Overview of Touché 2023: Argument and Causal Retrieval

Extended Abstract

Alexander Bondarenko,[1] Maik Fröbe,[1] Johannes Kiesel,[2] Ferdinand Schlatt,[3] Valentin Barriere,[4] Brian Ravenet,[5] Léo Hemamou,[6,*] Simon Luck,[7] Jan Heinrich Reimer,[3] Benno Stein,[2] Martin Potthast,[8] and Matthias Hagen[1]

[1] Friedrich-Schiller-Universität Jena [2] Bauhaus-Universität Weimar
[3] Martin-Luther-Universität Halle-Wittenberg
[4] Centro Nacional de Inteligencia Artificial (CENIA) [5] Université Paris-Saclay
[6] Sanofi R&D France [7] Alma Mater Studiorum - Università di Bologna
[8] Leipzig University and ScaDS.AI

touche@webis.de touche.webis.de

Abstract The goal of Touché is to foster and support the development of technologies for argument and causal retrieval and analysis. For the fourth time, we organize the Touché lab featuring four shared tasks: (a) argument retrieval for controversial topics, where participants retrieve web documents that contain high-quality argumentation and detect the argument stance, (b) causal retrieval, where participants retrieve documents that contain causal statements from a generic web crawl and detect the causal stance, (c) image retrieval for arguments, where participants retrieve images showing support or opposition to some stance from a focused web crawl, and (d) intra-multilingual multi-target stance classification, where participants detect the stance of comments on proposals from the multilingual participatory democracy platform CoFE. In this paper, we briefly summarize the results of Touché 2022 and describe the planned setup for the fourth lab edition at CLEF 2023.

1 Introduction

Making informed decisions and forming opinions on a matter often involves not only weighing pro and con arguments towards different options but also considering cause-effect relationships for one's actions [1]. Nowadays, everybody has the chance to acquire knowledge and find any kind of information on the Web on almost any topic for these tasks. However, conventional search engines are primarily optimized for returning *relevant* results and do not address the deeper analysis of arguments (e.g., argument quality and stance), or analysis of causal relationships. To close this gap, with the Touché lab's four shared tasks,[1] we in-

[*] Independent view, not influenced by Sanofi R&D France.

[1] 'touché' is commonly "used to acknowledge a hit in fencing or the success or appropriateness of an argument, an accusation, or a witty point." [https://merriam-webster.com/dictionary/touche]

tend to solicit the research community to develop respective approaches. In 2023, we organize the four following shared tasks:

1. Argumentative document retrieval from a generic web crawl to provide an overview of arguments and opinions on controversial topics.
2. Retrieval of web documents from a generic web crawl to understand whether a causal relationship between two events/actions exists (*new task*).
3. Image retrieval to corroborate and strengthen textual arguments and to provide a quick overview of public opinions on controversial topics.
4. Stance classification of comments on proposals from the multilingual participatory democracy platform CoFE,[2] written in different languages to support opinion formation on socially important topics (*new task*).

After having organized three successful Touché labs on argument retrieval at CLEF 2020–2022 [5, 7, 6], we propose a fourth lab edition to bring together researchers from the fields of information retrieval, natural language processing, and computational linguistics working on argumentation and causality. During the previous Touché labs, we received more than 210 runs from 64 participating teams. We manually labeled the relevance and argument quality of more than 27,000 argumentative texts, web documents, and images for 200 search topics; the topics and judgments are publicly available at **https://touche.webis.de**.

The previous three labs explored different granularities of argument retrieval and analysis: debates on various topics crawled from several online debating portals and their gist, complete web documents, and text passages; in the current lab iteration, we plan to investigate argument retrieval from the large web crawl corpus ClueWeb22-B [13] and stance detection of web documents and human-written comments in different languages. With the new task on evidence retrieval for causal questions, we aim for exploring effective approaches to retrieve web documents relevant to causality-related information needs and to analyze if a document supports or refutes the causal relationship specified in the question. Additionally, by repeating the task on image retrieval for arguments, we intend to collect new ideas that improve over the achieved results and to expand the test collection with additional manual judgments. Thus, we plan to investigate different retrieval modalities: text and images. As in the previous Touché editions, we will encourage participants to deploy their software in our cloud-based evaluation-as-a-service platform TIRA [14] for better reproducibility.

