Overview of Touché 2022: Argument Retrieval

Extended Version*

Alexander Bondarenko[1], Maik Fröbe[1], Johannes Kiesel[2], Shahbaz Syed[3], Timon Gurcke[4], Meriem Beloucif[5], Alexander Panchenko[6], Chris Biemann[7], Benno Stein[2], Henning Wachsmuth[4], Martin Potthast[3] and Matthias Hagen[1]

[1]*Martin-Luther-Universität Halle-Wittenberg*
[2]*Bauhaus-Universität Weimar*
[3]*Leipzig University*
[4]*Paderborn University*
[5]*Uppsala University*
[6]*Skolkovo Institute of Science and Technology*
[7]*Universität Hamburg*
touche@webis.de https://touche.webis.de

Abstract

This paper is a report on the third year of the Touché lab on argument retrieval hosted at CLEF 2022. With the goal of supporting and promoting the research and development of new technologies for argument mining and argument analysis, we have organized three shared tasks: (a) argument retrieval for controversial topics, where the task is to find sentences that reflect the gist of arguments from online debates, (b) argument retrieval for comparative issues, where the task is to find argumentative passages from web documents that help in making a comparative decision, and (c) image retrieval for arguments, where the task is to find images that show support for or opposition to a particular stance.

Keywords

Argument retrieval, Controversial questions, Comparative questions, Image retrieval, Shared task

1. Introduction

Decision-making and opinion-forming are everyday tasks, often involving weighing pro and con arguments for or against different options. Considering the many arguments on almost any topic on the web, in principle anyone can come to an informed decision or opinion with the help of a search engine. However, large parts of the easily accessible arguments on the web are of low quality. They may contain incoherent logic, fail to substantiate a claim, or use inappropriate language. These arguments should not appear at the top of search results—regardless of whether a query is about socially important issues or "only" personal choices. Challenges arising from this observation range from evaluating the relevance of an argument to a query and assessing how well an implied stance is justified, to identifying the gist of an argument, to finding images that illustrate a particular stance. Commercial web search engines do not sufficiently address these challenges—a gap we aim to fill with the Touché labs.

*This overview extends the one published as part of the CLEF 2022 proceedings [1]. 'Touché' is commonly "used to acknowledge a hit in fencing or the success or appropriateness of an argument" (merriam-webster.com)

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

Following the two successful Touché labs on argumentation at CLEF 2020 and 2021 [2, 3], our third lab edition again brought together researchers from the fields of information retrieval and natural language processing who study argumentation. At Touché 2022, we have organized the following three shared tasks, the last of which is a completely new addition:

1. Argumentative sentence retrieval from a focused collection (crawled from debate portals) to support conversations about controversial topics.
2. Argument retrieval from a large collection of text passages to support answering comparative questions in personal decision making.
3. Argumentative image retrieval to support the illustration of arguments and getting an overview of the public opinion on controversial topics.

Touché follows the traditional TREC methodology: documents and topics are provided to participants, who then submit their results (up to five runs) for each topic to be assessed by human assessors. While the first two Touché editions focused on full argument and document retrieval, the third edition focused on more fine-grained retrieval units. The three shared tasks investigated whether argument retrieval can more directly support decision making and opinion formation by extraction of the gist of documents, classification of their stance on an issue as pro or con, and retrieval of images that support or oppose a particular stance.

The teams that participated in the third Touché lab were able to use the topics and assessments (relevance and quality of arguments) from the previous lab editions to train and optimize their approaches. In addition to traditional retrieval models such as BM25 [4], re-ranking approaches such as the recent transformer-based models T5 [5] and T0 [6] have been applied with the goal of combining topical relevance with "argumentativeness," argument quality, or stance. They are an essential part of the most effective approaches of all three Touché tasks, confirming the general trend in information retrieval and natural language processing that pre-trained transformers achieve good effectiveness [7] (cf. Sections 4—6). The most effective approach submitted to Task 1 re-ranks the DirichletLM model's search results by first using a BERT-based classifier [8] to decide on the argumentativeness of retrieved sentence pairs (i.e., whether they are premises or assertions), then estimating their coherence using the cosine similarity of their BERT embeddings. For Task 2, in terms of relevance, a TCT-ColBERT ranker [9] and, in terms of quality, a combination of query-dependent BM25F scores [10] and predicted argument quality were most effective. The most effective approach for Task 3 (across topic relevance, argumentativeness, and stance relevance) used BERT instead of a stance detection model to detect the sentiment of texts from web pages and texts in images and indexed both with BM25F.

Altogether, the most effective argument retrieval approaches used various strategies for query reformulation and expansion, and for re-ranking based on estimates of argument quality or "argumentativeness". Sentiment or emotion recognition was particularly useful for the argumentative image retrieval task, as well as OCR to retrieve image text for analysis.

The corpora, topics, and judgments created at Touché are freely available to the research community and can be found on the lab's website.[1] Parts of the data are also already available via the BEIR [11] and `ir_datasets` [12] resources.

[1] https://webis.de/events.html?q=Touche#shared-tasks

2. Related Work

Queries in argument retrieval often are phrases that describe a controversial topic, questions that ask to compare two options, or even complete claims or short arguments [13]. In the third edition of the Touché lab, we address the first two query types in three different shared tasks on argument retrieval in general, on comparative scenarios, and on image retrieval. Here, we briefly summarize the related work for all three tasks.

2.1. Argument Retrieval

The goal of argument retrieval is to find arguments that help when making a decision, when forming an opinion, or when trying to convince (or persuade) someone of a specific point of view. An argument is usually modeled as a conclusion with one or more supporting or attacking premises [14]. While a conclusion is a statement that can be accepted or rejected, a premise is a more grounded statement (e.g., statistical evidence or a referenced quote).

Adding argument retrieval components to a search engine poses challenges like identifying argumentative queries [15], mining arguments from documents, or assessing an argument's relevance and quality [14]. Different paradigms have been proposed for actual argument retrieval that perform argument mining and ranking in different order [16]. For instance, Wachsmuth et al. [14] use distant supervision and extract and index arguments from debate portals in a "pre-processing". Their argument search engine args.me[2] uses BM25F [10] to then only rank the extracted arguments at query time, giving more weight to conclusions than premises. Also Levy et al. [17] use distant supervision to mine arguments from Wikipedia in an offline pre-processing before ranking. Following a different paradigm, Stab et al. [18] retrieve documents from the Common Crawl[3] at query time (no prior offline argument mining) and use a topic-dependent neural network to then extract arguments from the retrieved documents. In our Touché tasks, we address both paradigms, the one of Wachsmuth et al. [14] in Task 1 (retrieval from a focused collection of pre-processed arguments) and the one of Stab et al. [18] in Task 2 (retrieval from some general collection with online argument mining).

Argument retrieval should take topical relevance into account but also argument quality. What makes a good argument has been studied since the time of Aristotle [19]. Wachsmuth et al. [20] categorize the different aspects of argument quality into a taxonomy that covers three dimensions: logic, rhetoric, and dialectic. Logic concerns the strength of the internal structure of an argument (i.e., the conclusion and the premises along with their relations) while rhetoric covers the effectiveness of an argument in persuading an audience with its conclusion. Lastly, dialectic addresses the relations of an argument to other arguments on the topic. For example, an argument attacked by many others may be rather vulnerable in a debate. Note that an argument's relevance to a query is also categorized under dialectical quality [20].

Argument relevance has been typically assessed by an argument's similarity to a given topic and by incorporating the support and attack relations to other arguments. Potthast et al. [21] evaluate four standard retrieval models for ranking arguments with regard to topical relevance, logic, rhetoric, and dialectic. One of the main findings is that DirichletLM is better at

[2] https://www.args.me/
[3] http://commoncrawl.org

ranking arguments than BM25, DPH, and TF-IDF. Gienapp et al. [22] later proposed a pairwise annotation strategy that reduces the costs of crowdsourcing argument retrieval annotations by 93% (i.e., requiring the annotation of only a rather small subset of argument pairs).

As for argument ranking, several approaches exploit argument relations. For instance, Wachsmuth et al. [23] connect two arguments in a graph when one uses the other's conclusion as a premise and then compute an argument's PageRank [24] in this graph. In their study, taking PageRank into account improves upon baselines that only use an argument's content and internal structure (conclusion and premises) [23]. Later, Dumani et al. [25] used support and attack relations between clusters of premises and claims as well as between clusters of claims and a query. In an extended version, Dumani and Schenkel [26] also include the quality of a premise as a probability (fraction of premises that are worse with regard to cogency, reasonableness, and effectiveness). Using a pairwise quality estimator trained on the Dagstuhl-15512 ArgQuality Corpus [27], the approach with the argument quality component was more effective on the 50 topics of Task 1 from Touché 2020 than the one without taking argument quality into account.

2.2. Retrieval for Comparisons

Comparative information needs in web search have first been addressed with basic interfaces for comparing two products entered separately in two search boxes [28, 29]. Using opinion mining approaches, comparative sentences can then be identified from product reviews in favor of or against one or the other product [30, 31, 32]. Recently, identifying a comparison preference in a sentence (i.e., the "winning" option) has also been tackled more broadly (not just for product reviews) [33, 34] and forms the basis of the comparative argumentation machine CAM [35]. Similar to the early comparison interfaces, CAM takes two objects and some comparison aspect(s) as input, retrieves comparative sentences in favor of one or the other option using BM25, and then classifies the sentences' preferences for a final merged table-like result presentation. A proper argument ranking, however, was not included in CAM. Chekalina et al. [36] later extended the system to accept complete comparative questions as input and to return a natural language answer. From a comparative question, the comparison objects, aspect(s), and predicates are extracted and the system's answer is either generated directly based on transformers [8] or by retrieval from an index of comparative sentences. To identify comparative questions and information needs, Bondarenko et al. [37, 38] propose a cascading ensemble of classifiers (rule-based, feature-based, and neural models). They also propose improved approaches to extract the comparison objects, aspects, and predicates from comparative questions and to detect the stance of potential answers towards the comparison objects. The respective stance dataset could also be used by the participants of our Task 2.

2.3. Image Retrieval

Images can provide contextual information and express, underline, or popularize an opinion [39], thereby taking the form of subjective statements [40]. While some images can be complete arguments (i.e., expressing both, a premise and a conclusion) [41, 42] others provide contextual information only and have to be combined with a textual conclusion to form an argument. A recent SemEval task distinguished a total of 22 persuasion techniques in memes alone [43].

Moreover, argument quality dimensions like acceptability, credibility, emotional appeal, and sufficiency [27] all also apply to arguments that include images.

Pre-dated only by approaches relying on metadata and similarity measures [44], the actual content of images or videos has been analyzed and used for keyword-based image search for decades [45]. In a recent survey, Latif et al. [46] categorize image features into color, texture, shape, and spatial features but commercial search engines also index text found in images, surrounding text, alternative texts displayed when an image is unavailable, and the image URLs [47, 48]. As for the retrieval of argumentative images, a closely related concept is "emotional images", which is based on image features like color and composition [49, 50]. Since argumentation often goes hand in hand with emotions, emotional features may also be promising for retrieving images for arguments, a relatively new task recently proposed by Kiesel et al. [51] and now forming Task 3 of the Touché 2022 lab.

3. Lab Overview and Statistics

For the third edition of the Touché lab, we received 58 registrations, doubling the number from the previous year (29 registrations in 2021). Among the teams, 27 registered for more than one task, 17 registered particularly for Task 1, 10 for Task 2, and 4 for Task 3 (the new task this year). The majority of registrations came from Germany and Italy (13 each), followed by India (12), the United States (3), the Netherlands, France, Switzerland, Bangladesh (2 each), Pakistan, Portugal, United Kingdom, Indonesia, China, Russian Federation, Bulgaria, Nigeria, and Lebanon (1 each). Aligned with the lab's fencing-related title, the registered teams selected a real or fictional fencer or swordsman character (e.g., D'Artagnan) as their team name.

From the 58 registered teams, 23 actively participated in the tasks and submitted results[4] (27 teams submitted in 2021 and 17 teams in 2020). Using the setup of the previous Touché editions, we encouraged the teams to deploy their software in TIRA [52] for a better reproducibility of the developed approaches. The TIRA integrated research architecture is cloud-based evaluation-as-a-service platform where shared task participants can deploy their software in a dedicated virtual machine to which they have full administrative access. By default, the virtual machines run the server version of Ubuntu 20.04 with one Intel Xeon E5-2620 CPU, 4 GB RAM, 16 GB HDD, and the latest versions of often-used software packages pre-installed (e.g., Docker and Python). If needed, we tried to customize the resources as per a team's requirements. Providing GPUs was not possible, though.

