Overview of Touché 2023: Argument and Causal Retrieval

Extended Version*

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

[1]*Friedrich-Schiller-Universität Jena*
[2]*Bauhaus-Universität Weimar*
[3]*Centro Nacional de Inteligencia Artificial (CENIA)*
[4]*Université Paris-Saclay*
[5]*Sanofi R&D France*
[6]*Alma Mater Studiorum – Università di Bologna*
[7]*Leipzig University and ScaDS.AI*
touche@webis.de https://touche.webis.de

Abstract

This paper is an extended overview of Touché: the fourth edition of the lab on argument and causal retrieval that was held at CLEF 2023. With the goal to create a collaborative platform for research on computational argumentation and causality, we organized 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 from a focused web crawl images showing support or opposition to some stance, and (d) multilingual multi-target stance classification, where participants detect the stance of comments on proposals from an online multilingual participatory democracy platform.

Keywords

Argument retrieval, Causal retrieval, Image retrieval, Stance classification, Argument quality, Causality

1. Introduction

Making informed decisions and forming opinions on a matter often involves not only weighing pro and con arguments but also considering cause–effect relationships for one's actions [2]. Nowadays, everybody has the chance to acquire knowledge and find any kind of information on the Web (facts, opinions, arguments, etc.) on almost any topic, which can help make decisions

*This overview extends the one published as part of the CLEF 2023 proceedings [1].

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

† Independent view, not influenced by Sanofi R&D France.

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

or get an overview of different standpoints. However, conventional web search engines are primarily optimized for returning *relevant* results and hardly 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 intended to solicit the research community to develop respective approaches. In 2023, we organized 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 find evidence of whether a causal relationship between two events 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 written in different languages on proposals from the multilingual participatory democracy platform CoFE,[2] to support opinion formation on socially important topics (*new task*).

The three Touché retrieval tasks followed the traditional TREC[3] methodology: document collections and topics were provided to participants, who then submitted their results (up to five runs) for each topic to be assessed by human assessors.

All teams that participated in the fourth Touché lab used BM25(F) [3, 4] for first-stage retrieval (except Task 4). 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 their own feature-based and neural classifiers to predict argument quality and stance. Also, many often used ChatGPT with various prompt-engineering methods. To predict the stance for multilingual texts (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 created at Touché are freely available to the research community and can be found on the lab's website.[4] Parts of the data are also already available via the BEIR [5] and `ir_datasets` [6] resources.

2. Lab Overview and Statistics

In the fourth edition of the Touché lab, we received 41 registrations from 21 countries (vs. 58 registrations in 2022). The majority of the lab registrations came from Germany (10 registered teams), followed by China and India (4 teams each), France (3 teams), Italy, Malaysia, and Sweden (2 teams each), Bangladesh, Botswana, Bulgaria, Canada, Guinea, Ireland, Netherlands, Nigeria, Mexico, Romania, Spain, Syria, Thailand, and United Kingdom (1 team each). Out of the

[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]
[2]https://futureu.europa.eu
[3]https://trec.nist.gov/
[4]https://touche.webis.de/

41 registered teams, 7 actively participated (1 team submitted results for Task 1 and Task 2 each, 3 teams participated in Task 3, and 2 teams in Task 4) by making valid result submissions (previous editions had more active participants, with 23 active teams in 2022, 27 participating teams in 2021 and 17 teams in 2020).

We used TIRA [7] as the submission platform for Touché 2023 through which the participants could either submit software or upload run files.[5] Software submissions increase reproducibility, as the software can later be executed on different data of the same format. Overall, 5 out of the 7 active teams made software submissions. To submit software, a team implemented their approach in a Docker image that they then uploaded to their dedicated Docker registry in TIRA. Software submissions in TIRA are immutable, and after the docker image had been submitted, the teams specified the to-be-executed command—the same Docker image can thus be used for multiple software submissions (e.g., by changing some parameters). A team could upload as many Docker images or software submissions as needed (the images were not public while the shared tasks were ongoing). To improve reproducibility, TIRA executes software in a sandbox by blocking the internet connection. This ensures that the software is fully installed in the Docker image, which eases running the software later. For the execution, the participants could select the resources that their software had available for execution out of four options: (1) 1 CPU core with 10 GB RAM, (2) 2 cores with 20 GB RAM, (3) 4 cores with 40 GB RAM, or (4) 1 CPU core with 10 GB RAM and 1 Nvidia GeForce GTX 1080 GPU with 7 GB RAM. Also, the participants were able to run their software multiple times using different resources to investigate the scalability and reproducibility (e.g., whether the software executed on a GPU yields the same results as on a CPU). TIRA used a Kubernetes cluster with 1,620 CPU cores, 25.4 TB RAM, and 24 GeForce GTX 1080 GPUs to schedule and execute the software submissions, allocating the resources that the participants selected for their submissions.

3. Task 1: Argument Retrieval for Controversial Questions

The goal of the first task was to support individuals who search for opinions and arguments on socially important, controversial topics like "Are social networking sites good for our society?". The previous task iterations explored different granularities of argument retrieval and analysis: a focused crawl of debates on various controversial topics from several online debating portals and the arguments' concise gist [8, 9, 10]. For the fourth edition of the task, our focus shifted towards retrieving argumentative web documents from the web crawl corpus ClueWeb22 [11]. The topics and manual judgments from the previous task iterations were provided to the participants to enable approaches that leverage training and parameter tuning.

3.1. Task Definition

Given a controversial topic and a collection of web documents, the task was to retrieve and rank documents by relevance to the topic, ideally ranking higher documents that contain high-quality arguments, and to (optionally) detect the document's stance. Participants of Task 1 needed to retrieve documents from the ClueWeb22-B crawl for 50 search topics.

[5]https://tira.io

Table 1
Example topic for Task 1: Argument Retrieval for Controversial Questions.

Number	34
Title	Are social networking sites good for our society?
Description	Democracy may be in the process of being disrupted by social media, with the potential creation of individual filter bubbles. So a user wonders if social networking sites should be allowed, regulated, or even banned.
Narrative	Highly relevant arguments discuss social networking in general or particular networking sites, and its/their positive or negative effects on society. Relevant arguments discuss how social networking affects people, without explicit reference to society.

To lower the entry barrier for participants who could not index the whole ClueWeb22-B corpus on their side, we provided a first-stage retrieval possibility via the API of the BM25F-based search engine ChatNoir [12] and a smaller version of the corpus that contained one million documents per topic. To identify arguments (claims and premises) in documents, participants could use any existing argument tagging tool such as the TARGER API [13] hosted on our servers or develop their own tools if necessary.

3.2. Data Description

Topics. For the task on argument retrieval for controversial questions (Task 1), we provided 50 search topics representing various debated societal issues. These issues were chosen from the online debate portals (debatewise.org, idebate.org, debatepedia.org, and debate.org), with the largest number of user-generated comments and thus representing the highest societal interest. For each such case, we formulated a topic's *title* (i.e., a question on a controversial issue), a *description* specifying the particular search scenario, and a *narrative* that served as a guideline for the human assessors (see Table 1 for an example).

Document Collection. The retrieval collection was the ClueWeb22 (Category B) corpus [11] that contains 200 million multilingual most frequently visited web pages like Wikipedia articles or news websites. The indexed corpus was available via the ChatNoir API[6] and its Python module[7] integrated in PyTerrier [14].

