Overview of Touché 2025: Argumentation Systems

Extended Abstract

Johannes Kiesel[1], Çağrı Çöltekin[2], Marcel Gohsen[1], Sebastian Heineking[3], Maximilian Heinrich[1], Maik Fröbe[4], Tim Hagen[5,10], Mohammad Aliannejadi[6], Tomaž Erjavec[7], Matthias Hagen[4], Matyáš Kopp[8], Nikola Ljubešić[7], Katja Meden[7], Nailia Mirzakhmedova[1], Vaidas Morkevičius[9], Harrisen Scells[3], Ines Zelch[4], Martin Potthast[5,10,11], and Benno Stein[1]

[1] Bauhaus-Universität Weimar [2] University of Tübingen [3] Leipzig University [4] Friedrich-Schiller-Universität Jena [5] University of Kassel [6] University of Amsterdam [7] Jožef Stefan Institute [8] Charles University [9] Kaunas University of Technology [10] hessian.AI [11] ScaDS.AI

touche@webis.de touche.webis.de

Abstract Decision-making and opinion-forming are everyday tasks that involve weighing pro and con arguments. The goal of Touché is to foster the development of support-technologies for decision-making and opinion-forming and to improve our understanding of these processes. This sixth edition of the lab features four shared tasks, three of which include generative systems: (1) Retrieval-Augmented Debating (RAD), in which participants submit generative retrieval systems that argue against their users and evaluate such systems (new task); (2) Ideology and Power Identification in Parliamentary Debates, in which participants identify from a speech the political leaning of the speaker's party and whether it was governing at the time of the speech (2nd edition); (3) Image Retrieval/Generation for Arguments, in which participants find images to convey a written argument (4th edition, joint task with ImageCLEF); and (4) Advertisement in Retrieval-Augmented Generation, in which participants generate responses to queries with ads inserted and detect such inserted ads (new task). In this paper, we briefly describe the planned setup for the sixth lab edition at CLEF 2025 and summarize the results of the 2024 edition.

Keywords: Advertisement Detection · Argumentation · Ideology Identification · Image Generation · Image Retrieval · Retrieval-Augmented Generation · User Simulation.

1 Introduction

Decision-making and opinion-forming are everyday tasks that involve weighing pro and con arguments for or against different options. With ubiquitous access

to all kinds of information on the web, everybody has the chance to acquire knowledge for these tasks on almost any topic. However, current information systems are primarily optimized for returning *relevant* results and do not address deeper analyses of arguments or multi-modality. To close this gap, the Touché lab series, running since 2020, has several tasks to advance both argumentation systems and the evaluation thereof. Previous events and tasks, data, and publications are available at https://touche.webis.de/. In 2025, we organize the following shared tasks:

1. Retrieval-Augmented Debating (RAD; new task) features two subtasks in argumentative agent research of (1) generating responses to argue against a simulated debate partner and (2) evaluating systems of sub-task 1.
2. Ideology and Power Identification in Parliamentary Debates (2nd edition) features two subtasks in debate analysis of detecting the (1) ideology and (2) position of power of the speaker's party, respectively.
3. Image Retrieval/Generation for Arguments (4th edition; joint task with ImageCLEF [10]) is about finding images to help convey an argument.
4. Advertisement in Retrieval-Augmented Generation (new task) features two subtasks in retrieval-augmented generation of (1) generating responses with advertisements inserted and (2) detecting whether a response contains an advertisement.

After having organized five successful Touché labs on argument retrieval at CLEF 2020–2024 [1, 4, 3, 2, 13], we now organize a sixth 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 Touché labs, we received 324 runs from 94 teams. We manually labeled the relevance and quality of more than 35,000 argumentative texts, web documents, and images for 200 search topics (topics and judgments are publicly available at the lab's web page, https://touche.webis.de).