For teams that did not deploy their software in TIRA, we allowed run submissions similar to many TREC tracks. In case they preferred software submissions, the teams created their run using via web UI of TIRA by remote-executing their software inside their virtual machine. The software is fully installed in the virtual machine, and at execution time the virtual machine is shut down, disconnected from the internet, powered on again in a sandbox mode, and the test datasets for the respective tasks are mounted. Interrupting the internet connection ensures that the participants' software works without external web services that may disappear or become incompatible, which could reduce reproducibility (i.e., downloading additional external code or models during the execution is not possible). We offered support in case of problems

[4] Three teams did not submit a paper describing their approach, though.

during deployment and then archived the virtual machines that the participants used for their submissions. The respective systems can thus be re-evaluated or also applied to new datasets with the same input format.

Overall, 9 of the 23 teams submitted traditional run instead of deploying their software in TIRA. Per team, we allowed 5 runs and the run needed to follow the standard TREC format.[5] We checked the validity of each submitted run and asked participants to rerun their software or resubmit their files in case of problems while also offering support in such cases. In total, 84 runs were submitted—at least one from each team.

4. Task 1: Argument Retrieval for Controversial Questions

The goal of the Touché 2022 lab's 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?". Such scenarios benefit from obtaining the gists of various web resources that briefly summarize different stances (pro or con) on controversial topics. The task we considered in this regard followed the idea of extractive argument summarization [53].

4.1. Task Definition and Data

Task. Given a controversial topic and a collection of arguments, the task was to retrieve sentence pairs that represent the gist of their corresponding arguments (e.g., the main claim and a supporting premise). Sentences in such a pair may not contradict each other and ideally build upon each other in a logical manner comprising a coherent text.

Topics. We used 50 controversial topics from the previous iterations of Touché. Each topic is formulated as a question that the user might pose as a query to the search engine, accompanied by a description summarizing the information need and the search scenario, along with a narrative to guide assessors in recognizing relevant results (see Table 1).

Document collection. The document collection for Task 1 was based on the args.me corpus [16] which contains about 400,000 structured arguments (crawled from the online debate portals debatewise.org, idebate.org, debatepedia.org, and debate.org). It is freely available for download[6] and can also be accessed through the args.me API.[7] To account for this year's changes in the task definition (the focus on gists), we prepared a pre-processed version of the corpus. Preprocessing steps included sentence splitting and removing premises and conclusions shorter than two words, resulting in 5,690,642 unique sentences with 64,633 claims and 5,626,509 premises.

[5] The expected format was also described at the lab's web page: https://webis.de/events/touche-22/
[6] https://webis.de/data.html#args-me-corpus
[7] https://www.args.me/api-en.html

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.

4.2. Evaluation Setup

Participants submitted their rankings as traditional TREC-style runs where document IDs are sorted by descending relevance score for each search topic (i.e., the most relevant argument occurs at Rank 1). Given the large number of runs and the possibility of retrieving up to 1000 documents (in our case, these are sentence pairs) per topic in a run, using TrecTools [54], we created the pools using a top-5 pooling strategy, resulting in 6,930 unique sentence pairs for manual assessment of relevance, quality (argumentativeness), and textual coherence. Relevance was judged by our volunteer assessors on a three-point scale: 0 (not relevant), 1 (relevant), and 2 (highly relevant). For quality, annotators assessed whether a retrieved pair of sentences are rhetorically well-written on a three-point scale: 0 (low quality/non-argumentative), 1 (average quality), and 2 (high quality). Textual coherence (if the two sentences in a pair logically build upon each other) was also judged on a three-point scale: 0 (unrelated/contradicting), 1 (average coherence), and 2 (high coherence).

4.3. Submitted Approaches and Evaluation Results

This year's approaches included standard retrieval models such as TF-IDF, BM25, DirichletLM, and DPH. Participants also used third-party toolkits, such as the Project Debater API [55] (for stance and evidence detection in arguments), Apache OpenNLP[8] (for language detection), and BERT-based classifiers proposed by Reimers et al. [56] trained on the Webis Argument Quality Corpus [22] and the IBM Rank 30K dataset [57] for argument quality detection. Additionally, semantic similarity of word and sentence embeddings based on doc2vec [58], Spacy embeddings [59], and SBERT [60] have been employed for retrieving coherent sentence pairs as required by the task definition. One team leveraged the text generation capabilities of GPT-2 [61] to find subsequent sentences while another team similarly used the next sentence prediction (NSP) of BERT [8] for this. These toolkits augmented the document preprocessing and re-ranking of the retrieved results.

[8] https://opennlp.apache.org/

Table 2
Results of Task 1 (Argument Retrieval for Controversial Questions). Shown are the scores of a teams' best run for the three dimensions relevance, quality, and coherence of the retrieved sentence pairs with along a run's rank (results of all submitted runs in Tables 6–8). The teams are ordered alphabetically; baseline Swordsman emphasized. A † indicates statistically significant differences to the baseline (paired Student's t-test, $p = 0.05$, Bonferroni-correction).

Team	**nDCG@5**					
	Rank	Relevance	Rank	Quality	Rank	Coherence
Bruce Banner	3	0.651†	5	0.772†	4	0.378
D'Artagnan	4	0.642†	7	0.733†	5	0.378†
Daario Naharis	2	0.683†	1	0.913†	1	0.458†
Gamora	5	0.616†	3	0.785†	7	0.285
General Grevious	9	0.403	10	0.517	10	0.231
Gorgon	8	0.408	6	0.742†	8	0.282
Hit Girl	6	0.588†	4	0.776†	6	0.377
Korg	11	0.252	11	0.453†	11	0.168
Pearl	7	0.481	8	0.678	3	0.398†
Porthos	1	0.742†	2	0.873†	2	0.429†
Swordsman	10	**0.356**	9	**0.608**	9	**0.248**

We used nDCG@5 to evaluate of relevance, quality, and coherence. Table 2 shows the results of the best run per team. On all the evaluated dimensions at least eight out of ten teams managed to beat the provided baseline. Similar to previous years' results, quality is best covered by the approaches followed by relevance and the newly added coherence dimension.

Summarizing the results, for relevance, Team *Porthos* [62] achieved the highest rank followed by *Daario Naharis* [63] with nDCG@5 scores of 0.742 and 0.683, respectively. For the quality and coherence dimensions *Daario Naharis* obtained the highest scores (0.913 and 0.458) followed by *Porthos* (0.873 and+0.429). We believe that the two-stage re-ranking employed by *Daario Naharis* improved coherence and quality in comparison to the other approaches. They first ensured that retrieved pairs were relevant to their context in the argument alongside the topic which preserved high-quality arguments. Then, a second re-ranking based on stance to determine the final pairing of the retrieved sentences boosted coherence. Below, we briefly describe our baseline and summarize the submitted approaches.

Our baseline *Swordsman* employed a graph-based approach that ranks arguments' sentences by their centrality in the corresponding argument graph as proposed by Alshomary et al. [53]. The top two sentences per argument are used as the their gist. We retrieved 1000 pairs per topic.

Bruce Banner [64] employed the BM25 retrieval model implemented in the Pyserini toolkit [65] with its default parameters ($k_1 = 1.2$ and $b = 0.68$). For each argument, they indexed all possible sentence pairs. To speed up computation on such a large collection of sentence pairs, they specifically opted for the sparse representations in Pyserini that produce smaller indexes compared to the dense retrieval variants. Two query variants were used: original query (topic title) and an expanded query (narrative and description appended). Likewise, two variants of the sentence pairs were indexed: original pair and pair with the topic of a debate appended. They retrieved 1000 documents per query and did not apply any re-ranking.

D'Artagnan [66] also employed sparse retrieval together with text preprocessing and query expansion. For retrieval, they used two retrieval models from Lucene: BM25 [10] ($k_1 = 1.2$ and $b = 0.75$) and DirichletLM ($\mu = 2000$). For preprocessing, they experimented with both Porter [67] and Krovetz [68] stemmers. Additionally, they filtered both character and word n-grams (referred to as shingles) and used two stop word lists (SMART System [69], Glasgow IR.[9]) Query expansion was done using synonyms from WordNet [70] and word2vec [71]. Evaluation on the previous year's relevance judgments showed that a combination of the DirichletLM retrieval model, the Krovitz stemmer, and the Glasgow IR stop word list improved performance compared to their respective counterparts.

Daario Naharis [63] developed a standard Lucene-based document retrieval system using the TF-IDF model. Additionally, they introduced a new measure called *ICoefficient* for scoring the discriminant power of a term. This complements the standard TF-IDF weighting by additionally considering the number of documents that contain at least one occurrence of a given term. We refer readers to Bahrami et al. [63] for the mathematical formulation of the ICoefficient. For preprocessing, they created two custom stop lists, each composed of the 100 most frequent terms in the indexed collections of the argument contexts and individual arguments from the provided corpus. Document re-ranking was performed based on stance and evidence detection using the Project Debater API [55].

Gamora [72] developed Lucene-based approaches using deduplication and contextual feature-enriched indexing, adding the topic of a debate and the stance on the topic, to obtain document-level relevance and quality scores, following the approaches used in previous Touché editions [3]. To find relevant sentence pairs rather than relevant documents, these results were used to limit the number of documents by creating a new index for only the sentences of relevant documents (double indexing) or creating all possible sentence combinations and ranking them based on a weighted average of the argument quality (estimated using an SVM classifier) of the pair and its source document. BM25 [65] ($k_1 = 1.2$ and $b = 0.75$) and DirichletLM ($\mu = 2000$) were used for document similarity and SBERT [60] and TF-IDF for sentence similarity. The best approach is based on double indexing and a combination of a manual query reduction in which only the 2—6 main words of the query were kept, query boosting, query decorators, query expansion with respect to important keywords (GloVE [73]) and synonyms (WordNet [70]), and possessive removal, stemming (Krovetz stemmer [68]) and length filtering of the sentences.

General Grevious [74] used a conventional IR pipeline based on Lucene. First, documents were lowercased, tokenized and possessive words (with trailing ''s') were removed, keeping only tokens with a length between 3 and 20 characters. In addition, the team experimented with a variety of stemming approaches (S-stemmer [75], Krovetz stemmer [68], Porter stemmer [67], no stemming) and stop word lists (Core NLP [76], CountWordsFree [65], EBSCO,[10] GoogleStop,[11] and Ranks.[12]) To retrieve documents, BM25 [65] ($k_1 = 1.2$ and $b = 0.75$) and DirichletLM ($\mu \in \{1700, 1800\}$) were used together with query boosting, by assigning weights to the used inputs (argument, conclusion, debate title, and argument title), and query expansion, by finding

[9] https://github.com/igorbrigadir/stopwords/
[10] https://connect.ebsco.com/s/article/What-are-stop-words-and-how-does-EBSCO-s-search-engine-handle-them?
[11] https://www.semrush.com/blog/seo-stop-words/
[12] https://www.ranks.nl/stopwords

keywords (Rapid Automatic Keyword Extraction (RAKE) [77]) and synonyms (Datamuse[13]). This retrieval step was done once for the documents and once for all the potential sentence pairs within these retrieved documents to obtain a ranking of sentence pairs. Finally, sentiment analysis (Vader [78]) was used to boost documents that have a similar sentiment as the query, and readability analysis (Flesch-Kincaid [79]) was used for re-ranking. Their best model does not include re-ranking, stemming, and stop word removal but relies solely on the combination of query expansion and the BM25 retrieval model.

Gorgon [80] also used a Lucene-based IR pipeline and compared BM25 [65] ($k_1 = 1.2$ and $b = 0.75$) and DirichletLM ($\mu = 2000$) similarity measures, developing four different analyzers with different preprocessing steps including lowercasing, stemming (Krovetz stemmer [68]), removing possessive words (with trailing ''s') and filtering stop words (99webtools,[14] EBSCO). Sentence pairs were created from all combinations within a single document before indexing. The best approach is a combination of lowercasing, removing possessive words, and BM25.

Hit Girl [81] proposed a two-stage retrieval pipeline that combines semantic search and re-ranking via argument quality agnostic models. Documents were embedded to vectors using Spacy [59]. These were then indexed via Elasticsearch and its text similarity function used for semantic search. They experimented with three approaches for re-ranking: maximal marginal relevance [82], word mover's distance [83], and a novel method called structural distance which employs fuzzy matching between query and sentences based on POS tags. Preliminary evaluations showed that, while re-ranking improved the argument quality to varying degrees, it also affected relevance. Also, structural distance performed best for re-ranking.

Korg's [84] approaches are based on the Elasticsearch implementation of DirichletLM ($\mu = 2000$) to find the best matching argumentative sentences for a query after employing lowercasing, ASCII folding, stop word filtering (manually created stop word list) and stemming (Krovetz stemmer [68]). Then, either doc2vec [58] or SBERT [60] is trained on all sentences in the args.me corpus, which was used to find the most similar sentence pair within a document by direct comparison of the doc2vec embeddings. Alternatively, instead of directly comparing sentences, GPT-2 [61] was used to generate the next sentence for a given sentence to then find the most similar sentence to the generated sentence. The best approach is based on lowercasing, ASCII folding, stop word filtering, stemming, and doc2vec's similarity calculation without GPT-2.