3.3. Evaluation Setup

Our human assessors labeled the submitted by the task participants ranked lists of documents both for their general topical relevance and for the rhetorical argument quality [15], i.e., "well-writtenness": (1) whether the document contains arguments and whether the argument text has

[6]https://github.com/chatnoir-eu/chatnoir-api
[7]https://github.com/chatnoir-eu/chatnoir-pyterrier

a good style of speech, (2) whether the argument text has a proper sentence structure and is easy to follow, and (3) whether it includes profanity, has typos, etc. Also, the documents' stance towards the argumentative search topics was labeled as 'pro', 'con', 'neutral', or 'no stance'.

Analogously to the previous Touché editions, our volunteer assessors annotated the document's topical relevance with three labels: 0 (not relevant), 1 (relevant), and 2 (highly relevant). The argument quality was also labeled with three classes: 0 (low quality or no arguments in the document), 1 (average quality), and 2 (high quality). We provided the annotators with detailed annotation guidelines, including examples. In the training phase, we asked 4 annotators to label the same 20 randomly selected documents (initial Fleiss' kappa values: relevance κ=0.39 (fair agreement), argument quality κ=0.34 (fair agreement), and κ=0.51 (moderate agreement) for labeling the stance) and in the follow-up discussion clarified potential misinterpretations. Afterward, each annotator independently judged the results for disjoint subsets of the topics (i.e., each topic was judged by one annotator only). We used this annotation policy due to a high annotation workload. Our human assessors labeled in total 747 documents pooled from 8 runs using a top-10 pooling strategy implemented in the TrecTools library [16].

3.4. Submitted Approaches and Evaluation Results

In 2023, only one team participated in Task 1 and submitted seven runs. We, thus, decided to evaluate all the participant's runs and an additional baseline. Below, we summarize and describe the submitted approaches to the task and evaluation results.

The task's baseline run by *Puss in Boots* used the results that ChatNoir [12] returned for the topics' titles used as queries without any pre-processing. ChatNoir is an Elasticsearch-based search engine for the ClueWeb and Common Crawl web corpora that employs BM25F ranking (fields: document title, keywords, main content, and the whole document) and SpamRank scores [17]. The document stance for the baseline run was predicted by zero-shot prompting the Flan-T5 model [18][8] after summarizing the document's main content with BART [19].[9]

Team *Renji Abarai* [20] submitted seven runs in total. Their baseline run used the top-10 results returned by ChatNoir for the pre-processed topics' titles used as queries. During pre-processing, stop words were first removed using their own handcrafted list of terms; the remaining query terms were then lowercased and lemmatized with the Stanza NLP library [21]. For the other six runs, the results of the baseline run were re-ranked based on the predicted argument quality and predicted document stance. Argument quality was predicted using either a meta-classifier (random forests) trained on the class predictions and class probabilities of six base classifiers or by prompting ChatGPT. Each base classifier (feedforward neural network, LightGBM [22], logistic regression, naïve Bayes, SVM, and random forests) was trained in two variants: (1) using a set of 32 handcrafted features (e.g., sentiment, spelling errors, the ratio of arguments in documents, etc.) and (2) using documents represented with the instruction-based fine-tuned embedding model `INSTRUCTOR` [23]. All the classifiers were trained on the manual argument quality labels from the Touché 2021 Task 1 [9], which was also used to select examples for a few-shot prompting ChatGPT. The resulting ranked lists submitted by Renji Abarai differed

[8]Pre-trained model: https://huggingface.co/google/flan-t5-base; maximum generated tokens: 3; the prompt is given in Appendix A.
[9]Pre-trained model: https://huggingface.co/facebook/bart-large-cnn; minimum length: 64; maximum length: 256.

Table 2
Results of all runs submitted to Task 1: Argument Retrieval for Controversial Questions. Reported are the mean nDCG@10 for relevance and argument quality and macro-averaged F_1 for stance detection. Since Renji Abarai re-ranked the same set of documents for all the runs, this yields identical stance detection results. The task baseline run by Puss in Boots is shown in bold.

Team	Run Tag	nDCG@10		F_1 macro
		Rel.	Qua.	Stance
Puss in Boots	**ChatNoir [12]**	**0.834**	**0.831**	**0.203**
Renji Abarai	stance_ChatGPT	0.747	0.815	0.599
Renji Abarai	stance-certainNO_ChatGPT	0.746	0.811	0.599
Renji Abarai	ChatGPT_mmGhl	0.718	0.789	0.599
Renji Abarai	ChatGPT_mmEQhl	0.718	0.789	0.599
Renji Abarai	meta_qual_score	0.712	0.771	0.599
Renji Abarai	baseline	0.708	0.766	0.599
Renji Abarai	meta_qual_prob	0.697	0.774	0.599

in the type of argument quality classifiers used for re-ranking, whether predicted classes or probabilities were used, or if the predicted document stance was considered. The document stance for all the runs was predicted using ChatGPT.

Table 2 shows the results of all submitted runs with respect to relevance, argument quality, and stance detection (more detailed results for each submitted run, including the 95% confidence intervals, are in Tables 9 and 10 in Appendix B). In general, none of the submitted participant results outperformed the argumentation-agnostic BM25F-based task baseline. This is due to the worse effectiveness of the team's initial retrieval results ('baseline' run in Table 2) that were used in the re-ranking step. Five participants' re-ranking strategies were able to improve over their initial ranking. The most effective participant approach ('stance_ChatGPT' run in Table 2) exploited ChatGPT to predict the argument quality and stance. Then, a two-step re-ranking strategy was used: (1) move the 'no stance' documents to the bottom of the ranked list, and then (2) re-rank the remaining documents based on the predicted argument quality in the descending order. Thus, the promising future direction can be to apply the proposed re-ranking approach to the official task baseline run.

4. Task 2: Evidence Retrieval for Causal Questions

The goal of the Touché 2023 lab's second task was to support users who search for answers to causal yes-no questions like "Do microwave ovens cause cancer?", supported by relevant evidence instances. In general, such causal questions ask if something causes or does not cause something else.

4.1. Task Definition

Given a causality-related topic and a collection of web documents, the task was to retrieve and rank documents by relevance to the topic. For 50 search topics, participants of Task 2 needed to retrieve documents from the ClueWeb22-B crawl that contain relevant causal evidence. An

Table 3
Example topic for Task 2: Evidence Retrieval for Causal Questions.

Number	39
Title	Do microwave ovens cause cancer?
Cause	microwave ovens
Effect	cancer
Description	A user has recently learned that radiation waves can cause cancer. They are wondering if their microwave oven produces radiation waves and if these are dangerous enough to cause cancer.
Narrative	Highly relevant documents will provide information on a potential causal connection between microwave ovens and cancer. This includes documents stating or giving evidence that the first is (or is not) a cause of the other. Documents stating that there is not enough evidence to decide either way are also highly relevant. Relevant documents may contain implicit information on whether the causal relationship exists or does not exist. Documents are not relevant if they either mention one or both concepts, but do not provide any information about their causal relation.

optional task was to 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), or neutral (in some cases holds and in some does not). Like in Task 1, ChatNoir [12] could be used for first-stage retrieval.

4.2. Data Description

Topics. The 50 search topics for Task 2 described scenarios where users search for confirmation of whether some causal relationship holds. For example, a user may want to know the possible reason for a current physical condition. Each of these topics had 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 (see Table 3). The topics were manually selected from a corpus of causal questions [24] and a graph of causal statements [25] such that they spanned a diverse set of domains.

Document Collection. The same document collection as in Task 1 was used.

