Overview of Touché 2022: Argument Retrieval

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 touche.webis.de

Abstract This paper is a condensed report on the third year of the Touché lab on argument retrieval held at CLEF 2022. With the goal to foster and support the development of technologies for argument mining and argument analysis, we organized three shared tasks in the third edition of Touché: (a) argument retrieval for controversial topics, where participants retrieve a gist of arguments from a collection of online debates, (b) argument retrieval for comparative questions, where participants retrieve argumentative passages from a generic web crawl, and (c) image retrieval for arguments, where participants retrieve images from a focused web crawl that show support or opposition to some stance.

1 Introduction

Decision making and opinion formation are routine human tasks that often involve weighing pro and con arguments. Since the Web is full of argumentative texts on almost any topic, everybody has, in principle, the chance to acquire knowledge to come to informed decisions or opinions by simply using a search engine. However, large amounts of the arguments accessible easily may be of low quality. For example, they may be irrelevant, contain incoherent logic, provide insufficient support, or use foul language. Such arguments should rather remain "invisible" in search results which implies several retrieval challenges—regardless of whether a query is about socially important topics or "only" about personal decisions. The challenges range from assessing an argument's relevance to a query and estimating how well an implied stance is justified, to identifying what is the main "gist" of an argument's reasoning as well as finding images that help to illustrate some stance. Still, today's popular web search engines do not really

address these challenges, thus lacking a sophisticated support for searchers in argument retrieval scenarios—a gap we aim to close with the Touché labs.[8]

In the spirit of the two successful Touché labs on argument retrieval at CLEF 2020 and 2021 [13, 16], we organized a third lab edition to again bring together researchers from the fields of information retrieval and natural language processing who work on argumentation. At Touché 2022, we organized the following three shared tasks, the last of which being fully new to this edition:

1. Argumentative sentence retrieval from a focused collection (crawled from debate portals) to support argumentative conversations on controversial topics.
2. Argument retrieval from a large collection of text passages to support answering comparative questions in the scenario of personal decision making.
3. Image retrieval to corroborate and strengthen textual arguments and to provide a quick overview of public opinions on controversial topics.

In the Touché lab, we followed the classic TREC-style[9] methodology: documents and topics were provided to the participants who then submitted their ranked results (up to five runs) for every topic to be judged by human assessors. While the first two Touché editions focused on retrieving complete arguments and documents, the third edition focused on more refined problems. Three shared tasks explored whether argument retrieval can support decision making and opinion formation more directly by extracting the argumentative gist from documents, by classifying their stance as pro or con towards the issue in question, and by retrieving images that show support or opposition to some stance.

The teams that participated in the third year of Touché were able to use the topics as well as the relevance and argument quality judgments from the previous lab editions to improve their approaches. Only a few decided to train and optimize their pipelines using the judgments provided, though. Alongside sparse retrieval models like BM25 [70], this year approaches focus on more recent Transformer-based models, such as T5 [67] and T0 [76] in zero-shot settings, to predict relevance, argument quality, and stance. Also many re-ranking methods are proposed based on a wide range of diverse characteristics including a word mover's distance, linguistic properties of documents, as well as document "argumentativeness" and argument quality. A more comprehensive overview of all submitted approaches is covered in the extended overview [15].

2 Related Work

Queries in argument retrieval often may be phrases that describe a controversial topic, questions that ask to compare two options, or even statements that capture complete claims or short arguments [85]. In the Touché lab, we address the first two types in three different shared tasks. Here, we briefly summarize the related work for all three tasks.

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

[9] https://trec.nist.gov/tracks.html

2.1 Argument Retrieval

The goal of argument retrieval is to deliver arguments to support users in making a decision or in persuading an audience of a specific point of view. An argument is usually modeled as a conclusion with one or more supporting or attacking premises [83]. While a conclusion is a statement that can be accepted or rejected, a premise is a more grounded statement (e.g., statistical evidence).