Pearl [85] also proposed a two-stage retrieval pipeline using DirichletLM [86] and DPH [87] models to retrieve argumentative sentences. For both stages, they used the PyTerrier toolkit [88]. After retrieving the documents, two BERT-based argument quality models fine-tuned on the Webis Argument Quality Corpus [89], and the IBM-Rank-30k dataset [57] were employed to filter non-argumentative results. The resulting prototype from the first stage was considered the baseline model. On evaluating this on a set of 35 queries taken from the provided topics, they found that the DPH model assigned high relevance to sentences even if their terms are part of a URL, or other meta data in the corpus. Moreover, it was also susceptible to homonyms and thus negatively affecting the retrieval performance. To account for this, a refined prototype was developed that combined argument quality prediction with query expansion. For query expansion, they applied the Bo1 query expansion algorithm provided by PyTerrier which

[13] https://www.datamuse.com/api/

[14] https://99webtools.com/blog/list-of-english-stop-words/

weighs the terms based on divergence from randomness build on Bose-Einstein statistics [90]. Specifically, the Bo1 model extracts terms from the top-ranked documents retrieved for the original query, weighs them based on their informativeness, and appends the highest-weighted terms to the original query to expand it. Finally, a custom block list consisting of commonly repeated phrases such as "my opponent claims...", "PRO claims...", "I accept this debate" filtered further noisy sentences, leading to improved nDCG scores.

Porthos [62] used the Elasticsearch implementation of DirichletLM (with $\mu = 116$ being the average length of sentences in the corpus) and BM25 [65] (default Elasticsearch implementation with $k_1 = 1.2$ and $b = 0.75$) or retrieval after removing sentence duplicates and filtering non-relevant sentences by removing ones with low-quality language to retain only the ones that contain at least one verb. Another filtering step is based on the argumentativeness of sentences using the support vector machine (SVM) of [22] and the BERT approach of [56]. In addition, sentences were stemmed, lowercased and stop words were removed. The approaches are based on a search term as a composition of single terms and Boolean queries together with Reimers et al. [56] to reorder the retrieved sentences according to their argumentative quality. The sentences are paired with SBERT [60] and BERT [8] trained on Next Sentence Prediction (NSP). The best approach is based on DirichletLM, NSP, using the sentence classifier in preprocessing, Boolean queries with Noun Chunking for retrieval, and the BERT approach of [56] for re-ranking.

5. Task 2: Argument Retrieval for Comparative Questions

The goal of the Touché 2022 lab's second task was to support informed decisions in "everyday" or personal comparison situations—for instance for a question like "Should I major in philosophy or psychology?". Decision making in such situations benefits from finding balanced reasons for choosing one option over the other, usually in form of opinions or arguments.

5.1. Task Definition and Data

Task. Given a collection of text passages and a comparative topic with two comparison objects, the task was to retrieve relevant argumentative passages for or against one or both objects, and to detect the passages' stances with respect to the objects.

Topics. We provided 50 topics that describe scenarios of personal decision making. Each topic has a *title* formulated as a comparative question, a pair of *comparison objects* from the title that could be used for the stance detection of the retrieved passages, a *description* with some background on the particular search scenario, and a *narrative* that served as a guideline for our assessors (cf. Table 3 for an example).

Document collection. The retrieval collection for Task 2 was a corpus of 868,655 passages extracted from ClueWeb12.[15] We constructed this passage corpus using all 37,248 documents from the top-100 pool of all runs submitted to Task 2 in the previous Touché editions. Using the

[15] https://lemurproject.org/clueweb12/index.php

Table 3
Example topic for Task 2: Argument Retrieval for Comparative Questions.

Number	88
Title	Should I major in philosophy or psychology?
Objects	major in philosophy, psychology
Description	A soon-to-be high-school graduate finds themself at a crossroad in their life. Based on their interests, majoring in philosophy or in psychology are the potential options and the graduate is searching for information about the differences and similarities, as well as advantages and disadvantages of majoring in either of them (e.g., with respect to career opportunities or gained skills).
Narrative	Relevant documents will overview one of the two majors in terms of career prospects or developed new skills, or they will provide a list of reasons to major in one or the other. Highly relevant documents will compare the two majors side-by-side and help to decide which should be preferred in what context. Not relevant are study program and university advertisements or general descriptions of the disciplines that do not mention benefits, advantages, or pros/-cons.

TREC CAsT tools,[16] we split the documents at sentence boundaries into fixed-length passages of approximately 250 terms, since ranking fixed-length passages was shown to be more effective than that of variable-length passages [91]. From the initial 1,286,977 passages, we removed near-duplicates with CopyCat [92] to mitigate unwanted side-effects of near-duplicates on retrieval effectiveness [93, 94], resulting in the final collection of 868,655 passages. We also provided a second version of the corpus, in which the passages were expanded with queries generated by the docT5query model [95].

To lower the bar to entry of this task, we also provided the participants with a number of previously compiled resources. These included the document-level relevance and argument quality judgments from the previous Touché editions as well as the passage-level relevance judgments from a subset of MS MARCO [96] with about 40,000 comparative questions identified by an ALBERT-based [97] classifier [38]. Each question in MS MARCO is associated with 10 text passages (one is labeled as most relevant). To train stance detectors, an annotated dataset of 950 comparative questions and answers, extracted from Stack Exchange, was also provided [38]. For the identification of claims and premises, the participants could use any own or existing argument tagging tool, such as the API[17] of TARGER [98] hosted on our own servers.

5.2. Evaluation Setup

Similar to Task 1, we pooled the top-5 passages from the runs, resulting in 2,107 unique passages that were manually judged. Our volunteer human assessors labeled the passages' relevance

[16] https://github.com/grill-lab/trec-cast-tools
[17] Also available as a Python library: https://pypi.org/project/targer-api/

with three labels: 0 (not relevant), 1 (relevant), and 2 (highly relevant). They also assessed whether arguments are present in a passage and whether they are rhetorically well-written [27] with three labels: 0 (low quality, or no arguments in a passage), 1 (average quality), and 2 (high quality). Finally, we asked the assessors to label passages with respect to a topic's comparison objects as (a) pro first object, (b) pro second object, (c) neutral (both comparison objects are equally good or bad), and (d) no stance (no stance given). In Task 2, we used nDCG@5 for the relevance and argument quality dimensions and macro-averaged F_1 for the stance detection.

5.3. Submitted Approaches and Evaluation Results

Seven teams submitted their results to Task 2 (25 valid runs). Interestingly, only two teams used relevance judgments from the previous Touché editions to fine-tune their models or to optimize parameters. The others either manually labeled a sample of retrieved documents themselves or relied on zero-shot approaches like the transformer-based model T0++ [6]. Most teams used the standard passage collection, but two teams also used the docT5query-expanded [95] collection provided by us. Overall, the main trend of this year was the usage of transformer-based models for ranking and re-ranking (e.g., ColBERT [99] or monoT5 and duoT5 [100]) while our baseline approach was BM25, as in the previous years.

For the optional subtask of stance detection, five of the seven teams submitted results. They either trained their own classifiers on the provided stance dataset, fine-tuned pre-trained language models, or directly used pre-trained models as zero-shot classifiers. Our baseline stance detector was a simple always-'no stance' predictor (majority class).

Table 4 shows the results of each team's most effective runs with respect to relevance and argument quality (more detailed results for each submitted run can be found in Appendix A). For stance detection, for each team, we evaluated all passages that were part of the manual judgment pool and for which the team had predicted a stance (i.e., the stance of a passage returned at Rank 3 by some Team X (and thus part of the judgment pool) was also used in the stance evaluation of Team Y, even when the document was only on Rank 6 or lower (and thus not actually part of the pool for that run). Note that this potentially yields different numbers of passages used for the stance evaluation per team. Below, we briefly describe the teams' submitted approaches and their results (teams ordered by their relevance-wise best approach).

Captian Levi [101] submitted the relevance-wise most effective run. They first retrieved 2,000 documents using Pyserini's BM25 [65] ($k_1 = 1.2$ and $b = 0.68$) by combining top-1000 results for the original query (topic title) with the results for modified queries, where they used alternative strategies: (1) only removing stop words (using the NLTK [102] stop word list), (2) replacing comparative adjectives with synonyms and antonyms found in WordNet [70], (3) adding extra terms using pseudo-relevance feedback, (4) using queries generated with the docT5query model [95] provided by the Touché organizers. Queries and corpus were also processed by using stop words and punctuation removal and lemmatization (WordNet lemmatizer). The initially retrieved results were re-ranked using monoT5 and duoT5 [100]. Additionally, TCT-ColBERT [9] (a variant of ColBERT [99] with knowledge distillation) was also used for initial ranking for unmodified queries (topic titles). *Captain Levi* submitted in total five runs that differ in the aforementioned strategies of modifying queries, initial ranking models, and final re-ranking models. Their most effective run in terms of relevance and quality was

Table 4
Results of Task 2 (Argument Retrieval for Comparative Questions). (a) Evaluation results of a team's best run according to the results' relevance. (b) Best runs according to the results' quality. (c) Stance detection results (the teams' ordering is the same as in (b)). An asterisk (⋆) indicates that the runs with the best relevance and the best quality differ for a team. The baseline BM25 ranking is shown in bold; the baseline stance detector always predicts 'no stance'. A † indicates statistically significant differences to the baseline (paired Student's t-test, $p = 0.05$, Bonferroni-correction). Since stance detection results were calculated for different numbers of predictions for each team, we do not test statistical differences. Tables 9–11 show the results for all submitted runs.

(a) Best relevance score per team

Team	**nDCG@5**	
	Rel.	Qual.
Captain Levi	0.758†	0.744
Aldo Nadi⋆	0.709†	0.748
Katana⋆	0.618†	0.643
Captain Tempesta⋆	0.574†	0.589
Olivier Armstrong	0.492	0.582
Puss in Boots	**0.469**	**0.476**
Grimjack	0.422	0.403
Asuna	0.263†	0.332

(b) Best quality score per team

Team	**nDCG@5**	
	Qual.	Rel.
Aldo Nadi⋆	0.774†	0.695
Captain Levi	0.744†	0.758
Katana⋆	0.644†	0.601
Captain Tempesta⋆	0.597†	0.557
Olivier Armstrong	0.582	0.492
Puss in Boots	**0.476**	**0.469**
Grimjack	0.403	0.422
Asuna	0.332†	0.263

(c) Stance

F_1 macro	
Rank	Score
	—
1	0.261
3	0.220
	—
4	0.191
5	**0.158**
2	0.235
6	0.106

initial ranking by TCT-ColBERT. Finally, stance was detected using a RoBERTa-Large-MNLI model [103], pre-trained on the Multi-Genre Natural Language Inference corpus [104] without further fine-tuning in two steps: (1) detecting if the document has a stance, and then (2) for documents that were not classified as 'neutral' or 'no stance', detecting which comparison object the document favors. This stance detector achieved the highest macro-averaged F_1 score.

Aldo Nadi [105] submitted the quality-wise most effective run. They re-ranked passages that were initially retrieved with BM25F [10] (default Lucene implementation with $k_1 = 1.2$ and $b = 0.75$) on two fields: text of the original passages, and passages expanded with docT5query. All texts were processed with the Porter stemmer [67], removing stop words using different lists: (a) Snowball [106], (b) a default Lucene stop word list, (c) a custom list containing the 400 most frequent terms in the retrieval collection, excluding the comparison objects. Queries (topic titles) were expanded using a relevance feedback method based on the Rocchio Algorithm [107]. For the final ranking, the team experimented with two re-ranking techniques (involving up to the top-1000 documents from the initial results): (1) exploiting the argument quality estimation, i.e., they multiplied the document relevance and the quality scores, and (2) Reciprocal Ranking Fusion [108]. The quality scores were predicted using the IBM Project Debater API [55]. *Aldo Nadi* submitted five runs, which vary by different combinations of the proposed methods, e.g., using different stop word lists for pre-processing, using relevance feedback or not, using the quality-based re-ranking or fusion. The team's most effective run in terms of relevance used relevance feedback, and the most effective run in terms of quality was based on Reciprocal

Ranking Fusion. The did not detect the stance.