4.3. Evaluation Setup

Relevance assessments were gathered with volunteer human assessors. The assessors were instructed to label documents as *not relevant* (0), *relevant* (1), or *highly relevant* (2). The direction of causality was considered, i.e., a document stating that B causes A was considered off-topic (not relevant) for the topic "Does A cause B?". The document's stance was also labeled to evaluate the optional stance detection task. The labeling procedure was analogous to Task 1,

Table 4
Relevance results of all runs submitted to Task 2: Evidence Retrieval for Causal Questions. Reported are the mean nDCG@5 for relevance and macro-averaged F_1 for stance detection; Puss in Boots baseline is in bold. The dagger† indicates a statistically significant improvement ($p < 0.05$, Bonferroni corrected) over the Puss in Boots baseline. Team He-Man did not detect the stance.

Team	Run Tag	nDCG@5 Relevance	F_1 macro Stance
He-Man	no_expansion_rerank	0.657†	–
Puss in Boots	**ChatNoir [12]**	**0.585**	**0.256**
He-Man	gpt_expansion_rerank	0.374	–
He-Man	causenet_expansion_rerank	0.268	–

where volunteer assessors participated in training and a discussion. Agreement on the same 20 randomly selected documents across 4 annotators was measured with Fleiss' kappa. Before the discussion, the agreement was $\kappa = 0.58$ for relevance and $\kappa = 0.55$ for stance assessment (both indicate a moderate agreement). After discussing discrepancies, similar to Task 1, each annotator labeled a disjoint set of topics. We pooled the top-5 documents from each submitted run (plus additional baseline) and labeled 718 documents in total.

4.4. Submitted Approaches and Evaluation Results

One team *He-Man* [26] participated in Task 2 and submitted three runs. Like the baseline run *Puss in Boots*, all three participant runs used ChatNoir [12] for first-stage retrieval. For two runs, first, the cause and effect events were extracted from the topic title field using dependency tree parsing. Next, query expansion and query reformulation approaches were applied. In the query expansion approach, the topic title was expanded with semantically related concepts from the CauseNet, a graph of causal relations [25]. For this, all relations in the CauseNet-Precision variant were embedded using BERT [27]. Next, the embedding's cosine similarity was compared with the embedding of the topic's relation. The top-5 terms from the documents linked to the matched CauseNet relation were then used to expand the query. The second approach, the query reformulation technique, fed the deconstructed topic title in a semi-structured JSON format to ChatGPT. The chatbot was then prompted to generate new query variants, exchanging causes, effects, and causal phrases. All three query variants (original topic title, expanded query, and reformulated query) were then submitted to ChatNoir. Finally, all approaches re-ranked the results using a position bias. Documents containing the causal relationship from the topic earlier in the document were ranked higher. To detect the position of the relation, the same dependency tree parsing developed for the query deconstruction was used.

The task's baseline run of *Puss in Boots* additionally predicted the document stance by first summarizing a document's main content with BART [19],[10] and then zero-shot prompting the Flan-T5 model [18].[11]

[10]Pre-trained model: https://huggingface.co/facebook/bart-large-cnn; minimum length: 64; maximum length: 256.
[11]Pre-trained model: https://huggingface.co/google/flan-t5-base; maximum generated tokens: 3; the prompt is given in Appendix A.

Table 4 shows the evaluation results for Task 2 (more detailed results for each submitted run, including the 95% confidence intervals, is in Table 11 in Appendix B). We report nDCG@5 for relevance-based retrieval effectiveness and macro-averaged F_1 for stance detection. The Puss in Boots baseline was more effective in terms of relevance than the two participant runs that used query expansion. However, the participant run which only applied re-ranking, statistically significantly outperformed the baseline. This suggests that the participants' query expansion techniques degrade the first-stage retrieval results and the re-ranking approach applied afterwards cannot compensate for the substantially worse performance of the query expansion. The participating team opted to not detect the stance. Therefore, only the baseline run could be evaluated for stance detection, achieving an F_1-score of 0.256.

5. Task 3: Image Retrieval for Arguments

The goal of the third task was to provide argumentation support through image search. The retrieval of relevant images should provide both a quick visual overview of frequent arguments on some topic and compelling images to support one's argumentation. To this end, the second edition of this task continued with the retrieval of images which can be posted to either indicate an agreement or disagreement to some stance on a given topic. Images should be retrieved as two separate lists, similar to a textual argument search (e.g., https://args.me).

5.1. Task Definition

Given a controversial topic and a collection of web documents with images, the task was to retrieve for each stance (pro and con) images that indicate support for that stance. Participants of Task 3 should retrieve and rank images, possibly utilizing the corresponding web documents, from a focused crawl of 55,691 images and for a given set of 50 topics (which were used by other tasks in previous years) [28]. 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 were able to use the approximately 6,000 relevance judgments from the last edition of the task for training supervised approaches [29].[12] Similar to the other tasks, participants were free to use any additional existing tools and datasets or develop their own.

5.2. Data Description

Topics. Task 3 employs 50 controversial topics from earlier Touchè editions (e.g., used in 2021), but which were not used in the first edition of this task. As for Task 1 (cf. Section 3), we provided for each topic a title, description, and narrative. The description and narrative were adapted as needed to fit the image retrieval setting.

Document Collection. This task's document collection stems from a focused crawl of 55,691 images and associated web pages from late 2022. We downloaded the top-100 images

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

and associated web pages from Google's image search for 2,209 queries. Nearly half of the queries (namely 1,050) were created like in the first edition of this task, by appending filter words like "good," "meme," "stats," "reasons," or "effects" to a manually created query for each topic. The remaining 1,159 queries were collected from participants in an open call, which allowed anyone to submit queries until the end of December 2022. Of these queries, 557 were created manually (57 by team Neville Longbottom, 250 by team Hikaru Sulu, and 250 by us), and the remaining were created using ChatGPT by team Neville Longbottom: they asked ChatGPT for a list of pro and con arguments for each topic, then for an image description illustrating the respective arguments, and then for a search query to match the description. From the search results we attempted to download 147,264 images, but discarded 5,666 for which we could not download the image, 6,619 for which the image was more than 2,000 pixels wide or high,[13] 20,696 for which an initial text recognition using Tesseract[14] yielded more than 20 words,[15] 8,538 for which the web page could not be downloaded, 484 for which the web page contained no text, and 45,254 for which we could not find the image URL in the web page DOM. After a duplicate detection using pHash,[16] the final dataset contains 55,691 images. The dataset contains various resources for each image, including the associated page for which it was retrieved as an HTML page and as a detailed web archive,[17] information on how Google ranked the image, and information from Google's Cloud Vision API,[18] e.g., detected text and objects.

5.3. Evaluation Setup

Our two volunteer human assessors labeled the ranked results by the task participants (i.e., the images) for their relevance to the topic's narrative. First, assessors decided whether an image is on topic (yes or no). If so, they also decided whether an image is relevant according to the pro-side of the narrative, its con-side, or both: 0 (not relevant), 1 (relevant), and 2 (highly relevant), though we did not distinguish between levels 1 and 2 in our evaluation. However, assessors were instructed that an image could not be highly relevant for both pro and con to indicate a preference. We provided the assessors with guidelines, discussed several examples, and discussed edge cases as they came up. Achieved Fleiss' κ values (measured on three topics for which all assessors labeled all images) were for on-topic 0.38 (fair), for pro 0.34 (fair), and for con 0.31 (fair). Without distinguishing levels 1 and 2, the agreement increases to 0.45 for pro (moderate) and 0.52 for con (moderate). Our human assessors labeled in total 6692 images.