The development of an argument search engine is faced with challenges that range from identifying argumentative queries [2] to mining arguments from unstructured text to assessing their relevance and quality [83]. Argument retrieval follows several paradigms that start from different sources and perform argument mining and retrieval tasks in different orders [3]. Wachsmuth et al. [83], for instance, extract arguments offline using heuristics that are tailored to online debate portals. Their argument search engine args.me uses BM25F [71] to rank the indexed arguments, giving conclusions more weight than premises. Also Levy et al. [48] use distant supervision to mine arguments offline for a set of topics from Wikipedia before ranking. Following a different paradigm, Stab et al. [79] retrieve documents from the Common Crawl[10] in an online fashion (no prior offline argument mining) and use a topic-dependent neural network to extract arguments from the retrieved documents at query time. With the three Touché tasks, we address the paradigms of Wachsmuth et al. [83] (Task 1) and Stab et al. [79] (Tasks 2 and 3), respectively.

Argument retrieval should rank arguments according to their topical relevance but also to their quality. What makes a good argument has been studied since the time of Aristotle [6]. Wachsmuth et al. [81] categorized 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. Rhetoric covers the effectiveness of the argument in persuading an audience with its conclusion. Dialectic, finally, 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. The relevance of an argument to a query's topic is categorized under dialectical quality [81].

Researchers assess argument relevance by measuring an argument's similarity to a query's topic or by incorporating its support and attack relations to other arguments. Potthast et al. [63] evaluate four standard retrieval models for ranking arguments with regard to four quality dimensions: relevance, logic, rhetoric, and dialectic. One of the main findings is that DirichletLM is better at ranking arguments than BM25, DPH, and TF-IDF. Gienapp et al. [32] extend this work by proposing a pairwise strategy that reduces the costs of crowdsourcing argument retrieval annotations in a pairwise fashion by 93% (i.e., annotating only a small subset of argument pairs).

Wachsmuth et al. [84] create a graph of arguments by connecting two arguments when one uses the other's conclusion as a premise. They exploit this

[10] http://commoncrawl.org

structure to rank the arguments in the graph using PageRank scores [58]. This method is shown to outperform baselines that only consider the content of the argument and its internal structure (conclusion and premises). Dumani et al. [25] introduce a probabilistic framework that operates on semantically similar claims and premises and that utilizes support and attack relations between clusters of premises and claims as well as between clusters of claims and a query. It is found to outperform BM25 in ranking arguments. Later, Dumani and Schenkel [26] also proposed an extension of the framework to include the quality of a premise as a probability by using the fraction of premises which are worse with regard to three quality dimensions: cogency, reasonableness, and effectiveness. Using a pairwise quality estimator trained on the Dagstuhl-15512 ArgQuality Corpus [82], their probabilistic framework with the argument quality component outperformed the one without on the 50 Task 1 topics of Touché 2020.

2.2 Retrieval for Comparisons

Comparative information needs in web search have first been addressed by basic interfaces where two products to be compared are entered separately in a left and a right search box [55, 80]. Comparative sentences are then identified and mined from product reviews in favor or against one or the other product using opinion mining approaches [39, 40, 42]. Recently, the identification of the comparison preference (the "winning" entity) in comparative sentences has been tackled in a more broad domain (not just product reviews) by applying feature-based and neural classifiers [60, 52]. Such preference classification forms the basis of the comparative argumentation machine CAM [77] that takes two entities and some comparison aspect(s) as input, retrieves comparative sentences in favor of one or the other entity using BM25, and then classifies their preference for a final merged result table presentation. A proper argument ranking, however, is still missing in CAM. Chekalina et al. [18] later extend the system to accept comparative questions as input and to return a natural language answer to the user. A comparative question is parsed by identifying the comparison objects, aspect(s), and predicate. The system's answer is either generated directly based on Transformers [22] or by retrieval from an index of comparative sentences. Identifying comparative information needs in question queries is proposed by Bondarenko et al. [12] and Bondarenko et al. [11] who study such information needs in a search engine log, propose a cascading ensemble of classifiers (rule-based, feature-based, and neural models) that identifies comparative questions, and label a respective dataset. They also propose an approach to identify entities of interest such as comparison objects, aspects, and predicates in comparative questions and to detect the stance of potential answers towards the comparison objects. The respective stance dataset is provided for Touché Task 2 participants to train their approaches for the stance classification of retrieved passages.