Katana [109] submitted three runs that all used different variants of ColBERT [99]: (1) pre-trained on MS MARCO [96] by the University of Glasgow,[18] (2) pre-trained by *Katana* from scratch on MS MARCO, replacing a cosine similarity between a query and a document representation with L2 distance, and (3) the latter model fine-tuned on the relevance and quality judgments from the previous Touché editions. As queries the team used topic titles without additional processing. The team's most effective run in terms of relevance used ranking by pre-trained ColBERT, and the most effective run in terms of quality used ranking by training ColBERT from scratch (without further fine-tuning). For stance detection, *Katana* used a pre-trained XGBoost-based classifier that is part of Comparative Argumentation Machine [35, 33].

Captain Tempesta [110] exploited linguistic properties of text such as a non-informative symbol frequency (hashtags, emojis, etc.), a difference between a short words' (less or equal than 4 characters) frequency and a long words' (more than 4 characters) frequency, and adjective as well as comparative adjective frequencies. Based on these properties for each document in the retrieval corpus, a quality score was computed as a weighted sum (weights were assigned manually). At query time, the relevance score of BM25 (Lucene; default: $k_1 = 1.2$ and $b = 0.75$) was multiplied with the quality score, used as ranking criterion. Queries (topic titles) were processed by removing stop words (Lucene default list) and lowercasing query terms except for brand names,[19] stemming them using Lovins stemmer [111]. The team's five submitted runs differ in the weights manually assigned for the different quality properties. The team's most effective run in terms of relevance used document quality estimation with linguistic properties, and the most effective run in terms of quality did not. The team did not detect stance.

Olivier Armstrong [112] submitted one run. They first identified the comparison objects, aspects, and predicates in queries (topic titles) using a RoBERTa-based classifier proposed by Bondarenko et al. [38]. After removing stop words, queries were expanded with synonyms of the objects, aspects, and predicates found using WordNet. Then 100 documents were retrieved using Elasticsearch's BM25 ($k_1 = 1.2$ and $b = 0.75$) as initial ranking. Using a DistilBERT-based classifier [113], fine-tuned by Alhamzeh et al. [114] (a Touché 2021 participant), *Olivier Armstrong* identified premises and claims in the retrieved documents. For ranking, the following scores were calculated for each candidate document: (1) the arg-BM25 score returned by querying the new re-indexed corpus (only premises and claims are kept) using the unmodified queries (topic titles), (2) the argument support score, i.e., the ratio of premises and claims in the document, (3) the similarity score, i.e., the averaged cosine similarity between the original query and every premise and claim in the document represented using the SBERT embeddings [60]. The final score for each candidate document was calculated as sum of the normalized individual scores. Their final ranking included 25 documents. For stance detection, the team used an LSTM-based neural network with one hidden layer that was pre-trained on the provided stance dataset.

Puss in Boots was our baseline retrieval model that used the BM25 implementation in Pyserini [65] with default parameters ($k_1 = 0.9$ and $b = 0.4$) and original topic titles as queries. The baseline stance detector simply assigned 'no stance' to all documents in the ranked list.

Grimjack [115] submitted five runs using query expansion and query reformulation, argument

[18] http://www.dcs.gla.ac.uk/∼craigm/colbert.dnn.zip

[19] https://github.com/MatthiasWinkelmann/english-words-names-brands-places

quality estimation, stance detection, and axiomatic re-ranking. For the first ranking result, the team simply retrieved 100 passages ranked with a Pyserini implementation of DirichletLM (default $\mu = 1000$), using original, unmodified queries (topic titles). Another approach re-ranked the top-10 of the initially retrieved passages using (1) argument axioms that "prefer" documents with more premises and claims (identified with TARGER [98]) or earlier occurrence of query terms in premises and claims [116, 117], (2) newly proposed comparative axioms that "prefer" documents with more comparison objects or their earlier occurrence in premises and claims, and (3) an argument quality axiom that ranks higher documents with higher argument quality scores calculated using the IBM Project Debater API [55]. For another result ranking, document positions (from the previous run) were changed based on the predicted stance, such as the 'pro first object' document was followed by the 'pro second object' followed by 'neutral' stance. The document stance was predicted using the IBM Project Debater API [55]. The last two runs used T0++ [6] (1) to expand queries, e.g., by combining topic titles with newly generated queries, where T0++ was prompted to generate a question given a topic's description, (2) to assess the argument quality, and (3) to detect the stance in zero-shot settings. These two runs differed in whether a stance balancing was used. The team's most effective run in terms of relevance and quality used axiomatic re-ranking, and re-ranking based on the detected stance.

Asuna [118] preprocessed each document (passage) in the retrieval corpus by (1) creating a one-sentence extractive summary using LexRank [119], (2) identifying premises and claims with TARGER [98], and (3) looking up the spam score in the Waterloo Spam Rankings dataset [120].[20] The modified corpus was indexed, and initial retrieval of the top-40 documents was performed with the Pyserini [65] implementation of BM25F (default $k_1 = 0.9$ and $b = 0.4$) using the unmodified queries (topic titles) over the index fields with original passages, summaries, and premises and claims. Next, the queries were lemmatized and stop words were removed using the NLTK library, and expanded with the most frequent terms coming from LDA topics [121] for the initially retrieved documents. The expanded queries were used to, again, retrieve the top-40 passages with BM25F. Finally, *Asuna* re-ranked the retrieved documents using a random forest classifier [122] with the following features: BM25F score, number of times the document was retrieved for different queries (original, three extended with the LDA topics for documents, and one extended with the LDA topic for the task topic description), number of tokens in documents, number of sentences in documents, number of premises in documents, number of claims in documents, spam-score, predicted argument quality score, and predicted stance. The classifier was trained on the Touché 2020 and 2021 relevance judgments. The argument quality was predicted using DistilBERT, fine-tuned on the Webis-ArgQuality-20 corpus [89]. The stance was also predicted using DistilBERT, fine-tuned on the provided stance dataset.

6. Task 3: Image Retrieval for Arguments

The goal of the Touché 2022 lab's 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 for compelling images to support one's argumentation. The goal of the third task was thus to retrieve images that indicate an agreement or disagreement

[20] https://lemurproject.org/clueweb12/related-data.php

to some stance on a given topic as two separate lists similar to textual argument search.

6.1. Task Definition and Data

Task. Given a controversial topic, the task was to retrieve images (from web pages) for each stance (pro and con) that show support for that stance.

Topics. Task 3 uses the same 50 controversial topics as Task 1 (cf. Section 4).

Document collection. This task's document collection stems from a focused crawl of 23,841 images and associated web pages from late 2021. For each of the 50 topics, we issued 11 queries (with different filter words like "good," "meme," "stats," "reasons," or "effects") to Google's image search and downloaded the top 100 images and associated web pages; 868 duplicate images were identified and removed using pHash[21] and manual checks. The dataset contains for each image: (1) the image itself in both WebP and PNG format, (2) its URL; (3) its pHash. Moreover, the dataset contains for each page: (1) its URL; (2) the Google rank of the page for each query for which the image was retrieved; (3) a WARC web archive;[22] (4) a DOM HTML snapshot; (5) its complete text; (6) a screenshot; (7) meta-information of each DOM node, including the node's xPath, CSS attributes, and position on the screenshot; and (8) the xPath of the corresponding image in the DOM HTML snapshot. The full dataset is 368 GB large.[23] To kickstart machine learning approaches, we provided 334 relevance judgments from Kiesel et al. [51].

6.2. Evaluation Setup

We employed crowdsourcing on Amazon Mechanical Turk[24] to evaluate the topical relevance, argumentativeness, and stance of the 6,607 image-topic pairs from all runs, employing 5 independent annotators each. Specifically, we asked for each topic for which an image was retrieved: (1) Is the image in some manner related to the topic? (2) Do you think most people would say that, if someone shares this image without further comment, they want to show they approve of the pro-side to the topic? (3) Or do you think most people would rather say the one who shares this image does so to show they disapprove? We described each topic using the topic's title, modified as necessary to convey the description and narrative (cf. Table 1) and to clarify which stance is approve (pro) and disapprove (con). We then iteratively employed MACE [123] to identify image–topic pairs with low annotator agreement (MACE confidence ≤ 0.55) and re-judged them ourselves, employing our judgments as check instances for another iteration of MACE. We repeated this procedure until MACE predicted the labels for all image–topic pairs from the runs with a confidence above 0.55 (re-judging 2,056 images total).

[21] https://www.phash.org/
[22] Archived using https://github.com/webis-de/scriptor
[23] Available at https://webis.de/data.html#touche22-image-retrieval-for-arguments
[24] https://www.mturk.com

Table 5
Results of Task 3 (Image Retrieval for Arguments) in terms of Precision@10 (per stance) for topic relevance, argumentativeness, and stance relevance. The table shows the best run for each team across all three measures. Results for the baseline are shown in bold. A † indicates statistically significant differences to the baseline (paired Student's t-test, $p = 0.05$, Bonferroni-correction). Table 12 shows the results for all submitted runs.

Team	Run	Precision@10		
		Topic	Arg.	Stance
Boromir	BERT, OCR, query-processing	0.878†	0.768†	0.425
Minsc	**Baseline**	**0.736**	**0.686**	**0.407**
Aramis	Argumentativeness:formula, stance:formula	0.701†	0.634	0.381
Jester	With emotion detection	0.696†	0.647†	0.350†

6.3. Submitted Approaches and Evaluation Results

In total, 3 teams submitted 12 runs to this task. The teams pursued quite different approaches. However, all participants employed OCR (specifically Tesseract[25]) to extract image text. The teams Boromir and Jester also used the associated web page's text, but Team Jester restricted to text close to the image on the web page. Each team used sentiment or emotion features, based on image colors (*Aramis*), faces in the images (*Jester*), image text (all), and the web page text (*Boromir*, *Jester*). *Boromir* used the ranking information for internal evaluation.

We used Precision@10 for evaluation: the ratio of relevant images among 10 retrieved images for each topic and stance. Table 5 shows the results of each team's most effective run. For each team, the best runs were the same with respect to all three measures.

Minsc represents our baseline run, which ranks images in the same order as our original Google queries, namely of the query that includes the filter word "good" for pro and of the query that includes "anti" for con. We considered this a tough baseline, especially for on-topic relevance, as topical relatedness is similar for argumentative and "standard" web image search. However, *Boromir* beat this baseline—with a considerable margin for on-topic relevance.

Aramis [124] focused on image features. No retrieval model was employed, but all images evaluated for each topic. They tested the use of a heuristic formula vs. fully-connected neural network classifiers for both argumentativeness and stance detection. Features were based on OCR (text length in characters, text area size, and cells in an 8×8 grid with high text density, VADER sentiment score [78]), image color (average color, dominant color, and percentages of pixels with each of these color ranges as per self-defined RGB-buckets: red, green, blue, yellow, light, and dark), image category (graphic vs. photo [125]; percentage of area covered by diagrams[26]), and query–text similarity (whether the query is fully contained, the overlap for an optimal query alignment, and VADER sentiment score of words in a six-token radius around occurrences of query terms in the text). However, the query–text similarity features were not used for argumentativeness classification, as the team assumed this sub-task to be query-independent. In our evaluation, the formula performed better than the neural approaches, which

[25] https://github.com/tesseract-ocr/tesseract
[26] Based on a Stackoverflow answer, archived as https://perma.cc/KE6J-KMQT

Aramis traced back to the formula being slightly better at handling off-topic images—with topical relevance not being the team's focus, they had trained and internally evaluated the network on on-topic images only. However, their worst runs still achieved a similar Precision@10 as their best one, namely 0.664 (topic; -0.037 compared to best run), 0.609 (argumentativeness; -0.025), and 0.344 (stance; -0.037). Moreover, for an evaluation that ignores the problem of topical relevance, the ratio of argumentative images among topical relevant images for their runs is between 0.904 (using both formulas) and 0.927 (using both networks), and thus very close to the baseline, which reaches a ratio of 0.932.

Boromir [126] indexed both image text (boosted five-fold) and web page text (Elasticsearch BM25 with default settings, $k_1 = 1.2$ and $b = 0.75$), using lowercasing, URL, punctuation and number removal, NLTK's WordNet lemmatization [102], removal of tokens consisting of exactly one letter, stop word removal (using the list from NLTK), and min-frequency filtering (removing tokens that appear less than three times in the text). They clustered images into 13 clusters (as determined by the elbow criterion) using k-Means and manually assigned retrieval boosts per cluster to favor more argumentative images, especially diagrams. For example, clusters with the highest boost of 5.0 were found to contain, upon manual inspection, "graphics with text (e.g., memes, quotes, twitter posts)," "graphics with round forms and text (e.g., pie charts)," "statistical graphics but with better quality [...] (e.g., bar plots, tables, line plots)," and "statistical plots (bar plots and line plots)." On the other hand, not boosted were images from clusters that were found to contain mostly photos (five clusters). They employed textual sentiment detection for stance detection, using either a dictionary (AFINN [127]) or a BERT classifier. Their approach performed best and convincingly improved over the baseline. The BERT classifier improved over the dictionary-based classifier whereas image clustering was detrimental. Specifically, the image clustering seemed to introduce more off-topic images into the ranking: the same setup as the best run but using image clusters achieved a Precision@10 of 0.822 (topic; -0.056), 0.728 (argumentativeness; -0.040), and 0.411 (stance; -0.014).