Although rank-based metrics for single image grids exist [30], none have been proposed so far for a 'pro-con' layout. Therefore, participants' submitted results were evaluated by the ratio of relevant images among 20 retrieved images, namely 10 images per stance (precision@10). We again used 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.

[13]As one use case for our task is getting a quick overview of arguments, we excluded overly large images
[14]https://github.com/tesseract-ocr/tesseract
[15]To sharpen our focus on images, this year we tried to exclude images that are merely screenshots of text documents
[16]https://www.phash.org/; same procedure as in the previous year
[17]Archived using https://github.com/webis-de/scriptor
[18]https://cloud.google.com/vision

Table 5
Relevance results of all runs submitted to Task 3: Image Retrieval for Argumentation. Reported are the mean precision@10 for all three definitions of relevance; Minsc baseline is in bold. The dagger† indicates a statistically significant improvement ($p < 0.05$, Bonferroni corrected) over the baseline.

Team	Run Tag	Precision@10		
		On-topic	Arg.	Stance
Neville Longbottom	clip_chatgpt_args.raw	0.785†	0.338†	0.222†
Neville Longbottom	clip_chatgpt_args.debater	0.684†	0.341†	0.216†
Hikaru Sulu	Keywords	0.664†	0.350†	0.185†
Hikaru Sulu	Topic-title	0.770†	0.335†	0.179†
Neville Longbottom	bm25_chatgpt_args.raw	0.572	0.274	0.166†
Jean-Luc Picard	No stance detection	0.523†	0.292†	0.162
Neville Longbottom	bm25_chatgpt_args.diff	0.442	0.240	0.150
Jean-Luc Picard	Text+image text stance detection	0.502†	0.272	0.144
Jean-Luc Picard	BM25 Baseline	0.536†	0.268†	0.141
Jean-Luc Picard	Text stance detection	0.498†	0.262†	0.136
Neville Longbottom	bm25_chatgpt_args.debater	0.416	0.201	0.128
Minsc	**Aramis**	**0.376**	**0.194**	**0.102**
Jean-Luc Picard	Image text stance detection	0.369	0.196	0.098

5.4. Submitted Approaches and Evaluation Results

In total, three teams participated in Task 3 and submitted 12 runs in total, not counting the submitted queries described above. Table 5 shows the results of all submitted runs (more detailed results for each submitted run, including the 95% confidence intervals, are in Tables 12, 13, and 14 in Appendix B). Overall, scores are considerably lower than last year, where precision@10 for stance relevance was as high as 0.425. We attribute this to the new set of topics, which contained much more questions that were hard to picture.

As a baseline (team *Minsc*), we used the model of [31], which was developed by a collaboration of two teams that participated in last year's task: Aramis and Boromir.[19] The approach employed standard retrieval and a set of handcrafted features for argumentativeness detection. For retrieval, the approach used Elasticsearch's BM25 (default settings: k_1=1.2 and b=0.75) with each image (document) represented by the text from the web page around the image and text recognized in the image using Tesseract.[14] For argumentativeness detection, the approach used a neural network classifier based on thirteen different features (color properties, image type, and textual features), and trained on the ground-truth annotations from last year. The features are calculated from, amongst others, the query, the image text, the HTML text around the image, the interrelation and sentiments of the mentioned texts, and the colors in the image. The approach used random stance assignment. Since this baseline performed much worse than anticipated, we expect a bug in the implementation.

Team *Hikaru Sulu* submitted two valid runs. Their approach used CLIP [32] to calculate the similarity between keywords and images and retrieved, per topic, the images most similar to

[19]Since no stance model convincingly outperformed naive baselines in their evaluation, we use the simple both-sides baseline that assigns each image to both stances

one of the keywords. For the first run, they used the topic title as a keyword, but for the second run, they extracted all nouns and verbs from the topic title and extended that list with synonyms and antonyms from WordNet [33]. The stance was determined randomly, which performed in their internal evaluation better than using different keywords for pro and con. As Table 5 shows, the extended list lead to retrieving more on-topic images, but less argumentative ones.

Team *Jean-Luc Picard* [34] submitted five valid runs. Their first run used the web page text indexed by PyTerrier's BM25 [14] (default settings: k_1=1.2 and b=0.75). For the other runs, they used a pipeline of query preprocessing, the same BM25-based retrieval as their first run, stance detection, and re-ranking. For query preprocessing, they created a parse tree of the topic and filtered out frequent words to create a short query. The runs correspond to four different stance detection approaches: (1) random or (2) using a zero-shot classification based on the pre-trained BART MultiNLI model[20] that assigns the image to pro, contra, or neutral (i.e., will be discarded) based on the (a) web page text, (b) the image text, or (c) both texts. After that, images were re-ranked: for each topic, images were generated with Stable Diffusion [35] using the preprocessed query as prompt, then SIFT keypoints were identified [36] in both retrieved and generated image and matched between the two images, and then the result list was re-ranked as per the number of matched keypoints in descending order. Similar to the internal evaluation of team Hikaru Sulu, a random stance assignment performed best.

Team *Neville Longbottom* [37] submitted five valid runs. They first employed ChatGPT[21] to generate image descriptions for each topic and stance (neither description nor narrative was used). Then, they retrieved images with these descriptions, either (1) using the web page text close to the image indexed via PyTerrier's BM25 [14] (default settings: k_1=1.2 and b=0.75) or (2) using CLIP [32] for ranking images by their similarity to the description. For runs 3–5, the approach continued by re-ranking the result list, either (a) by penalizing the BM25-score of an image with the BM25-score of the image for the respective other stance's description (re-ranking the results of run (1)) or (b) by using IBM's debater pro-con score [38] between the topic title and the text close to the image on the web page (2 runs; re-ranking results of run (1) or (2)). The CLIP method without re-ranking performed best.

6. Task 4: Multilingual Multi-Target Stance Classification

In this edition of the Touché lab, we proposed a new task on multilingual multi-target stance classification of comments to proposals from an online participatory democracy platform. The goal of the fourth task was to build technologies that help analyze opinions on a wide range of socially important topics. Large-scale deployment of such technologies faces challenges like multilingualism or high variability of the topics of interest and hence is the target of this task.

6.1. Task Definition

Given a proposal on a socially important issue, its title, and its topic, the task was to classify whether a comment on the proposal is 'in favor', 'against, or 'neutral' towards the commented

[20]https://huggingface.co/facebook/bart-large-mnli
[21]https://chat.openai.com/chat

Table 6
A data instance for Task 4: Multilingual Multi-Target Stance Classification.

Number	34
Title	Set up a program for returnable food packaging made from recyclable materials
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 [. . .]
Comment	Ja, wir müssen den Verpackungsmül reduzieren
Label	In favor

proposal. The participants needed to classify multilingual comments written in 6 different languages[22] into the 3 stance classes. Comments to the proposals could be written in a different language than the proposal itself, and multiple comments could target the same proposal.

Within the task, we organized two subtasks: (1) *Cross-debate Classification*, where the participants were not allowed to use for training comments on proposals that also had comments in the test set, and (2) *All-data-available Classification*, where the participants could use all the available data. Also, the participants could use any additional existing tools or previously published datasets like Debating Europe [39] or X-Stance [40].