2.3 Image Retrieval

Images can provide contextual information and express, underline, or popularize an opinion [24], thereby taking the form of subjective statements [27]. Some images express both a premise and a conclusion, making them full arguments [73, 35]. Other images may provide contextual information only and have to be combined with a textual conclusion to form a complete argument. In this regard, a recent SemEval task distinguished a total of 22 persuasion techniques in memes alone [23]. Moreover, argument quality dimensions like acceptability, credibility, emotional appeal, and sufficiency [82] all apply to arguments that include images as well.

Keyword-based image search by analyzing the content of images or videos has been studied for decades [1], pre-dated only by approaches relying on metadata and similarity measures [17]. Early approaches exploited keyword-based web search (e.g., by Yanai [88]). In a recent survey, Latif et al. [46] categorize image features into color, texture, shape, and spatial features. Current commercial search engines also index text found in images, surrounding text, alternative texts displayed when an image is unavailable, and their URLs [87, 34]. As for the retrieval of argumentative images, a closely related concept is "emotional images", which is based on image features like color and composition [86, 78]. Since argumentation goes hand in hand with emotions, those emotional features may be promising for retrieving images for arguments in the future. To retrieve images for arguments is a relatively new task that has been recently proposed by Kiesel et al. [44], which forms the basis of the Touché Task 3.

3 Lab Overview and Statistics

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

Out of 58 registered teams, 23 actively participated in the tasks[11] and submitted their results (27 teams submitted in 2021 and 17 teams in 2020). Using the setup of the previous Touché editions, we encouraged the participants to submit software in TIRA [64] to improve the reproducibility of the developed approaches. TIRA is an integrated cloud-based Evaluation-as-a-Service research architecture where shared task participants can install their software on a dedicated virtual machine to which they have a full administrative access. By default,

[11] Three teams declined to proceed in the task after submitting the results.

the virtual machines run the server version of Ubuntu 20.04 with one CPU (Intel Xeon E5-2620), 4 GB RAM, and 16 GB HDD. However, we customized the resources as needed to meet participants' requirements. We pre-installed the latest versions of reasonable software in the virtual machines (e.g., Docker and Python) to simplify the deployment of the approaches within TIRA.

We allowed participants to submit software submissions and run file submissions in TIRA. For software submissions, participants created the run files with their software using the web UI of TIRA. The process for software submissions ensured that the software is fully installed in the virtual machine: the respective virtual machine is shut down, disconnected from the internet, powered on again in a sandbox mode, mounting the test datasets for the respective tasks. The interruption of the internet connection ensured that the participants' software worked 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 during deployment. Later, we archived the virtual machines that the participants used for their submissions such that the respective systems can be re-evaluated or applied to new datasets.

Overall, 9 of the 23 teams submitted traditional run files instead of software in TIRA. We allowed each team to submit up to 5 runs that should follow the standard TREC-style format.[12] We checked the validity of all submitted run files, asking participants to resubmit their run files (or software) if there were any problems—again, also offering our support in case of problems. All 23 teams submitted valid runs, resulting in 84 valid runs.

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 standpoints (pro or con) on controversial topics. The task we considered in this regard followed the idea of extractive argument summarization [5].

4.1 Task Definition

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 premise). Sentences in a pair may not contradict each other and ideally build upon each other in a logical manner comprising a coherent text.

4.2 Data Description

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

[12] The expected format of submissions was also described at https://touche.webis.de

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.