2 Task Definition

The first three Touché 2023 lab's shared tasks follow the classic TREC-style methodology: documents and search topics are provided to the participants who submit their ranked results (up to five runs) to be judged by human assessors. For the fourth task, the participants will submit the results with a predicted stance for respective data entries. The fourth Touché lab's edition will include the four shared tasks that are outlined below in detail.

[2] https://futureu.europa.eu

Task 1: Argument Retrieval for Controversial Questions. Given a controversial topic and a collection of web documents, the task is to retrieve and rank documents by relevance to the topic, by argument quality, and to detect the document's stance. Participants of Task 1 will retrieve documents from the ClueWeb22-B crawl for 50 search topics. Our human assessors will label the ranked results both for their general topical relevance and for the rhetorical argument quality [16], i.e., "well-writtennes": (1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the argument text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc. Optionally, participants will detect the documents' stance: pro, con, neutral, or no stance.

Analogously to the previous Touché editions, our volunteer assessors will annotate the document's topical relevance with three levels: 0 (not relevant), 1 (relevant), and 2 (highly relevant). The argument quality will also be labeled with three labels: 0 (low quality, or no arguments in the document), 1 (average quality), and 2 (high quality). The annotators will be provided with detailed annotation guidelines, including examples, and will participate in a training phase with an initial kappa test and a follow-up discussion to clarify potential misinterpretations. Afterwards, each annotator will independently judge the results for disjoint subsets of the topics (i.e., each topic will be judged by one annotator only). We use this annotation policy due to a high annotation workload.

To lower the entry barrier for participants who cannot index the whole ClueWeb22-B corpus on their side, we provide a first-stage retrieval possibility via the API of the search engine ChatNoir [4] and a smaller version of the corpus that contains one million documents per topic. Additionally, participants are provided with a number of previously compiled resources that include the document-level relevance and quality judgments from the previous Touché editions.[3] For the identification of claims and premises in documents, participants can use any existing argument tagging tool such as the TARGER API [9] hosted on our own servers or develop their own tools if necessary. We will use nDCG@k[4] to evaluate rankings and accuracy to evaluate stance detection.

Topics. For the tasks on controversial questions (Task 1) and image retrieval (Task 3), we provide 50 search topics that represent various debated societal matters. The topics were chosen from the online debate portals (debatewise.org, idebate.org, debatepedia.org, and debate.org) having the largest number of user-generated comments, and thus representing the matters of the highest societal interest. Each of these topics has a *title* (i.e., a question on a controversial issue), a *description* specifying the particular search scenario, and a *narrative* that serves as a guideline for the human assessors. The example topic is shown below:

[3] https://webis.de/data.html#touche-corpora

[4] The value of k will depend on the number of result submissions and, thus, the annotation workload (nDCG@5 was used in the previous Touché editions).

```
<title> Should teachers get tenure? </title>
<description> A user has heard that some countries do give teachers
tenure and others don't. Interested in the reasoning for or against
tenure, the user searches for arguments [...] </description>
<narrative> Highly relevant statements clearly focus on tenure for
teachers in schools or universities. Relevant statements consider tenure
more generally, not specifically for teachers, or [...] </narrative>
```

Task 2: Evidence Retrieval for Causal Questions. Given a causality-related topic and a collection of web documents, the task is to retrieve and rank documents by relevance to the topic. For 50 search topics, participants of Task 2 will retrieve documents from the ClueWeb22-B crawl that contain relevant causal evidence. Optionally, participants will detect the document's *causal* stance. A document can provide supportive evidence (a causal relationship between the cause and effect from the topic holds), refutative (a causal relationship does not hold), neutral (in some cases holds and in some does not), or no evidence is entailed.

Our volunteer assessors will label the topical relevance documents according to three relevance levels: 0 (not relevant), 1 (relevant), and 2 (highly relevant). The direction of causality will be considered, e.g., a document stating that B causes A will be considered as off-topic (not relevant) for the topic 'Does A cause B?'. The document's stance will also be labeled to evaluate the optional stance detection task. In general, the labeling procedure will be analogous to Task 1, where volunteer assessors will participate in training and a discussion.