This year's edition of Touché again intends to widen its scope. After having explored ethical questions in last year's edition, our two new tasks explore retrieval-augmented generation. The first new task, Retrieval-Augmented Debating, investigates retrieval-augmented generation for argumentative discussions. Participating systems will debate with simulated users over up to five turns, and their performance will be evaluated based on the quality of their responses. Moreover, participants can develop and submit automated measures to mirror human judgments of the system performance. The second new task, Advertisement in Retrieval-Augmented Generation, looks at advertisements as a monetization option for generative retrieval systems. Participants submit systems to either insert advertisements unobtrusively into generated text or to detect such insertions. Moreover, two of our tasks, Ideology and Power Identification in Parliamentary Debates and Image Retrieval/Generation for Arguments continue in 2025 with a refined setup. As in the previous Touché editions, we will encourage participants to deploy their software in our cloud-based evaluation-as-a-service platform TIRA [7] for better reproducibility.

2 Task Definitions

Task 1: Retrieval-Augmented Debating (RAD) Engaging in conversational argumentation enhances individuals' argumentation skills, which can also improve their performance in non-conversational contexts, such as writing argumentative essays [11]. In these dialogues, participants either defend their own positions or challenge their opponents' arguments. With the progress in the conversational capabilities of large language models (LLMs), we can now develop automated argumentation systems. These systems can be used to improve personal argumentation skills or to help individuals form or validate their opinions.

Overview The goal of this task is to create generative retrieval systems that engage in argumentative conversations by presenting counterarguments to users' claims. Participating systems will debate with simulated users over up to five turns, and their performance will be evaluated based on the quality of their responses. The conversation begins with a simulated user making a claim, to which the system must respond with counterarguments. Teams can participate in two sub-tasks: (1) developing debate systems, and (2) providing metrics to assess various quality criteria based on Grice's axioms of a cooperative dialog [8], specifically on the quantity (length), quality (faithfulness), relevance (cf. argumentative quality [17]), and manner (clarity) of system responses.

Data Participants will work with a dataset of about 300 000 arguments extracted from around 1 500 debates within the ClaimRev dataset [16]. Additionally, they will receive 100 curated claims sourced from the ChangeMyView subreddit. These claims will be used to simulate 100 debates between baseline systems and simulated users, which will then be evaluated by experts based on the quality criteria outlined in sub-task 2.

Evaluation Submissions for sub-task 1 will be evaluated using a new set of initial claims and various simulated users, each presenting different argument strategies, resulting in one simulated debate for each combination of claim, user, and system. All debates will be assessed using the evaluation systems submitted for sub-task 2 and our baseline metrics. A random subset of the debates will be judged by human experts according to the criteria of sub-task 2 to identify for each criterion the evaluation system that aligns best with human judgment. The respective evaluation systems will then be used to assess the debate systems from sub-task 1. To support participants in testing their systems, we provide metrics, baselines, and user simulators through GenIRSim [14].[1]

Task 2: Multilingual Ideology and Power Identification in Parliamentary Debates Parliaments are one of the most important institutions in modern democratic states where issues with high societal impact are discussed. The impact of the decisions made in a parliament often goes beyond their borders, and

[1] Demo: https://genirsim.webis.de/

may even have global effects. As a form of political debate, however, speeches in a parliament are often indirect and present challenges for automated systems for analyzing them.

Overview This task is concerned with predicting *ideology* and *power* in (transcribed) parliamentary speeches from multiple national parliaments, recorded in multiple languages. Both subtasks are formulated as binary classification tasks. Participating teams can submit software in one or both of two sub-tasks: (1) predicting the political orientation (`left`–`right`) of speakers from their speeches; and (2) predicting whether the speaker is a member of a governing party or the opposition.

Data The data for both tasks is a sample of ParlaMint [6], a corpus of parliamentary speeches from 29 national or regional parliaments with varying amounts of instances. The time span of the data is from 2015 to 2022 across all parliaments. To ease participation and balance the dataset, this task uses a sample of ParlaMint (full data is up to 90 million words per parliament). The dataset for both tasks includes at least speeches from national parliaments of Belgium, Iceland, Italy, Poland, Slovenia, Spain, The Netherlands, Turkey and United Kingdom. ParlaMint contains machine translation of all data to English, which participants can use as supporting data.

Evaluation Submissions are evaluated using macro F_1-score in both subtasks, for all languages. Even though the participants are encouraged to make use of multilingual data for improving results for individual languages, we do not evaluate zero- or few-shot settings separately.