Jester focused on emotion-based image retrieval via facial image recognition (using FER[27]), image text, and the associated web page's text that is close to the image in the HTML source code—for which they use the text within the image's parent element. Similar to stance detection in the args.me search engine [14], they assign positive-leaning images to the pro-stance and negative-leaning images to the con-stance. For comparison, they submitted a second run without emotion features (thus plain retrieval), which achieved a lower Precision@10: 0.671 (topic; -0.025), 0.618 (argumentativeness; -0.029), and 0.336 (stance; -0.014). Thus emotion features seem helpful but insufficient when taken alone.

7. Conclusion

The third edition of the Touché lab at CLEF 2022 featured three shared tasks: (1) argument retrieval for controversial questions, (2) argument retrieval for comparative questions, and (3) image retrieval for arguments. Compared to previous editions, retrieval units have been changed (sentences/passages instead of full arguments/documents and images as a completely new unit) and stance detection has been included. Of 58 registered teams, 23 participated in

[27] https://github.com/justinshenk/fer

the tasks and submitted at least one valid run. In addition to sparse retrieval and various query processing, reformulation, and expansion methods, approaches have increasingly focused on transformer models and re-ranking techniques. Not only was the quality of the documents and arguments evaluated, but also the predicted stance taken into account for the final rankings.

The most effective approaches to argument retrieval all share common characteristics. For example, most use various strategies for query reformulation and expansion, such as synonyms, relevance feedback, or generating new queries with pre-trained language models. An interesting observation is that re-ranking first-stage search results based on a quality assessment of the arguments almost always improves 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 assertions) 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 retrieval effectiveness. However, the overall still rather low effectiveness of the approaches to stance detection leaves room for future improvements. For Task 3 (image retrieval), the recognition of sentiment and emotion and the use of OCR to analyze the text in images were particularly helpful.

We plan to continue Touché as a collaborative platform for researchers in argument retrieval. All Touché resources are freely available, including topics, manual relevance and argument quality assessments, and submitted runs from participating teams. These resources, the submission and evaluation tools, and other events such as workshops will help to further foster the community working on argument retrieval. In the future, we plan to expand the evaluation pools and to include additional dimensions of argument quality. Improving stance detection and exploiting predicted stances better not only for ranking text arguments but also for images are also interesting tasks for future work.

Acknowledgments

We are very grateful to the CLEF 2022 organizers and the Touché participants, who allowed this lab to happen. We also want to thank our volunteer annotators who helped to create the relevance and argument quality assessments and our reviewers for their valuable feedback on the participants' notebooks.

This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) through the projects "ACQuA 2.0" (Answering Comparative Questions with Arguments; project number 376430233) and "OASiS" (Objective Argument Summarization in Search; project number 455913891) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999), and the German Ministry for Science and Education (BMBF) through the project "SharKI" (Shared Tasks as an Innovative Approach to Implement AI and Big Data-based Applications within Universities; grant FKZ 16DHB4021). We are also grateful to Jan Heinrich Reimer for developing the TARGER Python library and Erik Reuter for expanding a document collection for Task 2 with docT5query.

References

[1] 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2022.

[2] 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 Papers of the CLEF 2020 Evaluation Labs, volume 2696 of *CEUR Workshop Proceedings*, 2020. URL: http://ceur-ws.org/Vol-2696/.

[3] 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 Papers of the CLEF 2021 Evaluation Labs, volume 2936 of *CEUR Workshop Proceedings*, 2021. URL: http://ceur-ws.org/Vol-2936/.

[4] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: Proceedings of The Third Text REtrieval Conference, TREC 1994, volume 500-225 of *NIST Special Publication*, NIST, 1994, pp. 109–126. URL: https://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[6] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. V. Nayak, D. Datta, J. Chang, M. T. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Févry, J. A. Fries, R. Teehan, S. Biderman, L. Gao, T. Bers, T. Wolf, A. M. Rush, Multitask prompted training enables zero-shot task generalization, CoRR abs/2110.08207 (2021). URL: https://arxiv.org/abs/2110.08207. arXiv:2110.08207.

[7] J. Lin, R. Nogueira, A. Yates, Pretrained Transformers for Text Ranking: BERT and Beyond, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2021. URL: https://doi.org/10.2200/S01123ED1V01Y202108HLT053. doi:10.2200/S01123ED1V01Y202108HLT053.

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, ACL, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423.

[9] S. Lin, J. Yang, J. Lin, Distilling dense representations for ranking using tightly-coupled teachers, CoRR abs/2010.11386 (2020). URL: https://arxiv.org/abs/2010.11386. arXiv:2010.11386.

[10] S. E. Robertson, H. Zaragoza, M. J. Taylor, Simple BM25 extension to multiple weighted fields, in: Proceedings of the 13th International Conference on Information and Knowledge Management, CIKM 2004, ACM, 2004, pp. 42–49. URL: https://doi.org/10.1145/

1031171.1031181.

[11] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021. URL: https://openreview.net/forum?id=wCu6T5xFjeJ.

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

[13] H. Wachsmuth, S. Syed, B. Stein, Retrieval of the best counterargument without prior topic knowledge, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Association for Computational Linguistics, 2018, pp. 241–251. URL: https://www.aclweb.org/anthology/P18-1023/.

[14] H. Wachsmuth, M. Potthast, K. Al-Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, B. Stein, Building an argument search engine for the web, in: Proceedings of the Fourth Workshop on Argument Mining (ArgMining), Association for Computational Linguistics, 2017, pp. 49–59. URL: https://doi.org/10.18653/v1/w17-5106.

[15] Y. Ajjour, P. Braslavski, A. Bondarenko, B. Stein, Identifying argumentative questions in web search logs, in: 45th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2022), ACM, 2022. doi:10.1145/3477495.3531864.

[16] Y. Ajjour, H. Wachsmuth, J. Kiesel, M. Potthast, M. Hagen, B. Stein, Data acquisition for argument search: The args.me corpus, in: Proceedings of the 42nd German Conference on AI, KI 2019, volume 11793 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 48–59. URL: https://doi.org/10.1007/978-3-030-30179-8_4.

[17] R. Levy, B. Bogin, S. Gretz, R. Aharonov, N. Slonim, Towards an argumentative content search engine using weak supervision, in: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Association for Computational Linguistics, 2018, pp. 2066–2081. URL: https://www.aclweb.org/anthology/C18-1176/.

[18] C. Stab, J. Daxenberger, C. Stahlhut, T. Miller, B. Schiller, C. Tauchmann, S. Eger, I. Gurevych, ArgumenText: Searching for arguments in heterogeneous sources, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2018, Association for Computational Linguistics, 2018, pp. 21–25. URL: https://www.aclweb.org/anthology/N18-5005.

[19] Aristotle, G. A. Kennedy, On Rhetoric: A Theory of Civic Discourse, Oxford: Oxford University Press, 2006.

[20] H. Wachsmuth, N. Naderi, I. Habernal, Y. Hou, G. Hirst, I. Gurevych, B. Stein, Argumentation quality assessment: Theory vs. practice, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Association for Computational Linguistics, 2017, pp. 250–255. URL: https://doi.org/10.18653/v1/P17-2039.

[21] M. Potthast, L. Gienapp, F. Euchner, N. Heilenkötter, N. Weidmann, H. Wachsmuth, B. Stein, M. Hagen, Argument search: Assessing argument relevance, in: Proceedings of the 42nd International Conference on Research and Development in Information Retrieval, SIGIR 2019, ACM, 2019, pp. 1117–1120. URL: https://doi.org/10.1145/3331184.3331327.

[22] L. Gienapp, B. Stein, M. Hagen, M. Potthast, Efficient pairwise annotation of argument

quality, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5772–5781. URL: https://aclanthology.org/2020.acl-main.511. doi:10.18653/v1/2020.acl-main.511.

[23] H. Wachsmuth, B. Stein, Y. Ajjour, "PageRank" for argument relevance, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Association for Computational Linguistics, 2017, pp. 1117–1127. URL: https://doi.org/10.18653/v1/e17-1105.

[24] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web., Technical Report 1999-66, Stanford InfoLab, 1999. URL: http://ilpubs.stanford.edu:8090/422/.

[25] L. Dumani, P. J. Neumann, R. Schenkel, A framework for argument retrieval - ranking argument clusters by frequency and specificity, in: Proceedings of the 42nd European Conference on IR Research (ECIR 2020), volume 12035 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 431–445. URL: https://doi.org/10.1007/978-3-030-45439-5_29.

[26] L. Dumani, R. Schenkel, Quality aware ranking of arguments, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, 2020, pp. 335–344. URL: https://doi.org/10.1007/978-3-030-45439-5_29.

[27] 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 the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, 2017, pp. 176–187. URL: http://aclweb.org/anthology/E17-1017.

[28] A. Nadamoto, K. Tanaka, A comparative web browser (CWB) for browsing and comparing web pages, in: Proceedings of the 12th International World Wide Web Conference, WWW 2003, ACM, 2003, pp. 727–735. URL: https://doi.org/10.1145/775152.775254.

[29] J. Sun, X. Wang, D. Shen, H. Zeng, Z. Chen, CWS: A comparative web search system, in: Proceedings of the 15th International Conference on World Wide Web, WWW 2006, ACM, 2006, pp. 467–476. URL: https://doi.org/10.1145/1135777.1135846.

[30] N. Jindal, B. Liu, Identifying comparative sentences in text documents, in: Proceedings of the 29th Annual International Conference on Research and Development in Information Retrieval, SIGIR 2006, ACM, 2006, pp. 244–251. URL: https://doi.org/10.1145/1148170.1148215.

[31] N. Jindal, B. Liu, Mining comparative sentences and relations, in: Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference, AAAI 2006, AAAI Press, 2006, pp. 1331–1336. URL: http://www.aaai.org/Library/AAAI/2006/aaai06-209.php.

[32] W. Kessler, J. Kuhn, A corpus of comparisons in product reviews, in: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, European Language Resources Association (ELRA), 2014, pp. 2242–2248. URL: http://www.lrec-conf.org/proceedings/lrec2014/summaries/1001.html.

[33] A. Panchenko, A. Bondarenko, M. Franzek, M. Hagen, C. Biemann, Categorizing comparative sentences, in: Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Association for Computational Linguistics, 2019, pp. 136–145. URL: https://doi.org/10.18653/v1/w19-4516.

[34] N. Ma, S. Mazumder, H. Wang, B. Liu, Entity-aware dependency-based deep graph attention network for comparative preference classification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Association for Computational Linguistics, 2020, pp. 5782–5788. URL: https://www.aclweb.org/anthology/2020.acl-main.512/.

[35] M. Schildwächter, A. Bondarenko, J. Zenker, M. Hagen, C. Biemann, A. Panchenko, Answering comparative questions: Better than ten-blue-links?, in: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, ACM, 2019, pp. 361–365. URL: https://doi.org/10.1145/3295750.3298916.

[36] V. Chekalina, A. Bondarenko, C. Biemann, M. Beloucif, V. Logacheva, A. Panchenko, Which is better for deep learning: Python or matlab? answering comparative questions in natural language, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021, Association for Computational Linguistics, 2021, pp. 302–311. URL: https://www.aclweb.org/anthology/2021.eacl-demos.36/.

[37] A. Bondarenko, P. Braslavski, M. Völske, R. Aly, M. Fröbe, A. Panchenko, C. Biemann, B. Stein, M. Hagen, Comparative web search questions, in: Proceedings of the 13th ACM International Conference on Web Search and Data Mining, WSDM 2020, ACM, 2020, pp. 52–60. URL: https://dl.acm.org/doi/abs/10.1145/3336191.3371848.

[38] A. Bondarenko, Y. Ajjour, V. Dittmar, N. Homann, P. Braslavski, M. Hagen, Towards understanding and answering comparative questions, in: Proceedings of the 15th ACM International Conference on Web Search and Data Mining, WSDM 2022, ACM, 2022, pp. 66–74. doi:10.1145/3488560.3498534.

[39] I. J. Dove, On images as evidence and arguments, in: F. H. van Eemeren, B. Garssen (Eds.), Topical Themes in Argumentation Theory: Twenty Exploratory Studies, Argumentation Library, Springer Netherlands, Dordrecht, 2012, pp. 223–238. doi:10.1007/978-94-007-4041-9_15.

[40] F. Dunaway, Images, emotions, politics, Modern American History 1 (2018) 369–376. doi:10.1017/mah.2018.17.

