retrieval-Augmented Generation (continues in 2026).

Touché 2025 received 62 registrations, of which 12 teams actively participated in the tasks and submitted 60 results (runs). Unsurprisingly, large language models (zero-shot, fine-tuned, etc.) were used across tasks, especially for retrieval-augmented debating. But also classic and more efficient approaches like SVMs were used for ideology and power identification. One team successfully submitted generated images for task 3, surpassing the still strong CLIP-baseline used in retrieval. For the Advertisement in Retrieval-Augmented Generation task, teams primarily used encoder models like MiniLM, MPNet, RoBERTa and DeBERTa-v3 for advertisement detection and Qwen and Mistral for generation. The corpora, topics, and judgments are available on the Touché website.[5]

4 Conclusion

At Touché, we continue to foster research on argumentation systems, building respective test collections, and bringing the research community together. During the previous six years, the submitted approaches developed from sparse to dense retrieval and zero-shot models (both for text and images), combined with assessments of document "argumentativeness," argument quality, stance detection, and sentiment analysis. Among others, argumentation systems can effectively contribute to generation systems, since in generative systems the task of reasoning (of which argumentation is the explication) is often a crucial but currently not sufficiently effective part of the system.

Touché 2026 brings in new tasks and refines existing ones. We continue our investigation into the detection of advertisements in generated search result text (injected subtle argumentation). Moreover, with fallacies and causality we investigate two exciting elements of argumentation. Furthermore, we aim to find robust argument identification systems—a tool that is still lacking today despite argument identification being the basis for many analyses.

[4] https://www.deepeval.com
[5] https://touche.webis.de/

Acknowledgements This work was partially supported by the European Commission under grant agreement GA 101070014 (https://openwebsearch.eu) and by the German Federal Ministry of Research, Technology and Space (BMFTR) through the project " DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B).

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Alhamzeh, A., Fonck, R., Versmée, E., Egyed-Zsigmond, E., Kosch, H., Brunie, L.: It's time to reason: Annotating argumentation structures in financial earnings calls: The FinArg dataset. In: Chen, C.C., Huang, H.H., Takamura, H., Chen, H.H. (eds.) Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pp. 163–169, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), https://doi.org/10.18653/v1/2022.finnlp-1.22, URL https://aclanthology.org/2022.finnlp-1.22/

[2] Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2020: Argument Retrieval. In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum (CLEF 2020), CEUR Workshop Proceedings, vol. 2696, CEUR-WS.org (2020), URL http://ceur-ws.org/Vol-2696/paper_261.pdf

[3] Bondarenko, A., Fröbe, M., Kiesel, J., Schlatt, F., Barriere, V., Ravenet, B., Hemamou, L., Luck, S., Reimer, J., Stein, B., Potthast, M., Hagen, M.: Overview of Touché 2023: Argument and Causal Retrieval. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) 14th International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science, vol. 14163, pp. 507–530, Springer, Berlin Heidelberg New York (Sep 2023), https://doi.org/10.1007/978-3-031-42448-9_31

[4] Bondarenko, A., Fröbe, M., Kiesel, J., Syed, S., Gurcke, T., Beloucif, M., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2022: Argument Retrieval. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (CLEF 2022), CEUR Workshop Proceedings, vol. 3180, pp. 2867–2903, CEUR-WS.org (2022), URL http://ceur-ws.org/Vol-3180/paper-247.pdf

[5] Bondarenko, A., Gienapp, L., Fröbe, M., Beloucif, M., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2021: Argument Retrieval. In: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum

(CLEF 2021), CEUR Workshop Proceedings, vol. 2936, pp. 2258–2284, CEUR-WS.org (2021), URL http://ceur-ws.org/Vol-2936/paper-205.pdf

[6] Cheng, L., Bing, L., He, R., Yu, Q., Zhang, Y., Si, L.: IAM: A comprehensive and large-scale dataset for integrated argument mining tasks. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2277–2287, Association for Computational Linguistics, Dublin, Ireland (May 2022), https://doi.org/10.18653/v1/2022.acl-long.162, URL https://aclanthology.org/2022.acl-long.162/

