Overview of Touché 2024: Argumentation Systems

Extended Abstract

Johannes Kiesel[1], Çağrı Çöltekin[2], Maximilian Heinrich[1], Maik Fröbe[3], Milad Alshomary[4], Bertrand De Longueville[5], Tomaž Erjavec[6], Nicolas Handke[7], Matyáš Kopp[8], Nikola Ljubešić[6], Katja Meden[6], Nailia Mirzhakhmedova[1], Vaidas Morkevičius[9], Theresa Reitis-Munstermann[5], Mario Scharfbillig[5], Nicolas Stefanovitch[5], Henning Wachsmuth[4], Martin Potthast[7,10], and Benno Stein[1]

[1] Bauhaus-Universität Weimar [2] University of Tübingen
[3] Friedrich-Schiller-Universität Jena [4] Leibniz University Hannover
[5] European Commission, Joint Research Centre (JRC) [6] Jožef Stefan Institute
[7] Leipzig University [8] Charles University [9] Kaunas University of Technology
[10] 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 fifth edition of the lab features three shared tasks: (1) Human value detection (ValueEval), where participants detect (implicit) references to human values and their attainment in text; (2) Multilingual Ideology and Power Identification in Parliamentary Debates, where participants identify from a speech the political leaning of the speaker's party and whether it was governing at the time of the speech (new task); and (3) Image retrieval or generation in order to convey the premise of an argument with visually. In this paper, we briefly describe the planned setup for the fifth lab edition at CLEF 2024 and summarize the results of the 2023 edition.

Keywords: Argumentation · Human values · Ideology · Image retrieval.

1 Introduction

Decision-making and opinion-forming are everyday tasks, for which everybody has the chance to acquire knowledge on the Web on almost every topic. However, conventional search engines are primarily optimized for returning *relevant* results, which is insufficient for collecting and weighing the pros and cons for a topic. To close this gap of technologies that support people in decision-making and opinion-forming, the Touché lab's shared tasks[1] (https://touche.webis.de)

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

call for the research community to develop respective approaches. In 2024, we organize the three following shared tasks:

1. Human Value Detection (a continuation of ValueEval'23 @ SemEval [8][2]) features two subtasks in ethical argumentation of detecting human values in texts and their attainment, respectively.
2. Ideology and Power Identification in Parliamentary Debates features two subtasks in debate analysis of detecting the ideology and position of power of the speaker's party, respectively (new task).
3. Image Retrieval for Arguments (third edition, now joint task with ImageCLEF) is about the retrieval or generation of images to help convey an argument's premise.

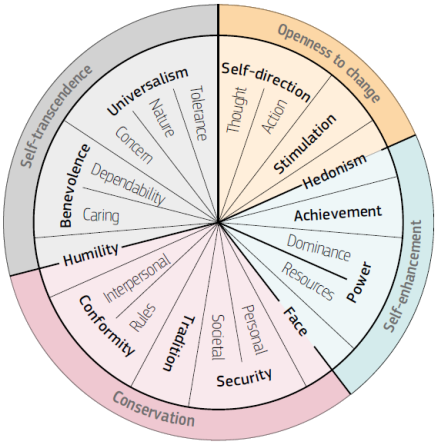
After having organized four successful Touché labs on argument retrieval at CLEF 2020–2023 [1, 2, 3, 4], we now organize a fifth lab edition to bring together researchers from the fields of information retrieval, natural language processing, and computational linguistics working on argumentation. During the previous Touché labs, we received 243 runs from 74 teams. We manually labeled the relevance and quality of more than 30,000 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 intends to widen its scope. After having explored causal questions in last year's edition, we now explore ethical argumentation in the task of human value detection. Compared to ValueEval'23 @ SemEval, this year's task features a larger dataset that also considers multiple languages and is created in a joint effort of over 70 value scholars. The second task targets deep linguistic analyses of debates, by analyzing the language of different ideologies and positions of power in parliamentary speeches. It is based on the multilingual ParlaMint corpus.[3] In addition, we have further developed the task of finding images for arguments, which this year focuses on finding images for specific arguments rather than topics. Moreover, for the first time we allows participants to use a text-to-image generative AI and submit generated 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 [11] for better reproducibility.