[41] G. Roque, Visual argumentation: A further reappraisal, in: F. H. van Eemeren, B. Garssen (Eds.), Topical Themes in Argumentation Theory, volume 22, Springer Netherlands, Dordrecht, 2012, pp. 273–288. URL: http://link.springer.com/10.1007/978-94-007-4041-9_18. doi:10.1007/978-94-007-4041-9_18, series Title: Argumentation Library.

[42] I. Grancea, Types of visual arguments, Argumentum. Journal of the Seminar of Discursive Logic, Argumentation Theory and Rhetoric 15 (2017) 16–34.

[43] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, in: 15th International Workshop on Semantic Evaluation (SemEval'2021), Association for Computational Linguistics, Online, 2021, pp. 70–98. URL: https://aclanthology.org/2021.semeval-1.7. doi:10.18653/v1/2021.semeval-1.7.

[44] N. Chang, K. Fu, Query-by-pictorial-example, IEEE Transactions on Software Engineering 6 (1980) 519–524. doi:10.1109/TSE.1980.230801.

[45] P. Aigrain, H. Zhang, D. Petkovic, Content-based representation and retrieval of visual media: A state-of-the-art review, Multimedia Tools and Applications 3 (1996) 179–202.

doi:10.1007/BF00393937.
[46] A. Latif, A. Rasheed, U. Sajid, J. Ahmed, N. Ali, N. I. Ratyal, B. Zafar, S. H. Dar, M. Sajid, T. Khalil, Content-based image retrieval and feature extraction: A comprehensive review, Mathematical Problems in Engineering 2019 (2019) 21. doi:10.1155/2019/9658350.
[47] A. Wu, Learn more about what you see on google images, Google Blog, 2020. URL: https://support.google.com/webmasters/answer/114016.
[48] Google, Google images best practices, Google Developers, 2021. URL: https://support.google.com/webmasters/answer/114016.
[49] W. Wang, Q. He, A survey on emotional semantic image retrieval, in: International Conference on Image Processing (ICIP 2008), IEEE, 2008, pp. 117–120. doi:10.1109/ICIP.2008.4711705.
[50] M. Solli, R. Lenz, Color emotions for multi-colored images, Color Research & Application 36 (2011) 210–221. doi:10.1002/col.20604.
[51] J. Kiesel, N. Reichenbach, B. Stein, M. Potthast, Image retrieval for arguments using stance-aware query expansion, in: Proceedings of the 8th Workshop on Argument Mining, ArgMining 2021 at EMNLP, ACL, 2021, pp. 36–45.
[52] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA integrated research architecture, in: Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, volume 41 of *The Information Retrieval Series*, Springer, 2019, pp. 123–160. URL: https://doi.org/10.1007/978-3-030-22948-1_5.
[53] M. Alshomary, N. Düsterhus, H. Wachsmuth, Extractive snippet generation for arguments, in: Proceedings of the 43nd International ACM Conference on Research and Development in Information Retrieval, SIGIR 2020, ACM, 2020, pp. 1969–1972. URL: https://doi.org/10.1145/3397271.3401186.
[54] 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 the 42nd International Conference on Research and Development in Information Retrieval, SIGIR 2019, ACM, 2019, pp. 1325–1328. URL: https://doi.org/10.1145/3331184.3331399.
[55] R. Bar-Haim, Y. Kantor, E. Venezian, Y. Katz, N. Slonim, Project debater apis: Decomposing the AI grand challenge, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 267–274. URL: https://doi.org/10.18653/v1/2021.emnlp-demo.31.
[56] N. Reimers, B. Schiller, T. Beck, J. Daxenberger, C. Stab, I. Gurevych, Classification and clustering of arguments with contextualized word embeddings, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 567–578. URL: https://aclanthology.org/P19-1054. doi:10.18653/v1/P19-1054.
[57] S. Gretz, R. Friedman, E. Cohen-Karlik, A. Toledo, D. Lahav, R. Aharonov, N. Slonim, A large-scale dataset for argument quality ranking: Construction and analysis, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, AAAI Press, 2020, pp. 7805–7813. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6285.

[58] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.
[59] M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, To appear 7 (2017) 411–420.
[60] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410.
[61] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[62] P. Sülzle, N. Wenzlitschke, Using BERT to retrieve relevant and argumentative sentence pairs, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[63] S. Bahrami, G. P. Goli, A. Pasin, N. Rajkumari, M. M. Sohail, P. Tahan, N. Ferro, SEUPD@CLEF: Team INTSEG on argument retrieval for controversial questions, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[64] B. Moreira, H. Cardoso, B. Martins, F. Goularte, Team Bruce Banner at Touché 2022: Argument retrieval for controversial questions, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[65] J. Lin, X. Ma, S. Lin, J. Yang, R. Pradeep, R. Nogueira, Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021, ACM, 2021, pp. 2356–2362. URL: https://doi.org/10.1145/3404835.3463238.
[66] L. Cappellotto, M. Lando, D. Lupu, M. Mariotto, R. Rosalen, N. Ferro, SEUPD@CLEF: Team 6musk on argument retrieval for controversial questions by using pairs selection and query expansion, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[67] M. F. Porter, An algorithm for suffix stripping, Program 14 (1980) 130–137. URL: https://doi.org/10.1108/eb046814.
[68] R. Krovetz, Viewing morphology as an inference process, in: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93, Association for Computing Machinery, New York, NY, USA, 1993, p. 191–202. URL: https://doi.org/10.1145/160688.160718. doi:10.1145/160688.160718.
[69] D. D. Lewis, Y. Yang, T. G. Rose, F. Li, RCV1: A new benchmark collection for text categorization research, J. Mach. Learn. Res. 5 (2004) 361–397. URL: http://jmlr.org/papers/volume5/lewis04a/lewis04a.pdf.
[70] G. A. Miller, WordNet: A lexical database for English, Communications of the ACM 38 (1995) 39–41.
[71] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, 2013. URL: http://arxiv.org/abs/1301.3781.

[72] A. Benetti, M. D. Togni, G. Foti, R. Lacini, A. Matteazzi, E. Sgarbossa, N. Ferro, SEUPD@CLEF: Team Gamora on argument retrieval for controversial questions, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[73] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543. URL: https://doi.org/10.3115/v1/d14-1162.
[74] M. Barusco, G. D. Fiume, R. Forzan, M. G. Peloso, N. Rizzetto, E. Soleymani, N. Ferro, SEUPD@CLEF: Team Lgtm on argument retrieval for controversial questions, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[75] D. Harman, How effective is suffixing?, Journal of the american society for information science 42 (1991) 7–15.
[76] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky, The stanford corenlp natural language processing toolkit, in: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.
[77] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, Text mining: applications and theory 1 (2010) 10–1002.
[78] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the international AAAI conference on web and social media, volume 8, 2014, pp. 216–225.
[79] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, B. S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, Technical Report, Naval Technical Training Command Millington TN Research Branch, 1975.
[80] M. S. Ebrahimi, A. Crivellari, A. Sah, M. Hansen, S. Mehrbanou, P. Ashok, N. Ferro, SEUPD@CLEF: SPAM on argument retrieval for controversial questions, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[81] J. Wuerf, Similar but different: Simple re-ranking approaches for argument retrieval, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[82] J. G. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobel (Eds.), SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, ACM, 1998, pp. 335–336. URL: https://doi.org/10.1145/290941.291025. doi:10.1145/290941.291025.
[83] M. J. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger, From word embeddings to document distances, in: F. R. Bach, D. M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of *JMLR Workshop and Conference Proceedings*, JMLR.org, 2015, pp. 957–966. URL: http://proceedings.mlr.press/v37/kusnerb15.html.
[84] C. V. Ta, F. Reiner, I. von Detten, F. Stöhr, Finding pairs of argumentative sentences using embeddings, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR

Workshop Proceedings, 2022.
[85] S. Schmidt, J. Probst, B. Bartelt, A. Hinz, Two-stage retrieval for pairs of argumentative sentences, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[86] C. Zhai, J. D. Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval, in: Proceedings of the 24th International Conference on Research and Development in Information Retrieval, SIGIR 2001, ACM, 2001, pp. 334–342. URL: https://doi.org/10.1145/383952.384019.
[87] G. Amati, Frequentist and bayesian approach to information retrieval, in: M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, A. Yavlinsky (Eds.), Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006, Proceedings, volume 3936 of *Lecture Notes in Computer Science*, Springer, 2006, pp. 13–24. URL: https://doi.org/10.1007/11735106_3. doi:10.1007/11735106_3.
[88] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using pyterrier, in: K. Balog, V. Setty, C. Lioma, Y. Liu, M. Zhang, K. Berberich (Eds.), ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020, ACM, 2020, pp. 161–168. URL: https://doi.org/10.1145/3409256.3409829. doi:10.1145/3409256.3409829.
[89] L. Gienapp, B. Stein, M. Hagen, M. Potthast, Efficient pairwise annotation of argument quality, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Association for Computational Linguistics, 2020, pp. 5772–5781. URL: https://www.aclweb.org/anthology/2020.acl-main.511/.
[90] G. Amati, C. J. van Rijsbergen, Probabilistic models of information retrieval based on measuring the divergence from randomness, ACM Trans. Inf. Syst. 20 (2002) 357–389. URL: http://doi.acm.org/10.1145/582415.582416. doi:10.1145/582415.582416.
[91] M. Kaszkiel, J. Zobel, Passage retrieval revisited, in: N. J. Belkin, A. D. Narasimhalu, P. Willett, W. R. Hersh, F. Can, E. M. Voorhees (Eds.), Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1997, Philadelphia, PA, USA, July 27-31, 1997, ACM, 1997, pp. 178–185. URL: https://doi.org/10.1145/258525.258561.
[92] M. Fröbe, J. Bevendorff, L. Gienapp, M. Völske, B. Stein, M. Potthast, M. Hagen, Copycat: Near-duplicates within and between the clueweb and the common crawl, in: Proceedings of the 44th International ACM Conference on Research and Development in Information Retrieval, SIGIR 2021, ACM, 2021, pp. 2398–2404. URL: https://dl.acm.org/doi/10.1145/3404835.3463246.
[93] M. Fröbe, J. Bevendorff, J. Reimer, M. Potthast, M. Hagen, Sampling bias due to near-duplicates in learning to rank, in: Proceedings of the 43rd International ACM Conference on Research and Development in Information Retrieval, SIGIR 2020, ACM, 2020, pp. 1997–2000. URL: https://dl.acm.org/doi/10.1145/3397271.3401212.
[94] M. Fröbe, J. Bittner, M. Potthast, M. Hagen, The effect of content-equivalent near-duplicates on the evaluation of search engines, in: Proceedings of the 42nd European Conference on IR Research (ECIR 2020), volume 12036 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 12–19. URL: https://link.springer.com/chapter/10.1007%2F978-3-030-45442-5_2.

[95] R. Nogueira, J. Lin, A. Epistemic, From doc2query to doctttttquery, Online preprint (2019). URL: https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery-v2.pdf.

[96] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated MAchine Reading COmprehension dataset, in: Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 at NIPS, volume 1773 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016. URL: http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

[97] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

[98] A. Chernodub, O. Oliynyk, P. Heidenreich, A. Bondarenko, M. Hagen, C. Biemann, A. Panchenko, TARGER: Neural argument mining at your fingertips, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, ACL, 2019, pp. 195–200. URL: https://doi.org/10.18653/v1/p19-3031.

[99] O. Khattab, M. Zaharia, ColBERT: efficient and effective passage search via contextualized late interaction over BERT, in: J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu (Eds.), Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, ACM, 2020, pp. 39–48. URL: https://doi.org/10.1145/3397271.3401075.

[100] R. Pradeep, R. Nogueira, J. Lin, The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models, CoRR abs/2101.05667 (2021). URL: https://arxiv.org/abs/2101.05667. arXiv:2101.05667.

[101] A. Rana, P. Golchha, R. Juntunen, A. Coajă, A. Elzamarany, C.-C. Hung, S. P. Ponzetto, LeviRank: Limited query expansion with voting integration for document retrieval and ranking, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.

[102] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly, 2009. URL: http://www.oreilly.de/catalog/9780596516499/index.html.

[103] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[104] A. Williams, N. Nangia, S. R. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Association for Computational Linguistics, 2018, pp. 1112–1122. URL: https://doi.org/10.18653/v1/n18-1101.

[105] M. Aba, M. Azra, M. Gallo, O. Mohammad, I. Piacere, G. Virginio, N. Ferro, Aldo Nadi at Touché 2022: Argument retrieval for comparative question, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.

[106] M. F. Porter, Snowball: A language for stemming algorithms, 2001. URL: http://snowball.tartarus.org/texts/introduction.html.