6.2. Data Description

The proposals and comments used in Task 4 stem from the Conference on the Future of Europe (CoFE),[23] an online debating platform where users can write proposals and comment on the suggested ideas. The initial dataset was comprised of 4,247 proposals and 20,102 comments written in 26 languages (24 official languages of the European Union plus Catalan and Esperanto) [41, 42]. As shown in Figure 1, English, German, and French were the most commonly used languages to write proposals and comments on the platform. An example of a proposal, a corresponding comment, and the stance of the comment is shown in Table 6.[24]

For training stance classifiers, participants were provided with three datasets: (1) $CF_{E\text{-}D}$, a small set of comment–proposal pairs manually annotated by expert native speakers with three stance labels, (2) CF_S, a larger set of comment–proposal pairs where comment authors selected either 'in favor' or 'against' stance (no 'neutral' label was available for selection), and (3) CF_U, a large set of unlabeled comment–proposal pairs. The fourth dataset, $CF_{E\text{-}T}$, was built in the same way as the $CF_{E\text{-}D}$ dataset but was held out for testing the submitted approaches (see Table 7). The annotations of the $CF_{E\text{-}D}$ and $CF_{E\text{-}T}$ datasets were made based on the untranslated comments and English translations of the proposals, to understand the context. The test set, $CF_{E\text{-}T}$, contained comments from the most common languages except for Spanish (see Figure 1).

[22] German, English, Greek, French, Italian, and Hungarian.
[23] https://futureu.europa.eu
[24] From https://futureu.europa.eu/en/processes/GreenDeal/f/1/proposals/83.

Table 7
Number of languages, comments, proposals, and stance label distribution of the datasets used in Task 4.

Dataset	**# Languages**	**# Comments**	**# Proposals**	**Stance**		
				In favor	Against	Neutral
CF_S	25	7,002	2,731	77.7 %	22.3 %	0.0 %
CF_U	25	13,213	2,892	0.0 %	0.0 %	0.0 %
$CF_{E\text{-}D}$	4	1,414	936	53.3 %	8.3 %	38.4 %
$CF_{E\text{-}T}$	6	1,228	771	55.2 %	17.7 %	27.1 %

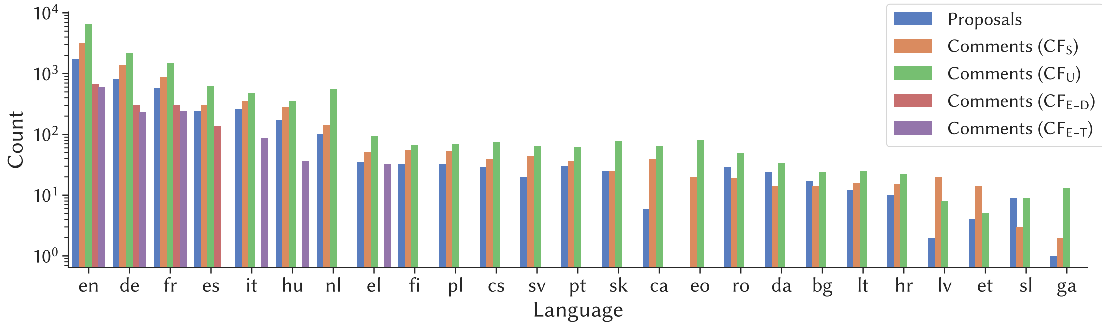

Figure 1: Number of proposals and comments per language (using ISO 3166-1 alpha-2 country codes) for the 4 datasets used in Task 4.

6.3. Submitted Approaches and Evaluation Results

Two teams participated in Task 4 and submitted 8 runs in total. Below, we briefly describe the participants' approaches plus additional baseline runs.

Team *Cavalier* was our baseline that implemented three stance classifiers. For Subtask 1 (cross-debate classification) we implemented two baseline classifiers: The first one ('Cavalier Simple') always predicts the majority class ('in favor'). The second baseline ('Cavalier Subtask 1') is based on the transformer-based multilingual masked language model XLM-R [43, 42]. This model was first fine-tuned on the X-Stance dataset [40] and the CF_S dataset to classify just two stance classes ('in favor' or 'against') and subsequently fine-tuned again on the Debating Europe dataset [39] to classify all three stance classes ('in favor', 'against', or 'neutral'). All comments or proposals appearing in the test set, $CF_{E\text{-}T}$, were removed before fine-tuning. The baseline classifier for Subtask 2 (all-data-available classification) used the same model and training scheme as 'Cavalier Subtask 1', but comments or proposals that appeared in the test set were not removed from the training data.

Team *Silver Surfer* [44] submitted six valid runs to Subtask 2. Their approaches were based on fine-tuning pre-trained English and multilingual transformer models: a RoBERTa model [45],[25]

[25]https://huggingface.co/roberta-base

an XLM-R model [43][26] and two BERT models [27].[27] To increase the size of the training data, the team applied data augmentation using back-translation [46] (i.e., translating texts to other languages and then back to the original language) and used label spreading [47] to transfer labels from the $CF_{E\text{-}D}$ dataset to the CF_U dataset. The team first fine-tuned a RoBERTa model (Run 2, comments translated to English) and an XLM-R model (Run 3, no translation) on the CF_S dataset as well as on the CF_U dataset after applying label spreading. Run 4 used the $CF_{E\text{-}D}$ dataset after data augmentation using back-translation to fine-tune an XLM-R model. For Run 5, the team fine-tuned a RoBERTa model on the comments from the $CF_{E\text{-}D}$ dataset, translating all comments to English. The team's Run 6 used a two-step training approach, where they first fine-tuned an English BERT model on binary stance classification based on the translated comments from the CF_S dataset and subsequently fine-tuned the model to classify all three stance classes on translated comments from the $CF_{E\text{-}D}$ dataset. Finally, Team Silver Surfer combined comment metadata features (e.g., number of upvotes/downvotes, endorsements) and the output probabilities from six fine-tuned transformer models in an XGBoost classifier (Run 1): (1) RoBERTa fine-tuned on the $CF_{E\text{-}D}$ dataset (comments translated to English, same as Run 5), (2) XLM-R fine-tuned on the $CF_{E\text{-}D}$ dataset (no translation), (3) RoBERTa fine-tuned on the CF_S dataset (translation to English), (4) XLM-R fine-tuned on the CF_S dataset (no translation), (5) English BERT fine-tuned on the CF_S and $CF_{E\text{-}D}$ datasets (two-step fine-tuning, comments translated to English, same as Run 6), and (6) multilingual BERT fine-tuned on the CF_S and $CF_{E\text{-}D}$ datasets (two-step fine-tuning, no translation, analogous to Run 6).

Team *Queen of Swords* [48] submitted two valid runs to Subtask 1. Their runs used English and multilingual BERT models [27] that were fine-tuned on a combination of the labeled (CF_S and $CF_{E\text{-}D}$) and unlabeled datasets (CF_U). To derive labels for the CF_U dataset, the team first fine-tuned a BERT model only on the CF_S and $CF_{E\text{-}D}$ datasets and used the fine-tuned model to predict labels on the CF_U dataset. The final BERT model was trained on the predicted labels from the CF_U dataset and the ground-truth labels from the CF_S and $CF_{E\text{-}D}$ datasets. For training the English BERT model, all non-English texts were translated to English using the `deep-translator` Python package.[28]

The submitted approaches were evaluated using the macro-averaged F_1-score and the accuracy of classifying the correct stance for each comment–proposal pair. Table 8 shows the systems' classification effectiveness for the most common languages and across all languages. Due to the class imbalance of the test set (see the stance label distribution for $CF_{E\text{-}T}$ in Table 7), we use the overall F_1-score as the main measure for evaluation. None of the submitted runs outperformed the baseline (Cavalier) in the two subtasks. For cross-debate classification (Subtask 1), the best participant approach (Queen of Swords, Run 1) achieved a macro-averaged F_1-score of 41.7 slightly worse than the baseline (Cavalier Subtask 1, macro-avg. F_1: 59.3). With all data available for training (Subtask 2), the best participant-submitted run (Silver Surfer, Run 3) achieved an F_1 of 35.0 and was again outperformed by the baseline (Cavalier Subtask 2, macro-avg. F_1: 57.7).