2 Task Definitions

Task 1: Human Value Detection (ValueEval) In argumentation, one has to consider that people have different beliefs and priorities of what is generally worth striving for (e.g., personal achievements vs. humility) and how to do so (e.g., being self-directed vs. respecting traditions), referred to as (human) values.

[2]Demo of best-performing approach: https://values.args.me
[3]https://www.clarin.eu/parlamint



Inner circle: 19 human values
(see https://valueeval.webis.de)

Outer circle: four general directions
(not used in this task)

- **Openness to change**
  Being independent and exploring
- **Self-enhancement**
  Seeking pleasure, wealth, and esteem
- **Conservation**
  Preserving group cohesion, order, and security
- **Self-transcendence**
  Helping others, close ones, and nature

Figure 1. The 19 values used in this task, shown in the Schwartz value taxonomy [16].

Overview The task is to identify the values of the widely accepted value taxonomy of Schwartz [16] (cf. Figure 1) and their attainment in long texts of eight languages (Bulgarian, Dutch, English, French, German, Hebrew, Italian, and Turkish). This taxonomy has been replicated in over 200 samples in 80 countries and is the backbone of value research [15]. A value can either be mentioned as something that is or should be attained (i.e., lead towards fulfilling the value) or something that is not attained or constrained. For example, for Security, (partial) attainment would mean that something is made safer or healthier. In contrast, an event can be stated in a way that thwarts or constrains safety or health. Participating teams can submit software in one or both of two sub-tasks: (1) Given a text, for each sentence, detect which human values the sentence refers to; and (2) Given a text, for each sentence and value this sentence refers to, detect whether this reference (partially) attains or constrains the value.

Data The task employs a collection of 3000 human-annotated texts between 400 and 800 words (across eight languages) from news articles and political texts (excerpts of speeches, debates, and party manifestos). Texts are sampled to reflect diverse opinions (different parties; mainstream news and not; from 2019 to 2023). The data is annotated as part of the ValuesML project[4] by over 70 value scholars. Dedicated team leaders per language train the respective annotators, consolidate annotations, and discuss disagreement (measured continuously by the organizers) in their language teams. The team leaders discuss issues with the organizers in bi-weekly meetings. The test set covers 20% of this data.

[4]https://knowledge4policy.ec.europa.eu/projects-activities/valuesml-unravelling-expressed-values-media-informed-policy-making_en

Evaluation Submissions are evaluated using macro F_1-score over all values. To facilitate quick develop-and-test cycles, the report facilities in TIRA provide participants with detailed feedback on the prediction errors in their submissions.

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 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. The first subtask is about predicting the political orientation (`left`–`right`) of speakers from their speeches. The second subtask is about 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 of producing and exchanging reasons to support claims. Though mostly associated with the exchange of words, argumentation often involves also images, either for exemplification, illustration, or evoking emotions. This task investigates how images can be used to convey an argument. Whereas the first two editions of this task followed the setup of Kiesel et al. [9] to retrieve images for a topic, this year's edition focuses on images for specific arguments.

Topic:	Photo identification at polling stations		
Claim:	Legislation to impose restrictive photo ID requirements has the potential to block millions of American voters.		
Premise:	People will forget their ID cards and be denied their right to vote.		
Submissions:			
Images:	 Image: Freepik.com		
Rationale:	Woman who forgot her ID	Embarrassed man who forgot his ID	Retired nuns barred from voting
Relevance:	1	2	0

Figure 2. Three possible submissions for the specified argument. The first (retrieved) image could help to convey the "forget"-part of the premise but does not relate to voting, unlike the second image (which was generated) that is thus rated higher on relevance (1 vs. 2). The third image (which was generated) does not indicate that someone forgets their ID or is barred from voting, and is thus rated irrelevant (0).

Overview Given a set of arguments, the task is to return for each argument several images that help convey the argument's premise. A suitable image could depict the argument or show a generalization or specialization. Participants can optionally add a short rationale that explains the meaning of the image.

Data The task data consists of 50 arguments, each consisting of a claim and a premise (cf. Figure 2). Premises are either facts or anecdotal. As document collection we provide a focused crawl of at least 1000 images per argument. Following the idea of the infinite index [5], we also provide an API for a Stable Diffusion image generator [14].

Evaluation Images can be (1) retrieved from the focused crawl and (2) generated using the Stable Diffusion API. The task follows the classic TREC-style methodology of teams submitting ranked results to be judged by human assessors. For a metric, the task uses standard nDCG [7] to represent a user looking through a ranked list of images retrieved for the specific argument.