[7] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423/

[8] Dütting, P., Mirrokni, V., Paes Leme, R., Xu, H., Zuo, S.: Mechanism Design for Large Language Models. In: Proceedings of the ACM Web Conference 2024, pp. 144–155, ACM, Singapore Singapore (May 2024), https://doi.org/10.1145/3589334.3645511, URL https://dl.acm.org/doi/10.1145/3589334.3645511

[9] van Eemeren, F.H., Garssen, B., Krabbe, E.C.W., Snoeck Henkemans, A.F., Verheij, B., Wagemans, J.H.M.: Handbook of Argumentation Theory. Springer, Dordrecht, 1 edn. (Jul 2014), ISBN 978-90-481-9472-8, https://doi.org/10.1007/978-90-481-9473-5, URL https://doi.org/10.1007/978-90-481-9473-5, 61 b/w illustrations, 18 colour illustrations

[10] Feger, M., Boland, K., Dietze, S.: Limited generalizability in argument mining: State-of-the-art models learn datasets, not arguments (2025), URL https://arxiv.org/abs/2505.22137

[11] Fergadis, A., Pappas, D., Karamolegkou, A., Papageorgiou, H.: Argumentation mining in scientific literature for sustainable development. In: Al-Khatib, K., Hou, Y., Stede, M. (eds.) Proceedings of the 8th Workshop on Argument Mining, pp. 100–111, Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021), https://doi.org/10.18653/v1/2021.argmining-1.10, URL https://aclanthology.org/2021.argmining-1.10/

[12] Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., Hagen, M., Stein, B., Potthast, M.: Continuous Integration for Reproducible Shared Tasks with TIRA.io. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), pp. 236–241, Lecture Notes in

Computer Science, Springer, Berlin Heidelberg New York (Apr 2023), https://doi.org/10.1007/978-3-031-28241-6_20

[13] Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (11 2020), ISSN 2522-5839, https://doi.org/10.1038/s42256-020-00257-z, URL https://doi.org/10.1038/s42256-020-00257-z

[14] Haddadan, S., Cabrio, E., Villata, S.: Yes, we can! mining arguments in 50 years of US presidential campaign debates. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4684–4690, Association for Computational Linguistics, Florence, Italy (Jul 2019), https://doi.org/10.18653/v1/P19-1463, URL https://aclanthology.org/P19-1463/

[15] Hagen, T., Deckers, N., Wolter, F., Scells, H., Potthast, M.: Investigating Counterclaims in Causality Extraction from Text. CoRR **abs/2510.08224** (Oct 2025)

[16] Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., Schölkopf, B.: Logical fallacy detection. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pp. 7180–7198, Association for Computational Linguistics (2022), https://doi.org/10.18653/V1/2022.FINDINGS-EMNLP.532, URL https://doi.org/10.18653/v1/2022.findings-emnlp.532

[17] Kiesel, J., Çöltekin, Ç., Gohsen, M., Heineking, S., Heinrich, M., Fröbe, M., Hagen, T., Aliannejadi, M., Anand, S., Erjavec, T., Hagen, M., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Scells, H., Wolter, M., Zelch, I., Potthast, M., Stein, B.: Overview of Touché 2025: Argumentation Systems. In: de Albornoz, J.C., Gonzalo, J., Plaza, L., García Seco de Herrera, A., Mothe, J., Piroi, F., Rosso, P., Spina, D., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Sep 2025)

[18] Kiesel, J., Çöltekin, Ç., Heinrich, M., Fröbe, M., Alshomary, M., Longueville, B.D., Erjavec, T., Handke, N., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Reitis-Münstermann, T., Scharfbillig, M., Stefanovitch, N., Wachsmuth, H., Potthast, M., Stein, B.: Overview of Touché 2024: Argumentation Systems. In: Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Nunzio, G.M.D., Soulier, L., Galuscakova, P., Herrera, A.G.S., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Sep 2024)

[19] Lauscher, A., Glavaš, G., Ponzetto, S.P.: An argument-annotated corpus of scientific publications. In: Slonim, N., Aharonov, R. (eds.) Proceedings of the 5th Workshop on Argument Mining, pp. 40–46, Association for Computational Linguistics, Brussels, Belgium (Nov 2018), https://doi.org/10.18653/v1/W18-5206, URL https://aclanthology.org/W18-5206/