Even though Silver Surfer had access to the full dataset, their results were almost consistently outperformed by the approaches submitted by Team Queen of Swords, even though that

[26] https://huggingface.co/xlm-roberta-large
[27] https://huggingface.co/bert-base-uncased and https://huggingface.co/bert-base-multilingual-uncased
[28] https://pypi.org/project/deep-translator/#google-translate-1

Table 8
Results of Task 4 (Multilingual Multi-Target Stance Classification) for two subtasks evaluated using macro-averaged F_1 (per language and overall, using ISO 3166-1 alpha-2 country codes) and overall accuracy (Acc.). Sorted by overall F_1. Run IDs from the TIRA leaderboard included for reference. The Cavalier baseline is shown in bold.

		F_1 macro							Acc.
Team/Run	**Run ID**	**en**	**fr**	**de**	**it**	**hu**	**el**	**All**	
Subtask 1: Cross-Debate Classification									
Cavalier Subtask 1	—	**59.4**	**54.9**	**54.6**	**54.9**	**52.8**	**54.2**	**57.7**	**63.0**
Queen of Swords	`2023-05-19-07-51-03`	44.8	41.3	34.5	37.7	40.5	38.9	41.7	60.5
Queen of Swords	`2023-05-19-07-51-35`	35.1	31.5	26.2	40.9	43.0	35.7	32.4	61.6
Cavalier Simple	—	**24.4**	**24.2**	**20.3**	**25.1**	**29.3**	**17.1**	**23.7**	**55.2**
Subtask 2: All-Data-Available Classification									
Cavalier Subtask 2	—	**57.2**	**54.6**	**58.8**	**68.5**	**50.9**	**56.6**	**59.3**	**67.3**
Silver Surfer Run 6	`2023-05-12-18-56-56`	36.7	33.9	30.2	37.8	38.0	33.3	35.0	55.1
Silver Surfer Run 4	`2023-05-12-18-40-25`	35.3	30.4	26.1	35.3	34.8	27.8	32.9	53.7
Silver Surfer Run 1	`2023-05-12-17-49-58`	35.0	30.3	20.0	37.5	41.7	25.0	32.3	52.4
Silver Surfer Run 5	`2023-05-12-18-51-42`	28.5	25.6	24.3	32.9	21.5	22.8	27.0	46.3
Silver Surfer Run 2	`2023-05-12-18-29-46`	26.3	21.1	18.9	19.1	30.0	23.3	23.9	46.1
Silver Surfer Run 3	`2023-05-12-18-30-25`	41.4	23.2	21.2	14.1	22.8	32.8	17.7	21.6

team used less training data. The most difficult language for the baselines was Hungarian, which is the most morpho-syntactically distant from the other languages. Conversely, the participant approaches had the lowest scores when classifying German comments and did not have the same drop in effectiveness for Hungarian. Another language-specific difference can be observed for the runs of Team Silver Surfer: While the effectiveness of most of their runs was homogeneous between the languages, their Run 3 achieved a higher F_1-score for English compared to other languages, which might indicate a bias in the used multilingual XLM-R model. It is also interesting that the Cavalier baseline for Subtask 2 yielded better scores for Italian comments than for any other language, even though most of the other runs worked better on English comments. However, we could not observe patterns regarding the use of multilingual transformer models or English models with translation before classification. Both approaches seemed to work equally well.

The best runs of Subtask 2 (our baseline and Silver Surfer Run 2) used a two-step fine-tuning setting, where the model was first trained to learn binary stance classification and subsequently fine-tuned to three stance labels (including neutral). The best participant run of Subtask 1 also used two-stage fine-tuning, though not varying the number of labels to classify. These results indicate that breaking down the stance classification into several (easier-to-learn) steps can improve its effectiveness.

7. Conclusion

The fourth edition of the Touché lab on argument and causal retrieval featured four tasks: (1) argument retrieval for controversial topics, (2) causal retrieval, (3) image retrieval for arguments, and (4) multilingual multi-target stance classification. In contrast to the prior iterations of the Touché lab, the main challenge for the participant was to apply argument analysis methodology on long web documents from the ClueWeb22-B dataset. Furthermore, we expanded the lab's scope by introducing novel tasks that aimed to retrieve evidence for causal relationships and predict the stance of multilingual texts.

Out of the 41 registered teams, 7 participated in the tasks and submitted a total of 29 runs. The participants often used well-performing techniques from previous Touché iterations, such as using sparse retrieval for initial results and further re-ranking based on argument quality estimation and stance prediction, but now increasingly used generative language models like ChatGPT with various prompt-engineering techniques. All teams that participated in Tasks 1 and 2 used the search engine ChatNoir as their first-stage retrieval model, and then re-ranked documents based on the predicted argument quality and stance (Task 1), and based on the presence of causal relationships (Task 2). For Task 3, each of the top 4 most effective runs employed CLIP embeddings to find similar images to some text, which means dense retrieval approaches outperformed traditional approaches this year. However, still, none of the approaches were able to predict an image's stance better than random. To predict the stance of multilingual texts (Task 4), participants used BERT-based models and the most successful runs employed a two-step fine-tuning (first binary stance, then learn the neutral label). Overall, stance prediction remained the hardest problem across all four tasks in this edition.

We plan to continue Touché as a collaborative platform for researchers in argument retrieval. All Touché resources are freely available, including topics, manual relevance, argument quality, and stance judgments, and submitted runs from participating teams. These resources and other events such as workshops will continue supporting the research community working on argument and causal retrieval and analysis.

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. This work has been partially supported by the OpenWebSearch.eu project (funded by the EU; GA 101070014).

References

[1] A. Bondarenko, M. Fröbe, J. Kiesel, F. Schlatt, V. Barriere, B. Ravenet, L. Hemamou, S. Luck, J. Reimer, B. Stein, M. Potthast, M. Hagen, Overview of Touché 2023: Argument and causal retrieval, in: Proceedings of CLEF 2023, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023.

[2] I. Ajzen, The social psychology of decision making, in: Social Psychology: Handbook of Basic Principles, Guilford Press, 1996, pp. 297–325.

[3] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: Proceedings of TREC 1994, volume 500-225 of *NIST Special Publication*, NIST, 1994, pp. 109–126.

[4] S. E. Robertson, H. Zaragoza, M. J. Taylor, Simple BM25 extension to multiple weighted fields, in: Proceedings of CIKM 2004, ACM, 2004, pp. 42–49. doi:10.1145/1031171.1031181.

[5] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Proceedings of NeurIPS 2021, NeurIPS, 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

[6] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified data wrangling with `ir_datasets`, in: Proceedings of SIGIR 2021, ACM, 2021, pp. 2429–2436. doi:10.1145/3404835.3463254.

[7] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous integration for reproducible shared tasks with TIRA.io, in: Proceedings of ECIR 2023, Lecture Notes in Computer Science, Springer, 2023, pp. 236–241.

[8] A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2020: Argument retrieval, in: Working Notes of CLEF 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_261.pdf.

[9] A. Bondarenko, L. Gienapp, M. Fröbe, M. Beloucif, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2021: Argument retrieval, in: Working Notes of CLEF 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 2258–2284. URL: https://ceur-ws.org/Vol-2936/paper-205.pdf.