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 [3] that contains about 400,000 structured arguments (from debatewise.org, idebate.org, debatepedia.org, and debate.org). It is freely available for download[13] and can also be accessed through the args.me API.[14] To account for this year's changes in the task definition (the focus on gists), a pre-processed version of the corpus was created. Pre-processing 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.

4.3 Participant Approaches

This year's approaches included standard retrieval models such as TF-IDF, BM25, DirichletLM, and DPH. Participants also used multiple existing toolkits, such as the Project Debater API [7] for stance and evidence detection in arguments, Apache OpenNLP[15] for language detection, and classifiers proposed by Gienapp et al. [32] and Reimers et al. [69] trained on the IBM Rank 30K corpus [36] for argument quality detection. Additionally, semantic similarity of word and sentence embeddings based on doc2vec [47] and SBERT [68] was employed for retrieving coherent sentence pairs as required by the task definition. One team leveraged the text generation capabilities of GPT-2 [66] to find subsequent sentences while another team similarly used the next sentence prediction (NSP) of BERT [22] for this. These toolkits augmented the document pre-processing and re-ranking of the retrieved results.

[13] https://webis.de/data.html#args-me-corpus
[14] https://www.args.me/api-en.html
[15] https://opennlp.apache.org/

4.4 Task Evaluation

Participants submitted their rankings as classical 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, we created the pools using a top-5 pooling strategy for judgments with TrecTools [59], resulting in 6,930 unique documents for manual assessment of relevance, quality (argumentativeness), and textual coherence. Relevance was judged 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). Finally, 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).

4.5 Task Results

We used nDCG@5 for evaluation of relevance, quality, and coherence. Table 2 shows the results of the best run per team. For all evaluation categories at least eight out of ten teams managed to beat the provided baseline. Similar to previous years' results, quality appeared to be the evaluation category which is covered best by the approaches followed by relevance and the newly added coherence. A more comprehensive discussion including all teams' approaches is covered in the extended lab overview [15].

In terms of relevance Team *Porthos* achieved the highest results followed by Team *Daario Naharis* with nDCG@5 scores of 0.742 and 0.683 respectively. For quality and coherence Team *Daario Naharis* obtained the highest scores (0.913 and 0.458) followed by Team *Porthos* (0.873 and 0.429). The two-stage re-ranking employed by Team *Daario Naharis* improved coherence and quality in comparison to other approaches. They first ensured that retrieved pairs were relevant to their context in the argument alongside the topic that also boosted quality (argumentativeness). 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 argument's sentences by their centrality in the argument graph as proposed by Alshomary et al. [5]. The top two sentences are then retrieved as the final pair.

Team *Bruce Banner* employed BM25 retrieval model provided by the Pyserini toolkit [49][16] with its default parameters. Two query variants were used: standalone query and an expanded query (narrative and description appended). Likewise two variants of the sentence pairs were indexed: standalone pair and pair with the topic appended.

[16] https://pypi.org/project/pyserini/

Table 2. Results for Task 1 Argument Retrieval for Controversial Questions. Table shows the evaluation score of a team's best run for the three dimensions of relevance, quality, and coherence of the retrieved sentence pairs. Best scores per dimension are in bold. Team names are sorted alphabetically; the baseline Swordsman is emphasized.

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

Team *D'Artagnan* combined sparse retrieval with multiple text preprocessing and query expansion approaches. They used different combinations of retrieval models such as BM25 and DirichletLM, preprocessing steps, for instance, stemming, n-grams, and stopword removal, and query expansion with synonyms using WordNet [54] and word2vec [53]. Relevance judgments from the previous year were used for optimizing parameter values. Specifically, they used word and character n-grams (bi-grams and tri-grams) and built five different vocabularies for the word2vec model.

Team *Daario Naharis* developed a standard retrieval system using the Lucene TF-IDF implementation. Additionally, they introduced a new coefficient for scoring the discriminant power of a term. Re-ranking was performed based on stance detection using the Project Debater API. The highest nDCG@5 scores were achieved with a combination of the following components: Letter Tokenizer, English Stemmer, No Stop-List, POS Tag, WordNet, Evidence Detection, ICoefficient, and LMDirichlet Similarity.