Like in Task 1, ChatNoir [4] can also be used for first-stage retrieval, and we will provide a smaller version of the corpus that contains one million documents per topic. Participants are free to use any additional existing tools or datasets and are encouraged to develop their own.

Topics. The 50 search topics for Task 2 describe scenarios, when users search for confirmation of whether some causal relationship holds, e.g., to know the possible reason for a current physical condition. Each of these topics has a *title* (i.e., a causal question), *cause* and *effect* entities, a *description* specifying the particular search scenario, and a *narrative* serving as a guideline for the assessors. The topics were manually selected from a corpus of causal questions [8] and a graph of causal statements [10] such that they span a diverse set of domains:

```
<title>Can eating broccoli lead to constipation?</title>
<cause>broccoli</cause>
<effect>constipation</effect>
<description> A young parent has a child experiencing constipation
after eating some broccoli for dinner and is wondering whether broccoli
could cause constipation [...] </description>
<narrative> Relevant documents will discuss if broccoli and other
high-fiber foods can cause or ease constipation [...] </narrative>
```

Task 3: Image Retrieval for Arguments. Given a controversial topic and a collection of web documents with images, the task is to retrieve for each stance (pro and con) images that show support for that stance. Participants of Task 3 should retrieve and rank images, possibly utilizing the corresponding web documents, from a focused crawl of 30,000 images and for a given set of 50 search topics (the same as in Task 1) [12]. Like in the last edition of this task, the focus is on providing users with an overview of public opinions on controversial topics, for which we envision a system that provides not only textual but also visual support for each stance in the form of images. Participants are able to use the approximately 6,000 relevance judgments from the last edition of the task for training supervised approaches [11].[5] Similar to the other tasks, participants are free to use any additional existing tools and datasets or develop their own.

Although rank-based metrics for single image grids exist [17], none have been proposed so far for a 'pro-con' layout. Therefore, like the last year, participants' submitted results will be evaluated by the simple ratio of relevant images among 20 retrieved images, namely 10 images for each stance (precision@10). We will again use three increasingly strict definitions of relevance, corresponding to three precision@10 evaluation measures: being on-topic, being in support of some stance (i.e., an image is "argumentative"), and being in support of the stance for which the image was retrieved.

Task 4: Intra-Multilingual Multi-Target Stance Classification. Given a proposal on a socially important issue, its title, and topic, the task is to classify whether a comment on the proposal is *in favor*, *against*, or *other* towards the proposal. The data used for the evaluation of the participants' approaches comes from the CoFE participatory democracy platform; the respective dataset was created by Barriere et al. [3] and contains about 4,000 proposals and 20,000 comments written in 26 languages. The participants will have to classify into 3 classes multilingual comments from 6 different languages.[6] We also provide an automatic English translation of the proposals and titles that are written in any of the 24 official EU languages (plus Catalan and Esperanto) since a proposal can be written in one language and its corresponding comment in another.

For training their classifiers, the participants are provided with three datasets from the same debating platform: (1) CF_{E-D}: a small set of comments annotated with three stance labels, (2) CF_{S}: a larger set of comments that are self-annotated in a binary way (*in favor* or *against* only), and (2) CF_{U}: a large set of unlabeled comments. Since the class *other* cannot be put on the same scale as *in favor* or *against*, we will use a non-ordinal metric for evaluation widely used for the evaluation of stance classifiers. To account for the class imbalance we will evaluate submitted approaches using the macro-averaged F1-score.

Within the task, we organize two subtasks: (1) *Cross-debate Classification*: the participants should not use comments from debates that are in the test set and (2) *All-data-available Classification*: the participants can use all the available

[5] https://webis.de/data.html#touche-corpora

[6] German, English, Greek, French, Italian, and Hungarian.

data. Also, the participants can use any additional existing tools and datasets, e.g., the datasets that contain stance annotations created by Barriere et al. [2] and by Vamvas and Sennrich [15] for any of the subtasks.