[10] A. Bondarenko, M. Fröbe, J. Kiesel, S. Syed, T. Gurcke, M. Beloucif, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2022: Argument retrieval, in: Working Notes of CLEF 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 2867–2903. URL: https://ceur-ws.org/Vol-3180/paper-247.pdf.

[11] A. Overwijk, C. Xiong, J. Callan, ClueWeb22: 10 billion web documents with rich information, in: Proceedsings of SIGIR 2022, ACM, 2022, pp. 3360–3362. doi:10.1145/3477495.3536321.

[12] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Elastic ChatNoir: Search engine for the ClueWeb and the Common Crawl, in: Proceedings of ECIR 2018, volume 10772 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 820–824. doi:10.1007/978-3-319-76941-7_83.

[13] A. Chernodub, O. Oliynyk, P. Heidenreich, A. Bondarenko, M. Hagen, C. Biemann, A. Panchenko, TARGER: Neural argument mining at your fingertips, in: Proceedings of ACL 2019, ACL, 2019, pp. 195–200. doi:10.18653/v1/p19-3031.
[14] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, PyTerrier: Declarative experimentation in Python from BM25 to dense retrieval, in: Proceedings of CIKM 2021, ACM, 2021, pp. 4526–4533. doi:10.1145/3459637.3482013.
[15] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, B. Stein, Computational argumentation quality assessment in natural language, in: Proceedings of EACL 2017, ACL, 2017, pp. 176–187. doi:10.18653/v1/e17-1017.
[16] J. R. M. Palotti, H. Scells, G. Zuccon, TrecTools: an open-source Python library for information retrieval practitioners involved in TREC-like campaigns, in: Proceedings of SIGIR 2019, ACM, 2019, pp. 1325–1328. doi:10.1145/3331184.3331399.
[17] G. V. Cormack, M. D. Smucker, C. L. A. Clarke, Efficient and effective spam filtering and re-ranking for large web datasets, Information Retrieval Journal 14 (2011) 441–465. doi:10.1007/s10791-011-9162-z.
[18] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, arXiv (2022). doi:10.48550/arXiv.2210.11416.
[19] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of ACL 2020, ACL, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.
[20] M. Plenz, R. Buchmüller, A. Bondarenko, Argument quality prediction for ranking documents, in: Working Notes of CLEF 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.
[21] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: Proceedings of ACL 2020, ACL, 2020, pp. 101–108. doi:10.18653/v1/2020.acl-demos.14.
[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: Proceedings of NeurIPS 2017, NeurIPS, 2017, pp. 3146–3154. URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
[23] H. Su, W. Shi, J. Kasai, Y. Wang, Y. Hu, M. Ostendorf, W. Yih, N. A. Smith, L. Zettlemoyer, T. Yu, One embedder, any task: Instruction-finetuned text embeddings, arXiv (2022). doi:10.48550/arXiv.2212.09741.
[24] A. Bondarenko, M. Wolska, S. Heindorf, L. Blübaum, A.-C. N. Ngomo, B. Stein, P. Braslavski, M. Hagen, M. Potthast, CausalQA: A benchmark for causal question answering, in: Proceedings of COLING 2022, ICCL, 2022, pp. 3296–3308. URL: https://aclanthology.org/2022.coling-1.291.
[25] S. Heindorf, Y. Scholten, H. Wachsmuth, A.-C. Ngonga Ngomo, M. Potthast, CauseNet: Towards a causality graph extracted from the web, in: Proceedings of CIKM 2020, ACM, 2020, pp. 3023–3030. doi:10.1145/3340531.3412763.

[26] A. Gaden, B. Reinhold, L. Zeit-Altpeter, N. Rausch, Evidence retrieval for causal questions using query expansion and reranking, in: Working Notes of CLEF 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[27] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT 2019, ACL, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.

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

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

[30] X. Xie, J. Mao, Y. Liu, M. de Rijke, Y. Shao, Z. Ye, M. Zhang, S. Ma, Grid-based evaluation metrics for web image search, in: Proceedings of WWW 2019, ACM, 2019, pp. 2103–2114. doi:10.1145/3308558.3313514.

[31] M. L. Carnot, L. Heinemann, J. Braker, T. Schreieder, J. Kiesel, M. Fröbe, M. Potthast, B. Stein, On stance detection in image retrieval for argumentation, in: Proceedings of SIGIR 2023, ACM, 2023. doi:10.1145/3539618.3591917.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of ICML 2021, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[33] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[34] M. Möbius, M. Enderling, S. Bachinger, Jean-Luc Picard at Touché 2023: Comparing image generation, stance detection and feature matching for image retrieval for arguments, in: Working Notes of CLEF 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of CVPR 2022, IEEE, 2022, pp. 10674–10685. doi:10.1109/CVPR52688.2022.01042.

[36] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

[37] D. Elagina, B.-A. Heizmann, M. Koch, G. Lahmann, C. Ortlepp, Neville Longbottom at Touché 2023: Image retrieval for arguments using ChatGPT, CLIP and IBM Debater, in: Working Notes of CLEF 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[38] R. Bar-Haim, Y. Kantor, E. Venezian, Y. Katz, N. Slonim, Project debater APIs: Decomposing the AI grand challenge, in: Proceedings of EMNLP 2021, ACL, 2021, pp. 267–274. doi:10.18653/v1/2021.emnlp-demo.31.

[39] V. Barriere, A. Balahur, B. Ravenet, Debating Europe: A multilingual multi-target stance classification dataset of online debates, in: Proceedings of PoliticalNLP 2022, ELRA, 2022, pp. 16–21. URL: https://aclanthology.org/2022.politicalnlp-1.3.

[40] J. Vamvas, R. Sennrich, X-stance: A multilingual multi-target dataset for stance detection, in: Proceedings of SwissText/KONVENS 2020, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2624/paper9.pdf.

[41] V. Barriere, G. Jacquet, L. Hemamou, CoFE: A new dataset of intra-multilingual multi-

target stance classification from an online european participatory democracy platform, in: Proceedings of AACL-IJCNLP 2022, 2022, pp. 418–422.

[42] V. Barriere, A. Balahur, Multilingual multi-target stance recognition in online public consultations, Mathematics 11 (2023) 2161.

[43] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of ACL 2020, ACL, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[44] J. P. Avila, A. Rodrigo, R. Centeno, Silver Surfer team at Touché task 4: Testing data augmentation and label propagation for multilingual stance detection, in: Working Notes of CLEF 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[45] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv (2019). doi:10.48550/arXiv.1907.11692.

[46] A. Sugiyama, N. Yoshinaga, Data augmentation using back-translation for context-aware neural machine translation, in: Proceedings of DiscoMT@EMNLP 2019, ACL, 2019, pp. 35–44. doi:10.18653/v1/D19-6504.

[47] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: Proceedings of NIPS 2003, MIT Press, 2003, pp. 321–328. URL: https://proceedings.neurips.cc/paper/2003/hash/87682805257e619d49b8e0dfdc14affa-Abstract.html.