3 Touché at CLEF 2023: Brief Overview

In 2023, Touché at CLEF included the following four shared tasks [2]: (1) Retrieval of documents that contain arguments and opinions on some controversial topic. (2) Retrieval of documents that contain evidence on whether a causal relationship between two events exists. (3) Retrieval of images to visually corroborate 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[5] to support opinion formation on socially important topics. Touché 2023 received 41 registrations, from which 7 teams actively participated in the tasks and submitted 30 results (runs; every team could submit up to 5 results). The three retrieval tasks followed the traditional TREC methodology: the participants received document collections and topics, and submitted their results (up to five runs) for each topic to be judged by human assessors. In the retrieval tasks, all teams used BM25 or BM25F [12, 13] for first-stage retrieval. The final ranked lists (runs) were often created based on argument quality estimation and predicted stance (Task 1), based on the presence of causal relationships in documents (Task 2), and exploiting the contextual similarity between images and queries and using the predicted stance for images (Task 3). The participants trained feature-based and neural classifiers to predict argument quality or stance, and often used ChatGPT with various prompt-engineering methods. To predict the stance of multilingual texts in Task 4, the participants used transformer-based models exploiting a few-step fine-tuning, data augmentation, and label propagation techniques.

The corpora, topics, and judgments are available on the Touché website.[6] Parts of the data are also available in BEIR [17] and `ir_datasets` [10].

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 four 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 2024 brings in new tasks and refines existing ones, targeting more subtle aspects of argumentation. With ethical argumentation (human value detection) and the identification of ideology and power in speeches we focus on deep linguistic analyses of argumentation, the former continuing a very successful task at SemEval'23. The third year of the image retrieval task explores a more specific task and the opportunity to submit generated images.

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

[5] https://futureu.europa.eu
[6] 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] 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

[8] Kiesel, J., Alshomary, M., Mirzakhmedova, N., Heinrich, M., Handke, N., Wachsmuth, H., Stein, B.: SemEval-2023 Task 4: ValueEval: Identification of Human Values behind Arguments. In: Kumar, R., Ojha, A.K., Doğruöz, A.S., Martino, G.D.S., Madabushi, H.T. (eds.) 17th International Workshop on Semantic Evaluation (SemEval 2023), pp. 2287–2303, Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://doi.org/10.18653/v1/2023.semeval-1.313

[9] Kiesel, J., Reichenbach, N., Stein, B., Potthast, M.: Image Retrieval for Arguments Using Stance-Aware Query Expansion. In: 8th Workshop on Argument Mining (ArgMining 2021), ACL (2021)

[10] MacAvaney, S., Yates, A., Feldman, S., Downey, D., Cohan, A., Goharian, N.: Simplified data wrangling with `ir_datasets`. In: 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022), pp. 2429–2436, ACM (2021), https://doi.org/10.1145/3404835.3463254

[11] 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

[12] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: 3rd Text REtrieval Conference (TREC 1994), NIST Special Publication, vol. 500-225, pp. 109–126, NIST (1994)

[13] Robertson, S.E., Zaragoza, H., Taylor, M.J.: Simple BM25 Extension to Multiple Weighted Fields. In: 13th International Conference on Information and Knowledge Management (CIKM 2004), pp. 42–49, ACM (2004), URL https://doi.org/10.1145/1031171.1031181

[14] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), pp. 10674–10685, IEEE (2022), https://doi.org/10.1109/CVPR52688.2022.01042

[15] Scharfbillig, M., Smillie, L., Mair, D., Sienkiewicz, M., Keimer, J., Pinho Dos Santos, R., Vinagreiro Alves, H., Vecchione, E., Scheunemann, L.: Values and Identities - a Policymaker's Guide. Tech. Rep. KJ-NA-30800-EN-N, European Commission's Joint Research Centre, Luxembourg (2021), https://doi.org/10.2760/349527

[16] Schwartz, S.H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.E., Demirutku, K., et al.: Refining the Theory of Basic Individual Values. Journal of personality and social psychology **103**(4) (2012), https://doi.org/10.1037/a0029393

[17] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Vanschoren, J., Yeung, S. (eds.) Neural Information Processing Systems (NeurIPS 2021), NeurIPS (2021)