Team *Gamora* developed Lucene-based approaches using deduplication and contextual feature-enriched indexing, adding the title of a discussion and the stance on the topic, to obtain document-level relevance and quality scores following the approach used in previous Touché editions [16]. 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 argumentative quality (using an SVR) of the pair and its source document. BM25 and DirichletLM were used for document similarity and SBERT [68] and TF-IDF for sentence agreement. The best approach is based on double indexing and a combination of

query reduction, query boosting, query decorators, query expansion with respect to important keywords and synonyms, and using the EnglishPossessiveFilter, LengthFilter and the Krovetz stemmer.

Team *General Grevious* used a conventional IR pipeline based on Lucene, extended with a LowerCaseFilter, an EnglishPossessiveFilter (removes possessive words (trailing 's) from words), and a LengthFilter (retains tokens between 3 and 20 characters in length and removes the others). BM25 and Dirichlet-based document relevance and sentence relevance were used for retrieval along with Rapid Automatic Keyword Extraction (RAKE) [74] query expansion. Sentiment analysis and readability analysis were used for re-ranking. However, their best model does not include re-ranking, but relies solely on query expansion.

Team *Gorgon* used the Lucene project for document retrieval and compared BM25 and LMDirichlet similarity measures, developing four different analyzers with combinations of the following components: LowercaseFilter, Krovetz stemmer, EnglishPossessiveFilter, StopwordFilter. Sentence pairs were created by creating all combinations within a single document before indexing. The best approach is a combination of the LowercaseFilter, EnglishPossessiveFilter and the similarity measure BM25.

Team *Hit Girl* proposed a two-stage retrieval pipeline that combines semantic search and re-ranking via argument quality agnostic models. Internal evaluation results showed that while re-ranking improved the argument quality to varying degrees, it affected the relevance. Additionally, they proposed a novel re-ranking method called structural distance which employs a fuzzy matching between query and the sentences based on part of speech tags. This performed best in comparison to standard methods such as maximal marginal relevance and word mover's distance.

Team *Korg* proposed to first use Elasticsearch[17] with the LM-Dirichlet similarity measure to find the best matching argumentative sentences for a query. Then, either doc2vec [47], trained on all sentences in the argsme corpus, or GPT-2 [66] was used to find similar sentences by direct comparison and by generation, respectively. AsciifoldingFilter and LowercaseFilter were used together with the Krovetz stemmer and a user-defined stopword list to preprocess the sentences. Their best approach is based on doc2vec's similarity calculation.

Team *Pearl* also proposed a two-stage retrieval pipeline using DirichletLM and DPH models to retrieve argumentative sentences. Argument quality scores were used as a pre-processing step to remove noisy examples. First, a vertical prototype was developed as a baseline model for revealing the weakness of the DPH model. Specifically, they found that this model assigns high relevance to sentences even if their terms are a part of a URL, or other sources in the text and is susceptible to homonyms thus negatively affecting the retrieval performance. To account for this, a refined prototype was developed that combines an argument quality prediction model and query expansion.

Team *Porthos* used Elasticsearch with DirichletLM and BM25 for retrieval after removing sentence duplicates and filtering irrelevant sentences by removing

[17] https://www.elastic.co/

sentences in incorrect language based on POS heuristics and their argumentativeness using the support vector machine (SVM) of [32] and the BERT approach of [69]. The approaches are based on a search term as a composition of single terms and Boolean queries together with [69] to reorder the retrieved sentences according to their argumentative quality. The sentences are paired with SBERT [68] and BERT [22] trained for the next sentence prediction task (NSP). The best approach is based on DirichletLM, NSP, using the sentence classifier in preprocessing, Boolean query with Noun Chunking for retrieval, and the BERT approach of [69] for re-ranking.