Task 3: Image Retrieval/Generation for Arguments (joint task with ImageCLEF) Argumentation is a communicative activity in which reasons are exchanged. In addition to words, images are often used in argumentation, either to illustrate, to exemplify or to arouse emotions.

Overview Given a set of arguments, the task is to return for each argument several images that help convey the argument. A suitable image could depict the argument or show a generalization or specialization. Participants can optionally add a short caption that explains the meaning of the image. Images can be either retrieved from the focused crawl or generated using an image generator. Figure 1 shows two example images.

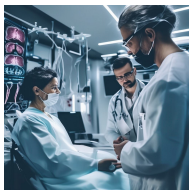
Data The task data consists of 200 arguments. As document collection we provide a focused crawl of at least 1000 images per argument [9]. Following the idea of the infinite index [5], we also provide an API for an image generator. The human judgments from previous years are available and can be used to train new approaches.

Figure 1. Two example submissions for the topic "AI in medicine" with the argument: "AI helps doctors with complicated operations."

Submission 1

Source: Retrieved

Relevance: 2
(shows AI, doctors and operation)



Submission 2

Source: Generated

Relevance: 0
(not topic-related, shows only doctors)

Evaluation The task follows the classic TREC-style methodology of teams submitting ranked results to be judged by human assessors. As a metric, nDCG [12] is used to represent a user looking through a ranked list of images retrieved for a specific argument.

Task 4: Advertisement in Retrieval-Augmented Generation The use of large language models (LLMs) in conversational search engines enables advertisements to be directly included into the generated responses as a form of native advertisement. This scenario has the potential of exerting an even greater influence on people than traditional advertisements as they can be more difficult to identify. Furthermore, the integration of ads in responses requires the development of a new kind of ad blocker.

Overview The goal of this task is to detect and hide advertisements in a generated text. Participating teams submit software in one or both of two sub-tasks. (1) Create relevant responses for a given query, based on a set of document segments. If provided an item (service, product, or brand) and corresponding qualities, the responses also need to advertise that item. This advertisement should be difficult to detect and fit seamlessly into the rest of the response. (2) Classify whether a given response contains an advertisement or not.

Data For the development of submissions, we provide the Webis Generated Native Ads 2024 dataset.[2] The dataset contains 4,868 queries, suitable items to be advertised, as well as 17,344 responses generated by Microsoft Copilot and YouChat. Into a third of the responses, we inserted advertisements with GPT-4. As context for the response generation, we provide up to 100 document segments for selected queries. These segments are retrieved from the segmented version of the MS MARCO v2.1 document corpus used for TREC 2024 RAG.[3]

Evaluation For the evaluation, we will create a new version of the Webis Generated Native Ads dataset. In subtask (1), we train a classifier for the detection of ads as we found this to perform sufficiently well for evaluation [15].

[2] https://zenodo.org/records/10802427
[3] https://trec-rag.github.io/about/

For each submitted system, we calculate the precision, recall, and F1 score. The primary score of a submission is the inverse F1 score - the lower the F1 score, the more difficult to detect are the inserted ads. We also report precision and recall as low recall values indicate subtly included ads, while low precision values suggest that responses without an advertisement also have an ad-like character (which should be avoided). For subtask (2), we will provide submitted systems with responses - both with and without inserted ads. Each submission will be scored based on its F1 score on the binary classification task. Again, we will also report precision and recall.

3 Touché at CLEF 2024: Brief Overview

In 2024, Touché at CLEF included the following three shared tasks [13]: (1) Human Value Detection (ValueEval'24), on detecting human values in texts and their attainment, respectively; (2) Ideology and Power Identification in Parliamentary Debates (continues in 2025); (3) Image Retrieval/Generation for Arguments (continues in 2025).

Touché 2024 received 68 registrations from 22 countries, of which 20 teams actively participated in the tasks and submitted 81 results (runs). Participants mainly used classification architectures, with BERT and variants still very dominant, although more classical machine learning models were also used in the Ideology and Power Identification in Parliamentary Debates task. Generative models, on the other hand, were rarely used. Approaches for the Image Retrieval/Generation task mainly used images and text embeddings for similarity search. The corpora, topics, and judgments are available on the Touché website.[4]

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 five years, the submitted approaches developed from sparse to dense retrieval and zero-shot models, combined with assessments of document "argumentativeness," argument quality, stance detection, and sentiment analysis.