[107] J. Rocchio, Relevance feedback in information retrieval, The Smart retrieval system-

experiments in automatic document processing (1971) 313–323.
[108] G. V. Cormack, C. L. A. Clarke, S. Büttcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, ACM, 2009, pp. 758–759. URL: https://doi.org/10.1145/1571941.1572114.
[109] V. Chekalina, A. Panchenko, Retrieving comparative arguments using deep language models, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[110] A. Chimetto, D. Peressoni, E. Sabbatini, G. Tommasin, M. Varotto, A. Zanardelli, N. Ferro, SEUPD@CLEF: Team Hextech on argument retrieval for comparative questions. the importance of adjectives in documents quality evaluation, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[111] J. B. Lovins, Development of a stemming algorithm, Mech. Transl. Comput. Linguistics 11 (1968) 22–31. URL: http://www.mt-archive.info/MT-1968-Lovins.pdf.
[112] P. Rajula, C.-C. Hung, S. P. Ponzetto, Stacked model based argument extraction and stance detection using embedded LSTM model, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[113] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108 (2019). URL: http://arxiv.org/abs/1910.01108. arXiv:1910.01108.
[114] A. Alhamzeh, M. Bouhaouel, E. Egyed-Zsigmond, J. Mitrovic, Distilbert-based argumentation retrieval for answering comparative questions, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 2319–2330. URL: http://ceur-ws.org/Vol-2936/paper-209.pdf.
[115] J. H. Reimer, J. Huck, A. Bondarenko, Grimjack at Touché 2022: Axiomatic re-ranking and query reformulation, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[116] A. Bondarenko, M. Fröbe, V. Kasturia, M. Völske, B. Stein, M. Hagen, Webis at TREC 2019: Decision track, in: E. Voorhees, A. Ellis (Eds.), Proceedings of the 28th International Text Retrieval Conference, TREC 2019, NIST, 2019.
[117] J. Bevendorff, A. Bondarenko, M. Fröbe, S. Günther, M. Völske, B. Stein, M. Hagen, Webis at TREC 2020: Health Misinformation track, in: E. Voorhees, A. Ellis (Eds.), Proceedings of the 29th International Text Retrieval Conference, TREC 2020, NIST, 2020.
[118] P. Rösner, N. Arnhold, T. Xylander, Quality-aware argument re-ranking for comparative questions, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.
[119] G. Erkan, D. R. Radev, LexRank: Graph-based lexical centrality as salience in text summarization, J. Artif. Intell. Res. 22 (2004) 457–479. URL: https://doi.org/10.1613/jair.1523.
[120] G. V. Cormack, M. D. Smucker, C. L. A. Clarke, Efficient and effective spam filtering and re-ranking for large web datasets, Inf. Retr. 14 (2011) 441–465. URL: https://doi.org/10.1007/s10791-011-9162-z.
[121] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003)

993–1022. URL: http://jmlr.org/papers/v3/blei03a.html.

[122] T. K. Ho, Random decision forests, in: Third International Conference on Document Analysis and Recognition, ICDAR 1995, August 14 - 15, 1995, Montreal, Canada. Volume I, IEEE Computer Society, 1995, pp. 278–282. URL: https://doi.org/10.1109/ICDAR.1995.598994.

[123] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, E. Hovy, Learning whom to trust with MACE, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HTL 2013), Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 1120–1130. URL: https://aclanthology.org/N13-1132.

[124] J. Braker, L. Heinemann, T. Schreieder, Aramis at Touché 2022: Argument detection in pictures using machine learning, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.

[125] M. Zaid, L. George, G. Al-Khafaji, Distinguishing cartoons images from real-life images, International Journal of Advanced Research in Computer Science and Software Engineering 5 (2015) 91–95.

[126] T. Brummerloh, M. L. Carnot, S. Lange, G. Pfänder, Boromir at Touché 2022: Combining natural language processing and machine learning techniques for image retrieval for arguments, in: Working Notes Papers of the CLEF 2022 Evaluation Labs, CEUR Workshop Proceedings, 2022.

[127] F. Å. Nielsen, A new ANEW: evaluation of a word list for sentiment analysis in microblogs, in: M. Rowe, M. Stankovic, A. Dadzie, M. Hardey (Eds.), Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts', volume 718 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2011, pp. 93–98. URL: http://ceur-ws.org/Vol-718/paper_16.pdf.

A. Full Evaluation Results of Touché 2022: Argument Retrieval

Table 6
Relevance results of all runs submitted to Task 1: Argument Retrieval for Controversial Questions. Reported are the mean nDCG@5 and the 95% confidence intervals. The baseline Swordsman is shown in bold.

Team	Run Tag	nDCG@5		
		Mean	Low	High
Porthos	scl_dlm_bqnc_acl_nsp	0.742	0.670	0.807
Daario Naharis	INTSEG-Letter-no_stoplist-Krovetz-Icoef-Evidence-Par	0.683	0.609	0.755
Daario Naharis	INTSEG-Run-Whitespace-Krovetz-Stoplist-Pos-Evidence-icoeff-Sep	0.676	0.587	0.762
Daario Naharis	INTSEG-Run-letter-english-2-20-no_stoplist-pos-evidence-icoef-An	0.670	0.589	0.751
Bruce Banner	Bruce-Banner_pyserinin_sparse_v3	0.651	0.573	0.720
D'Artagnan	seupd2122-6musk-kstem-stop-shingle3	0.642	0.575	0.705
Bruce Banner	Bruce-Banner_pyserinin_sparse_v1	0.641	0.575	0.705
Daario Naharis	INTSEG-Whitespace-Stoplist-Krovetz-Icoef-Sep	0.629	0.549	0.706
Gamora	seupd2122-javacafe-gamoraHeuristicsOnlyQueryReductionDoubleIndex	0.616	0.551	0.687
D'Artagnan	seupd2122-6musk-stop-wordnet-kstem-dirichlet	0.608	0.542	0.682
D'Artagnan	seupd2122-6musk-stop-kstem-concsearch	0.591	0.514	0.667
Hit Girl	Io	0.588	0.515	0.657
Gamora	seupd2122-javacafe-gamoraStandardDoubleIndex	0.588	0.518	0.655
Bruce Banner	Bruce-Banner_pyserinin_sparse_v4	0.586	0.509	0.658
D'Artagnan	seupd2122-6musk-word2vec-sentences-kstem	0.585	0.506	0.661
Gamora	seupd2122-javacafe-gamoraHeuristicsDoubleIndex	0.584	0.512	0.656
Hit Girl	Ganymede	0.583	0.513	0.650
Bruce Banner	Bruce-Banner_pyserinin_sparse_v2	0.580	0.507	0.654
Hit Girl	Jupiter	0.560	0.484	0.631
Hit Girl	Europa	0.546	0.477	0.615
Gamora	seupd2122-javacafe-gamora_tfidf_kstemstopengpos_multi_YYY	0.516	0.446	0.581
Gamora	seupd2122-javacafe-gamora_sbert_kstemstopengpos_multi_YYY	0.497	0.407	0.586
Pearl	PearlBlocklist_WeightedRelevance	0.481	0.399	0.560
Pearl	PearlArgRank8040_WeightedRelevance	0.479	0.403	0.556
Pearl	PearlArgRank7530	0.470	0.391	0.547
Pearl	PearlBlocklist	0.466	0.380	0.549
Pearl	PearlArgRank8040	0.465	0.389	0.551
Gorgon	GorgonA2Bm25	0.408	0.354	0.461
Daario Naharis	INTSEG-Run-Whitespace-Porter-Wordnet-Pos-no_stoplist-tfidf-An	0.406	0.305	0.510
General Grievous	seupd2122-lgtm_QE_NRR	0.403	0.335	0.471
General Grievous	seupd2122-lgtm_NQE_NRR	0.402	0.335	0.476
Gorgon	GorgonA1Bm25	0.396	0.350	0.442
Gorgon	GorgonBasicBM25	0.387	0.330	0.439
General Grievous	seupd2122-lgtm_NQE_NRR_ONLY_TITLE	0.386	0.314	0.451
General Grievous	seupd2122-lgtm_QE_NRR_ONLY_TITLE	0.386	0.317	0.450
Gorgon	GorgonKEBM25	0.378	0.329	0.428
Swordsman	**baseline_swordsman**	**0.356**	**0.296**	**0.412**
Gorgon	GorgonBasicLMD	0.315	0.269	0.362
D'Artagnan	seupd2122-6musk-stop-kstem-basic	0.300	0.229	0.369
Korg	korg9000	0.252	0.187	0.318
Porthos	scl_dlm_bqnc_acl_nsp_100_test	0.244	0.215	0.275

Table 7
Quality results of all runs submitted to Task 1: Argument Retrieval for Controversial Questions. Reported are the mean nDCG@5 and the 95% confidence intervals. The baseline Swordsman is shown in bold.

Team	**Run Tag**	**nDCG@5**		
		Mean	Low	High
Daario Naharis	INTSEG-Letter-no_stoplist-Krovetz-Icoef-Evidence-Par	0.913	0.870	0.947
Daario Naharis	INTSEG-Run-letter-english-2-20-no_stoplist-pos-evidence-icoef-An	0.898	0.855	0.941
Daario Naharis	INTSEG-Run-Whitespace-Krovetz-Stoplist-Pos-Evidence-icoeff-Sep	0.896	0.841	0.944
Porthos	scl_dlm_bqnc_acl_nsp	0.873	0.825	0.913
Gamora	seupd2122-javacafe-gamoraHeuristicsOnlyQueryReductionDoubleIndex	0.785	0.729	0.848
Gamora	seupd2122-javacafe-gamoraHeuristicsDoubleIndex	0.779	0.716	0.835
Daario Naharis	INTSEG-Whitespace-Stoplist-Krovetz-Icoef-Sep	0.776	0.712	0.839
Hit Girl	Ganymede	0.776	0.707	0.840
Bruce Banner	Bruce-Banner_pyserinin_sparse_v1	0.772	0.702	0.830
Bruce Banner	Bruce-Banner_pyserinin_sparse_v3	0.760	0.680	0.832
Gamora	seupd2122-javacafe-gamora_tfidf_kstemstopengpos_multi_YYY	0.755	0.686	0.823
Gamora	seupd2122-javacafe-gamora_sbert_kstemstopengpos_multi_YYY	0.743	0.656	0.823
Gorgon	GorgonA2Bm25	0.742	0.700	0.786
D'Artagnan	seupd2122-6musk-stop-wordnet-kstem-dirichlet	0.733	0.676	0.787
Gamora	seupd2122-javacafe-gamoraStandardDoubleIndex	0.731	0.672	0.786
Gorgon	GorgonA1Bm25	0.729	0.686	0.774
D'Artagnan	seupd2122-6musk-kstem-stop-shingle3	0.728	0.657	0.794
D'Artagnan	seupd2122-6musk-stop-kstem-concsearch	0.727	0.659	0.786
Hit Girl	Jupiter	0.725	0.651	0.796
Gorgon	GorgonKEBM25	0.724	0.677	0.769
D'Artagnan	seupd2122-6musk-word2vec-sentences-kstem	0.723	0.659	0.787
Hit Girl	Europa	0.721	0.643	0.793
Hit Girl	Io	0.719	0.643	0.797
Bruce Banner	Bruce-Banner_pyserinin_sparse_v4	0.709	0.624	0.783
Bruce Banner	Bruce-Banner_pyserinin_sparse_v2	0.701	0.610	0.783
Gorgon	GorgonBasicBM25	0.685	0.634	0.734
Gorgon	GorgonBasicLMD	0.679	0.634	0.726
Pearl	PearlArgRank7530	0.678	0.609	0.744
Daario Naharis	INTSEG-Run-Whitespace-Porter-Wordnet-Pos-no_stoplist-tfidf-An	0.671	0.585	0.753
Pearl	PearlBlocklist_WeightedRelevance	0.670	0.605	0.734
Pearl	PearlBlocklist	0.670	0.605	0.729
Pearl	PearlArgRank8040_WeightedRelevance	0.670	0.601	0.735
Pearl	PearlArgRank8040	0.668	0.595	0.737
Swordsman	**baseline_swordsman**	**0.608**	**0.543**	**0.671**
General Grievous	seupd2122-lgtm_QE_NRR	0.517	0.444	0.583
General Grievous	seupd2122-lgtm_NQE_NRR	0.517	0.442	0.591
General Grievous	seupd2122-lgtm_NQE_NRR_ONLY_TITLE	0.475	0.387	0.559
General Grievous	seupd2122-lgtm_QE_NRR_ONLY_TITLE	0.475	0.392	0.555
Korg	korg9000	0.453	0.384	0.529
D'Artagnan	seupd2122-6musk-stop-kstem-basic	0.441	0.357	0.517
Porthos	scl_dlm_bqnc_acl_nsp_100_test	0.274	0.247	0.301

Table 8
Coherence results of all runs submitted to Task 1: Argument Retrieval for Controversial Questions. Reported are the mean nDCG@5 and the 95% confidence intervals. The baseline Swordsman is shown in bold.