Proposals. For Task 4, the participants are given a proposal and its title as well as the comment to classify its stance that are exemplified below:[7]

```
<title> Set up a program for returnable food packaging made from
recyclable materials </title>
<proposal> The European Union could set up a program for returnable
food packaging made from recyclable materials (e.g. stainless steel,
glass). These packaging would be produced on the basis of open standards
and cleaned according to [...] </proposal>
<comment> Ja, wir müssen den Verpackungsmül reduzieren. </comment>
<label> In favor </label>
```

3 Touché at CLEF 2022: Brief Overview

For the Touché 2022 lab, we received 58 registrations, from which 23 teams actively participated in the tasks and submitted 84 results (runs; every team could submit up to 5 runs). Our evaluation of the submitted results showed that the most effective approaches to argument retrieval all share common characteristics. For instance, most use various strategies for query reformulation and expansion, such as using synonyms, relevance feedback, or generating new queries from scratch with pre-trained language models. For Task 1 (argument gist retrieval), the most challenging was identifying that a pair of sentences (premise and claim) is coherent. An interesting observation is that re-ranking first-stage retrieval results based on a quality assessment of arguments almost always improves the retrieval effectiveness. Specifically for Task 2 (comparative questions), re-ranking based on important terms such as comparison objects and aspects or argument units in documents (premises and claims) was successful. In Task 2, stance detection was a new subtask; and some participants included a re-ranking step based on the predicted stance in their retrieval pipelines, which had some promising effects on the overall retrieval effectiveness. However, the overall still rather low effectiveness of the stance detection approaches leaves room for future improvements. For Task 3 (image retrieval), the recognition of sentiment and emotion and the use of optical character recognition to analyze the text in images were particularly helpful. Stance detection for images was also very challenging. We also provide an online web service to visually explore the submitted runs to Task 3 (cf. Figure 1). For more details about the Touché 2022 lab, refer to the overview paper [6]. We expect that the relevance and argument quality judgments and stance labels collected at the Touché 2022 lab will help participating teams achieve higher effectiveness in the new lab iteration.

[7]Example from https://futureu.europa.eu/en/processes/GreenDeal/f/1/proposals/83

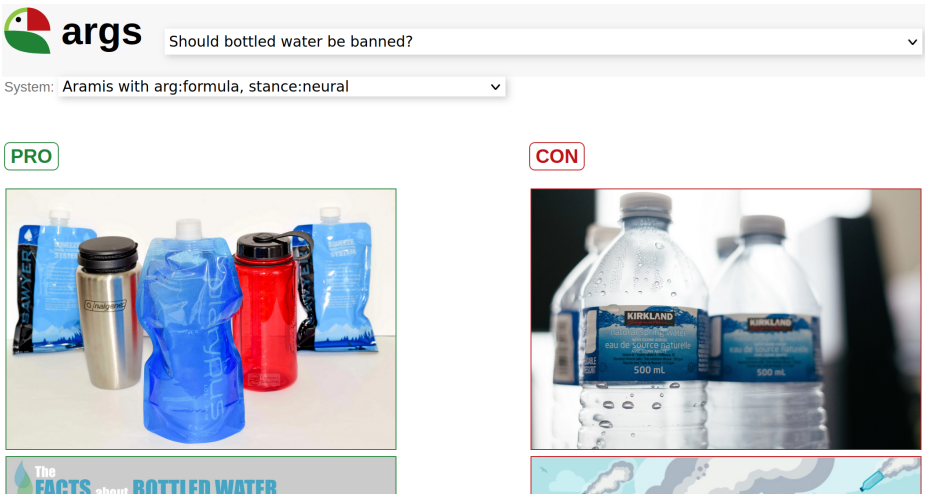

Figure 1. Screenshot of the Task 3 run browser (see the task page at https://touche.webis.de).

4 Conclusion

At Touché, we continue our activities aimed for fostering research in argument and causal retrieval and analysis, building respective test collections, and bringing the research community together. During the previous three years of organizing Touché, we have observed the development of the participants' submitted approaches from sparse to dense retrieval to the deployment of zero-shot models combined with extensive approaches to assess document "argumentativeness," argument quality, stance detection, and sentiment analysis in images.

With the new Touché lab, we plan to investigate how argument retrieval and argument analysis approaches can be applied to a large collection of web documents and to better understand the evoked challenges. By repeating the image retrieval task, we expect to collect more ideas for understanding how argument analysis techniques, historically developed for text, can be used for visual argument representation. Moreover, with the two new shared tasks, we want to explore web document retrieval and analysis for causality-related information needs and multilingual multi-target stance classification.