[20] Macagno, F.: Argumentation profiles and the manipulation of common ground. the arguments of populist leaders on twitter. Journal of Pragmatics **191**, 67–82 (2022), ISSN 0378-2166, https://doi.org/https://doi.org/10.1016/j.pragma.2022.01.022, URL https://www.sciencedirect.com/science/article/pii/S0378216622000285

[21] Mayer, T., Cabrio, E., Villata, S.: Transformer-based argument mining for healthcare applications. In: European Conference on Artificial Intelligence (2020), URL https://api.semanticscholar.org/CorpusID:221713735

[22] Misra, A., Ecker, B., Walker, M.: Measuring the similarity of sentential arguments in dialogue. In: Fernandez, R., Minker, W., Carenini, G., Higashinaka, R., Artstein, R., Gainer, A. (eds.) Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 276–287, Association for Computational Linguistics, Los Angeles (Sep 2016), https://doi.org/10.18653/v1/W16-3636, URL https://aclanthology.org/W16-3636/

[23] Panchenko, A., Bondarenko, A., Franzek, M., Hagen, M., Biemann, C.: Categorizing comparative sentences. In: Stein, B., Wachsmuth, H. (eds.) Proceedings of the 6th Workshop on Argument Mining, pp. 136–145, Association for Computational Linguistics, Florence, Italy (Aug 2019), https://doi.org/10.18653/v1/W19-4516, URL https://aclanthology.org/W19-4516/

[24] Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th International Competition on Plagiarism Detection. In: Forner, P., Navigli, R., Tufis, D. (eds.) Working Notes Papers of the CLEF 2013 Evaluation Labs, Lecture Notes in Computer Science, vol. 1179 (Sep 2013), ISBN 978-88-904810-3-1, ISSN 2038-4963, URL https://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-PotthastEt2013.pdf

[25] Schmidt, S., Zelch, I., Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Detecting Generated Native Ads in Conversational Search. In: Companion Proceedings of the ACM Web Conference 2024, p. 722–725, WWW '24, Association for Computing Machinery, New York, NY, USA (2024), https://doi.org/10.1145/3589335.3651489

[26] Stab, C., Gurevych, I.: Parsing argumentation structures in persuasive essays. Computational Linguistics **43**(3), 619–659 (Sep 2017), https://doi.org/10.1162/COLI_a_00295, URL https://aclanthology.org/J17-3005/

[27] Swanson, R., Ecker, B., Walker, M.: Argument mining: Extracting arguments from online dialogue. In: Koller, A., Skantze, G., Jurcicek, F., Araki, M., Rose, C.P. (eds.) Proceedings of the 16th Annual Meeting of

the Special Interest Group on Discourse and Dialogue, pp. 217–226, Association for Computational Linguistics, Prague, Czech Republic (Sep 2015), https://doi.org/10.18653/v1/W15-4631, URL https://aclanthology.org/W15-4631/

[28] Tan, F.A., Hettiarachchi, H., Hürriyetoglu, A., Oostdijk, N., Caselli, T., Nomoto, T., Uca, O., Liza, F.F., Ng, S.K.: RECESS: Resource for Extracting Cause, Effect, and Signal Spans. In: Park, J.C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., Krisnadhi, A.A. (eds.) Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023, pp. 66–82, Association for Computational Linguistics (2023), https://doi.org/10.18653/V1/2023.IJCNLP-MAIN.6

[29] Thorn Jakobsen, T.S., Barrett, M., Søgaard, A.: Spurious correlations in cross-topic argument mining. In: Ku, L.W., Nastase, V., Vulić, I. (eds.) Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, pp. 263–277, Association for Computational Linguistics, Online (Aug 2021), https://doi.org/10.18653/v1/2021.starsem-1.25, URL https://aclanthology.org/2021.starsem-1.25/

[30] Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press (2008), ISBN 9780521723749, URL http://www.cambridge.org/us/academic/subjects/philosophy/logic/argumentation-schemes