Team	**Run Tag**	**nDCG@5**		
		Mean	Low	High
Daario Naharis	INTSEG-Run-Whitespace-Krovetz-Stoplist-Pos-Evidence-icoeff-Sep	0.458	0.389	0.525
Daario Naharis	INTSEG-Letter-no_stoplist-Krovetz-Icoef-Evidence-Par	0.444	0.375	0.508
Porthos	scl_dlm_bqnc_acl_nsp	0.429	0.353	0.509
Daario Naharis	INTSEG-Run-letter-english-2-20-no_stoplist-pos-evidence-icoef-An	0.407	0.331	0.489
Pearl	PearlArgRank7530	0.398	0.311	0.485
Pearl	PearlArgRank8040	0.396	0.311	0.481
Pearl	PearlBlocklist	0.392	0.307	0.475
D'Artagnan	seupd2122-6musk-kstem-stop-shingle3	0.378	0.311	0.452
Bruce Banner	Bruce-Banner_pyserinin_sparse_v1	0.378	0.300	0.459
Hit Girl	Ganymede	0.377	0.303	0.456
Pearl	PearlBlocklist_WeightedRelevance	0.369	0.287	0.450
Pearl	PearlArgRank8040_WeightedRelevance	0.369	0.291	0.443
Hit Girl	Io	0.365	0.302	0.430
D'Artagnan	seupd2122-6musk-stop-wordnet-kstem-dirichlet	0.358	0.292	0.427
Bruce Banner	Bruce-Banner_pyserinin_sparse_v4	0.357	0.273	0.446
Bruce Banner	Bruce-Banner_pyserinin_sparse_v3	0.354	0.272	0.444
Bruce Banner	Bruce-Banner_pyserinin_sparse_v2	0.353	0.283	0.433
Hit Girl	Europa	0.349	0.287	0.415
D'Artagnan	seupd2122-6musk-stop-kstem-concsearch	0.336	0.270	0.400
D'Artagnan	seupd2122-6musk-word2vec-sentences-kstem	0.333	0.274	0.400
Hit Girl	Jupiter	0.330	0.269	0.394
Daario Naharis	INTSEG-Whitespace-Stoplist-Krovetz-Icoef-Sep	0.288	0.216	0.361
Gamora	seupd2122-javacafe-gamora_sbert_kstemstopengpos_multi_YYY	0.285	0.203	0.373
Gorgon	GorgonKEBM25	0.282	0.233	0.335
Gamora	seupd2122-javacafe-gamoraHeuristicsOnlyQueryReductionDoubleIndex	0.276	0.204	0.347
Gamora	seupd2122-javacafe-gamora_tfidf_kstemstopengpos_multi_YYY	0.276	0.200	0.372
Gorgon	GorgonBasicBM25	0.274	0.210	0.334
D'Artagnan	seupd2122-6musk-stop-kstem-basic	0.273	0.207	0.346
Gamora	seupd2122-javacafe-gamoraHeuristicsDoubleIndex	0.272	0.203	0.343
Gorgon	GorgonA2Bm25	0.259	0.209	0.314
Swordsman	**baseline_swordsman**	**0.248**	**0.193**	**0.303**
Gorgon	GorgonA1Bm25	0.246	0.197	0.301
General Grievous	seupd2122-lgtm_QE_NRR	0.231	0.162	0.313
General Grievous	seupd2122-lgtm_NQE_NRR	0.228	0.164	0.299
Gorgon	GorgonBasicLMD	0.225	0.162	0.289
General Grievous	seupd2122-lgtm_NQE_NRR_ONLY_TITLE	0.220	0.158	0.283
General Grievous	seupd2122-lgtm_QE_NRR_ONLY_TITLE	0.219	0.160	0.283
Daario Naharis	INTSEG-Run-Whitespace-Porter-Wordnet-Pos-no_stoplist-tfidf-An	0.203	0.137	0.280
Gamora	seupd2122-javacafe-gamoraStandardDoubleIndex	0.195	0.139	0.252
Korg	korg9000	0.168	0.117	0.223
Porthos	scl_dlm_bqnc_acl_nsp_100_test	0.105	0.070	0.144

Table 9
Relevance results of all runs submitted to Task 2: Argument Retrieval for Comparative Questions. Reported are the mean nDCG@5 and the 95% confidence intervals; Puss in Boots baseline in bold.

Team	**Run Tag**	**nDCG@5**		
		Mean	Low	High
Captain Levi	levirank_dense_initial_retrieval	0.758	0.708	0.805
Captain Levi	levirank_baseline_large_duo_t5	0.755	0.711	0.805
Captain Levi	levirank_psuedo_relevance_feedback+voting	0.753	0.713	0.797
Captain Levi	levirank_voting_retrieval	0.727	0.674	0.779
Captain Levi	levirank_psuedo_relevance_feedback	0.722	0.663	0.777
Aldo Nadi	seupd2122-kueri_rrf_reranked	0.709	0.648	0.766
Aldo Nadi	seupd2122-kueri_RF_reranked	0.695	0.629	0.756
Aldo Nadi	seupd2122-kueri_rrf	0.668	0.591	0.744
Aldo Nadi	seupd2122-kueri_[…]_porter_reranked	0.636	0.568	0.701
Katana	Colbert edinburg	0.618	0.553	0.678
Katana	Colbert trained by me	0.601	0.532	0.674
Captain Tempesta	hextech_run_1	0.574	0.499	0.641
Captain Tempesta	hextech_run_2	0.569	0.499	0.633
Captain Tempesta	hextech_run_3	0.564	0.488	0.635
Katana	Colbert fine tune on touche data	0.562	0.488	0.630
Captain Tempesta	hextech_run_5	0.557	0.483	0.624
Aldo Nadi	seupd2122-kueri_[…]_porter	0.546	0.473	0.620
Captain Tempesta	hextech_run_4	0.536	0.460	0.609
Olivier Armstrong	tfid_arg_similarity	0.492	0.422	0.564
Puss in Boots	**BM25-Baseline**	**0.469**	**0.403**	**0.535**
Grimjack	grimjack-fair-reranking-argumentative-axioms	0.422	0.349	0.500
Grimjack	grimjack-argumentative-axioms	0.376	0.299	0.455
Grimjack	grimjack-baseline	0.376	0.301	0.459
Grimjack	grimjack-fair-argumentative-reranking-with-t0	0.349	0.270	0.425
Grimjack	grimjack-all-you-need-is-t0	0.345	0.273	0.425
Asuna	asuna-run-5	0.263	0.198	0.328

Table 10
Quality results of all runs submitted to Task 2: Argument Retrieval for Comparative Questions. Reported are the mean nDCG@5 and the 95% confidence intervals; Puss in Boots baseline in bold.

Team	**Run Tag**	**nDCG@5**		
		Mean	Low	High
Aldo Nadi	seupd2122-kueri_RF_reranked	0.774	0.717	0.829
Aldo Nadi	seupd2122-kueri_[…]_porter_reranked	0.764	0.701	0.823
Aldo Nadi	seupd2122-kueri_rrf_reranked	0.748	0.687	0.807
Captain Levi	levirank_dense_initial_retrieval	0.744	0.694	0.804
Captain Levi	levirank_baseline_large_duo_t5	0.742	0.681	0.800
Captain Levi	levirank_psuedo_relevance_feedback+voting	0.730	0.672	0.789
Captain Levi	levirank_voting_retrieval	0.706	0.639	0.774
Captain Levi	levirank_psuedo_relevance_feedback	0.695	0.625	0.753
Aldo Nadi	seupd2122-kueri_rrf	0.664	0.589	0.735
Katana	Colbert trained by me	0.644	0.574	0.714
Katana	Colbert edinburg	0.643	0.577	0.709
Katana	Colbert fine tune on touche data	0.637	0.556	0.718
Captain Tempesta	hextech_run_5	0.597	0.521	0.676
Captain Tempesta	hextech_run_2	0.593	0.518	0.670
Captain Tempesta	hextech_run_1	0.589	0.508	0.667
Captain Tempesta	hextech_run_3	0.584	0.506	0.660
Olivier Armstrong	tfid_arg_similarity	0.582	0.502	0.656
Aldo Nadi	seupd2122-kueri_[…]_porter	0.570	0.490	0.647
Captain Tempesta	hextech_run_4	0.566	0.490	0.641
Puss in Boots	**BM25-Baseline**	**0.476**	**0.400**	**0.553**
Grimjack	grimjack-fair-reranking-argumentative-axioms	0.403	0.331	0.478
Grimjack	grimjack-fair-argumentative-reranking-with-t0	0.365	0.290	0.445
Grimjack	grimjack-argumentative-axioms	0.363	0.289	0.442
Grimjack	grimjack-baseline	0.363	0.287	0.443
Grimjack	grimjack-all-you-need-is-t0	0.344	0.266	0.428
Asuna	asuna-run-5	0.332	0.254	0.417

Table 11
Stance detection results of all runs submitted to Task 2: Argument Retrieval for Comparative Questions. Reported are a macro-averaged F_1 for each team and run and number of documents N for which the stance was predicted; Puss in Boots baseline that always predicts 'no stance' is in bold.

Team	Tag	F_1 run	N run	F_1 team	N team
Grimjack	grimjack-all-you-need-is-t0	0.313	1208	0.235	1386
Captain Levi	levirank_dense_initial_retrieval	0.301	1688	0.261	2020
Captain Levi	levirank_baseline_large_duo_t5	0.295	1960	0.261	2020
Captain Levi	levirank_psuedo_relevance_feedback	0.246	1948	0.261	2020
Captain Levi	levirank_voting_retrieval	0.236	1897	0.261	2020
Katana	Colbert edinburg	0.229	1027	0.220	1301
Katana	Colbert trained by me	0.221	1079	0.220	1301
Captain Levi	levirank_psuedo_relevance_feedback+voting	0.218	1822	0.261	2020
Katana	Colbert fine tune on touche data	0.212	940	0.220	1301
Grimjack	grimjack-argumentative-axioms	0.207	1282	0.235	1386
Grimjack	grimjack-baseline	0.207	1282	0.235	1386
Grimjack	grimjack-fair-reranking-argumentative-axioms	0.207	1282	0.235	1386
Grimjack	grimjack-fair-argumentative-reranking-with-t0	0.199	1180	0.235	1386
Olivier Armstrong	tfid_arg_similarity	0.191	551	0.191	551
Puss in Boots	**Always-NO-Baseline**	**0.158**	**1328**	**0.158**	**1328**
Asuna	asuna-run-5	0.106	578	0.106	578

Table 12
Results of all runs submitted to Task 3 Image Retrieval. Reported are the mean precision@10 (per stance) for topic relevance, argumentativeness, and stance relevance and the 95% confidence intervals (low and high). Results for the baseline are shown in bold.

		Precision@10								
		Topic			Arg.			Stance		
Team	Run	Mean	Low	High	Mean	Low	High	Mean	Low	High
Boromir	BERT, OCR, query-processing	0.878	0.847	0.904	0.768	0.733	0.799	0.425	0.398	0.451
Boromir	BERT, OCR, clustering, query-processing	0.822	0.782	0.863	0.728	0.685	0.772	0.411	0.383	0.442
Boromir	AFINN, OCR	0.814	0.774	0.851	0.726	0.680	0.768	0.408	0.379	0.436
Minsc	**Baseline**	**0.736**	**0.693**	**0.774**	**0.686**	**0.638**	**0.734**	**0.407**	**0.367**	**0.445**
Boromir	AFINN, OCR, clustering	0.749	0.705	0.792	0.674	0.625	0.721	0.384	0.354	0.414
Boromir	AFINN, OCR, clustering, query-processing	0.767	0.722	0.812	0.688	0.645	0.734	0.382	0.352	0.412
Aramis	Argumentativeness:formula, stance:formula	0.701	0.658	0.744	0.634	0.594	0.674	0.381	0.349	0.412
Aramis	Argumentativeness:neural, stance:formula	0.687	0.640	0.732	0.632	0.587	0.674	0.365	0.332	0.395
Aramis	Argumentativeness:neural, stance:neural	0.673	0.629	0.717	0.624	0.583	0.666	0.354	0.320	0.385
Jester	With emotion detection	0.696	0.654	0.736	0.647	0.601	0.688	0.350	0.316	0.382
Aramis	Argumentativeness:formula, stance:neural	0.664	0.622	0.710	0.609	0.568	0.646	0.344	0.317	0.371
Jester	Without emotion detection	0.671	0.635	0.712	0.618	0.577	0.656	0.336	0.308	0.366
Boromir	AFINN, clustering	0.600	0.549	0.649	0.545	0.495	0.595	0.319	0.285	0.351