Acknowledgments

This work has been partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the project "ACQuA 2.0: Answering Comparative Questions with Arguments" (project 376430233) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999). V. Barriere's work was funded by the National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

References

[1] Ajzen, I.: The social psychology of decision making. Social psychology: Handbook of basic principles pp. 297–325 (1996)

[2] Barriere, V., Balahur, A., Ravenet, B.: Debating Europe: A multilingual multi-target stance classification dataset of online debates. In: Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences, pp. 16–21, European Language Resources Association, Marseille, France (Jun 2022), URL https://aclanthology.org/2022.politicalnlp-1.3

[3] Barriere, V., Jacquet, G., Hemamou, L.: CoFE: A new dataset of intra-multilingual multi-target stance classification from an online european participatory democracy platform. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2022), Taipei, Taiwan (2022)

[4] Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Elastic ChatNoir: Search engine for the ClueWeb and the Common Crawl. In: Proceedings of the 40th European Conference on IR Research, (ECIR 2018), Lecture Notes in Computer Science, vol. 10772, pp. 820–824, Springer (2018), URL https://doi.org/10.1007/978-3-319-76941-7_83

[5] 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, CEUR Workshop Proceedings, vol. 2696, CEUR-WS.org (2020), URL http://ceur-ws.org/Vol-2696/paper_261.pdf

[6] 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: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 3180, pp. 2867–2903, CEUR-WS.org (2022), URL http://ceur-ws.org/Vol-3180/paper-247.pdf

[7] 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: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 2936, pp. 2258–2284, CEUR-WS.org (2021), URL http://ceur-ws.org/Vol-2936/paper-205.pdf

[8] Bondarenko, A., Wolska, M., Heindorf, S., Blübaum, L., Ngomo, A.C.N., Stein, B., Braslavski, P., Hagen, M., Potthast, M.: CausalQA: A benchmark for causal question answering. In: 29th International Conference on Computational Linguistics (COLING 2022), pp. 3296–3308, International Committee on Computational Linguistics (Oct 2022), URL https://aclanthology.org/2022.coling-1.291

[9] Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., Panchenko, A.: TARGER: Neural argument mining at your

fingertips. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, pp. 195–200, Association for Computational Linguistics (2019), URL https://doi.org/10.18653/v1/p19-3031

[10] Heindorf, S., Scholten, Y., Wachsmuth, H., Ngonga Ngomo, A.C., Potthast, M.: CauseNet: Towards a causality graph extracted from the web. In: 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), pp. 3023–3030, Association of Computation Machinery (Oct 2020), https://doi.org/10.1145/3340531.3412763

[11] Kiesel, J., Potthast, M., Stein, B.: Dataset Touché22-Image-Retrieval-for-Arguments (2022), https://doi.org/10.5281/zenodo.6786948

[12] Kiesel, J., Potthast, M., Stein, B.: Dataset Touché23-Image-Retrieval-for-Arguments (2023), https://doi.org/10.5281/zenodo.7497994

[13] Overwijk, A., Xiong, C., Callan, J.: ClueWeb22: 10 billion web documents with rich information. In: Proceedsings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022), pp. 3360–3362, Association for Computing Machinery (2022), URL https://doi.org/10.1145/3477495.3536321

[14] Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA integrated research architecture. In: Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, The Information Retrieval Series, vol. 41, pp. 123–160, Springer (2019), URL https://doi.org/10.1007/978-3-030-22948-1_5

[15] Vamvas, J., Sennrich, R.: X-stance: A multilingual multi-target dataset for stance detection. In: Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing (SwissText/KONVENS 2020), CEUR-WS.org (2020), URL http://ceur-ws.org/Vol-2624/paper9.pdf

[16] 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 the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), pp. 176–187, Association for Computational Linguistics (2017), URL https://doi.org/10.18653/v1/e17-1017

[17] Xie, X., Mao, J., Liu, Y., de Rijke, M., Shao, Y., Ye, Z., Zhang, M., Ma, S.: Grid-based evaluation metrics for web image search. In: Proceedings of the 28th International World Wide Web Conference (WWW 2019), pp. 2103–2114, Association for Computing Machinery (2019), https://doi.org/10.1145/3308558.3313514