Touché 2025 brings in new tasks and refines existing ones. We continue our investigation into the argumentation within political debates. Moreover, generative systems are now part of three of our tasks, exploring conversational aspects, the use of images, and advertisement insertion and detection. 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.

Acknowledgements This work was partially supported by the European Commission under grant agreement GA 101070014 (https://openwebsearch.eu)

[4] https://touche.webis.de/

Bibliography

[1] 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

[2] 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

[3] 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

[4] 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

[5] Deckers, N., Fröbe, M., Kiesel, J., Pandolfo, G., Schröder, C., Stein, B., Potthast, M.: The Infinite Index: Information Retrieval on Generative Text-To-Image Models. In: Gwizdka, J., Rieh, S.Y. (eds.) ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023), pp. 172–186, ACM (Mar 2023), https://doi.org/10.1145/3576840.3578327

[6] Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Calzada Pérez, M., de Macedo, L.D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darǵis, R., Ring, O., van Heusden, R., Marx, M., Fišer, D.: The parlamint corpora of parliamentary proceedings. Language resources and evaluation **57**, 415–448 (2022), https://doi.org/10.1007/s10579-021-09574-0

[7] 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

[8] Grice, H.: Studies in the Way of Words. William James lectures, Harvard University Press (1989), ISBN 9780674852716

[9] Heinrich, M., Kiesel, J., Wolter, M., Potthast, M., Stein, B.: Touché25-Image-Retrieval-and-Generation-for-Arguments (Dec 2024), https://doi.org/10.5281/zenodo.14258397

[10] Ionescu, B., Müller, H., Stanciu, D.C., Idrissi-Yaghir, A., Radzhabov, A., de Herrera, A.G.S., Andrei, A., Storås, A., Abacha, A.B., Bracke, B., Lecouteux, B., Stein, B., Macaire, C., Friedrich, C.M., Schmidt, C.S., Fabre, D., Schwab, D., Dimitrov, D., Esperança-Rodier, E., Constantin, G., Becker, H., Damm, H., Schäfer, H., Rodkin, I., Koychev, I., Kiesel, J., Rückert, J., Malvehy, J., Ştefan, L.D., Bloch, L., Potthast, M., Heinrich, M., Riegler, M.A., Dogariu, M., Codella, N., Halvorsen, P., Nakov, P., Brüngel, R., Novoa, R.A., Das, R.J., Hicks, S.A., Gautam, S., Pakull, T.M.G., Thambawita, V., Kovalev, V., Yim, W.W., Xie, Z.: ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications. In: Hauff, C., Macdonald, C., Jannach, D., Kazai, G., Nardini, F.M., Silvestri, F., Tonellotto, N. (eds.) Advances in Information Retrieval. 47th European Conference on IR Research (ECIR 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Apr 2025), https://doi.org/10.1007/978-3-031-56072-9_6

[11] Iordanou, K., Rapanta, C.: "Argue With Me": A Method for Developing Argument Skills. Frontiers in Psychology **12** (2021), ISSN 1664-1078, https://doi.org/10.3389/fpsyg.2021.631203

[12] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4), 422–446 (2002), https://doi.org/10.1145/582415.582418

[13] 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)

[14] Kiesel, J., Gohsen, M., Mirzakhmedova, N., Hagen, M., Stein, B.: Who Will Evaluate the Evaluators? Exploring the Gen-IR User Simulation Space. In: Proceedings of CLEF 2024, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Sep 2024)

[15] 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

[16] Skitalinskaya, G., Klaff, J., Wachsmuth, H.: Learning from revisions: Quality assessment of claims in argumentation at scale. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, pp. 1718–1729, Association for Computational Linguistics (2021), https://doi.org/10.18653/V1/2021.EACL-MAIN.147

[17] Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: Computational Argumentation Quality Assessment in Natural Language. In: Proceedings of EACL 2017, pp. 176–187 (Apr 2017), URL https://aclanthology.org/E17-1017/