[48] K. Schaefer, Queen of Swords at Touché 2023: Intra-multilingual multi-target stance classification using BERT, in: Working Notes of CLEF 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

A. Zero-shot Prompts

The zero-shot prompts used for the stance prediction baselines are given in Listing 1 (for Task 1, see Section 3) and in Listing 2 (for Task 2, see Section 4).

```
Given a query, predict the stance of a given text. The stance should be one of the following
     four labels:
PRO: The text contains opinions or arguments in favor of the query "<query>".
CON: The text contains opinions or arguments against the query "<query>".
NEU: The text contains as many arguments in favor of as it contains against the query "<query
     >".
UNK: The text is not relevant to the query "<query>", or it only contains factual information
     .

Text: <summary>
```

Listing 1: Zero-shot prompt to predict the stance of a document towards a query (Task 1). The placeholder `<query>` is replaced by the topic titles, and `<summary>` for a short summary of the retrieved document's text. The `UNK` label is mapped to `NO`.

```
Given a query, predict the stance of a given text. The stance should be one of the following
     four labels:
SUP: According to the text, <cause> causes <effect>.
REF: According to the text, <cause> does not cause <effect>.
UNK: The text is not relevant to <cause> and <effect>.

Text: <summary>
```

Listing 2: Zero-shot prompt to predict the causal stance of a document towards a query (Task 2). The placeholders `<cause>` and `<effect>` are replaced with the query's cause and effect entities, and `<summary>` with a short summary of the retrieved document's text. The `UNK` label is mapped to `NO`. The `NEU` label is not considered in the prompt.

B. Full Evaluation Results of Touché 2023: Argument and Causal Retrieval

Table 9
Relevance results of all runs submitted to Task 1: Argument Retrieval for Controversial Questions. Reported are the mean nDCG@10 and the 95% confidence intervals. The baseline Puss in Boots is shown in bold.

Team	Run Tag	nDCG@10		
		Mean	Low	High
Puss in Boots	**ChatNoir [12]**	**0.834**	**0.791**	**0.875**
Renji Abarai	stance_ChatGPT	0.747	0.687	0.812
Renji Abarai	stance-certainNO_ChatGPT	0.746	0.678	0.810
Renji Abarai	ChatGPT_mmGhl	0.718	0.653	0.775
Renji Abarai	ChatGPT_mmEQhl	0.718	0.650	0.779
Renji Abarai	meta_qual_score	0.712	0.641	0.782
Renji Abarai	baseline	0.708	0.632	0.775
Renji Abarai	meta_qual_prob	0.697	0.622	0.765

Table 10
Quality results of all runs submitted to Task 1: Argument Retrieval for Controversial Questions. Reported are the mean nDCG@10 and the 95% confidence intervals. The baseline Puss in Boots is shown in bold.

Team	Run Tag	nDCG@10		
		Mean	Low	High
Puss in Boots	**ChatNoir [12]**	**0.831**	**0.786**	**0.873**
Renji Abarai	stance_ChatGPT	0.815	0.764	0.862
Renji Abarai	stance-certainNO_ChatGPT	0.811	0.754	0.863
Renji Abarai	ChatGPT_mmEQhl	0.789	0.730	0.846
Renji Abarai	ChatGPT_mmGhl	0.789	0.731	0.842
Renji Abarai	meta_qual_prob	0.774	0.712	0.830
Renji Abarai	meta_qual_score	0.771	0.710	0.832
Renji Abarai	baseline	0.766	0.698	0.823

Table 11
Relevance results of all runs submitted to Task 2: Evidence Retrieval for Causal Questions. Reported are the mean nDCG@5 and the 95% confidence intervals. The baseline Puss in Boots is shown in bold.

Team	Run Tag	nDCG@5		
		Mean	Low	High
He-Man	no_expansion_rerank	0.657	0.564	0.740
Puss In Boots	**ChatNoir [12]**	**0.585**	**0.503**	**0.673**
He-Man	gpt_expansion_rerank	0.374	0.284	0.469
He-Man	causenet_expansion_rerank	0.268	0.172	0.368

Table 12
On-topic relevance results of all runs submitted to Task 3: Image Retrieval for Argumentation. Reported are the mean precision@10 and the 95% confidence intervals. The baseline Minsc is shown in bold.

Team	Run Tag	Precision@10		
		Mean	Low	High
Neville Longbottom	clip_chatgpt_args.raw	0.785	0.714	0.852
Hikaru Sulu	Keywords	0.770	0.704	0.831
Neville Longbottom	clip_chatgpt_args.debater	0.684	0.601	0.764
Hikaru Sulu	Topic-title	0.664	0.581	0.746
Neville Longbottom	bm25_chatgpt_args.raw	0.572	0.510	0.636
Jean-Luc Picard	BM25 Baseline	0.536	0.458	0.608
Jean-Luc Picard	No stance detection	0.523	0.442	0.598
Jean-Luc Picard	Text+image text stance detection	0.502	0.429	0.573
Jean-Luc Picard	Text stance detection	0.498	0.419	0.567
Neville Longbottom	bm25_chatgpt_args.diff	0.442	0.378	0.507
Neville Longbottom	bm25_chatgpt_args.debater	0.416	0.350	0.481
Minsc	**Aramis**	**0.376**	**0.310**	**0.442**
Jean-Luc Picard	Image text stance detection	0.369	0.301	0.433

Table 13

Argumentativeness results of all runs submitted to Task 3: Image Retrieval for Argumentation. Reported are the mean precision@10 and the 95% confidence intervals. The baseline Minsc is shown in bold.

Team	**Run Tag**	**Precision@10**		
		Mean	Low	High
Hikaru Sulu	Topic-title	0.350	0.291	0.415
Neville Longbottom	clip_chatgpt_args.debater	0.341	0.271	0.410
Neville Longbottom	clip_chatgpt_args.raw	0.338	0.273	0.404
Hikaru Sulu	Keywords	0.335	0.275	0.395
Jean-Luc Picard	No stance detection	0.292	0.220	0.367
Neville Longbottom	bm25_chatgpt_args.raw	0.274	0.211	0.338
Jean-Luc Picard	Text+image text stance detection	0.272	0.208	0.339
Jean-Luc Picard	BM25 Baseline	0.268	0.198	0.334
Jean-Luc Picard	Text stance detection	0.262	0.198	0.325
Neville Longbottom	bm25_chatgpt_args.diff	0.240	0.176	0.309
Neville Longbottom	bm25_chatgpt_args.debater	0.201	0.146	0.263
Jean-Luc Picard	Image text stance detection	0.196	0.149	0.247
Minsc	**Aramis**	**0.194**	**0.144**	**0.248**

Table 14

Stance relevance results of all runs submitted to Task 3: Image Retrieval for Argumentation. Reported are the mean precision@10 and the 95% confidence intervals. The baseline Minsc is shown in bold.

Team	**Run Tag**	**Precision@10**		
		Mean	Low	High
Neville Longbottom	clip_chatgpt_args.raw	0.222	0.174	0.268
Neville Longbottom	clip_chatgpt_args.debater	0.216	0.155	0.281
Hikaru Sulu	Topic-title	0.185	0.149	0.221
Hikaru Sulu	Keywords	0.179	0.140	0.219
Neville Longbottom	bm25_chatgpt_args.raw	0.166	0.127	0.208
Jean-Luc Picard	No stance detection	0.162	0.118	0.206
Neville Longbottom	bm25_chatgpt_args.diff	0.150	0.108	0.196
Jean-Luc Picard	Text+image text stance detection	0.144	0.108	0.185
Jean-Luc Picard	BM25 Baseline	0.141	0.105	0.183
Jean-Luc Picard	Text stance detection	0.136	0.101	0.177
Neville Longbottom	bm25_chatgpt_args.debater	0.128	0.091	0.170
Minsc	**Aramis**	**0.102**	**0.076**	**0.129**
Jean-Luc Picard	Image text stance detection	0.098	0.067	0.132