5 Task 2: Argument Retrieval for Comparative Questions

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

5.1 Task Definition

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

5.2 Data Description

Topics. For the task on comparative questions, we provided 50 search topics that described scenarios of personal decision making (cf. Table 3). Each of these topics had a *title* in terms of a comparative question, *comparison objects* for the stance detection of the retrieved passages, a *description* specifying the particular search scenario, and a *narrative* that served as a guideline for the assessors.

Document Collection. The collection for Task 2 was a focused collection of 868,655 passages extracted from the ClueWeb12[18] for the 50 search topics of the task. We constructed this passage corpus from 37,248 documents in the top-100 pools for all runs submitted in the previous Touché editions. Using the TREC CAsT tools[19]), we split the documents at the sentence boundary into fixed-length passages of approximately 250 terms since working with fixed-length passages is more effective than variable-length original passages [41]. From the initial 1,286,977 passages we removed near-duplicates with CopyCat [29] to mitigate negative impacts [30, 31], resulting in the final collection of 868,655 passages.

To lower the entry barrier of this task, we also provided the participants with a number of previously compiled resources. These included the document-level

[18] https://lemurproject.org/clueweb12/index.php
[19] https://github.com/grill-lab/trec-cast-tools

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.

relevance and argument quality judgments from the previous Touché editions as well as, for passage-level relevance judgments, a subset of MS MARCO [56] with comparative questions identified by our ALBERT-based [45] classifier (about 40,000 questions are comparative) [11]. Each comparative question in MS MARCO contains 10 text passages with relevance labels. For stance detection, a dataset comprising 950 comparative questions and answers extracted from Stack Exchange was provided [11]. For the identification of claims and premises, the participants could use any existing argument tagging tool, such as the API[20] of TARGER [19] hosted on our own servers, or develop an own method if necessary. Additionally, we provided the collection of 868,655 passages expanded with queries generated using the docT5query model [57].

5.3 Participant Approaches

For Task 2, seven teams submitted their results (25 valid runs). Interestingly, only two participants decided to use the relevance judgments from the previous Touché editions to fine-tune models or to optimize parameters. The others preferred to manually label a sample of retrieved documents themselves for the intermediate evaluation or relied on the zero-shot approaches such as the Transformer model T0++ [76]. Two teams also used the document collection expanded with docT5query [57] as a retrieval collection. Overall, the main trend of this year was using Transformer-based models for ranking and re-ranking such as ColBERT [43] and MonoT5 and DuoT5 [65]. The baseline retrieval approach was BM25. Five out of seven participants also submitted the results for stance

[20] Also available as a Python library: https://pypi.org/project/targer-api/

detection for retrieved passages (additional task). 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. The baseline stance detector simply output 'no stance' for all text passages.

5.4 Task Evaluation

Similar to Task 1, our volunteer human assessors labeled the relevance to a respective topic with three labels: 0 (not relevant), 1 (relevant), and 2 (highly relevant), and they assessed whether arguments are present in a result and whether they are rhetorically well-written [82] with three labels: 0 (low quality, or no arguments in a document), 1 (sufficient quality), and 2 (high quality). Additionally, we asked the assessors to label documents with respect to the comparison objects given in search topics as (a) pro first object (expresses a stronger positive attitude towards the first object), (b) pro second object (positive attitude towards the second object), (c) neutral (both comparison objects are equally good or bad), and (d) no stance (no attitude / opinion / argument towards the objects entailed). Following the strategy from Task 1, we pooled the top-5 documents from the runs resulting in 2,107 unique documents that were manually judged.

5.5 Task Results

For Task2, we used nDCG@5 to evaluate submitted rankings based on the relevance and argument quality judgments. The effectiveness of the stance detection approaches was evaluated using a macro-averaged F_1 score. Table 4 shows the results for the most effective runs of the participated teams based on the relevance and argument quality. For the stance detection (additional task) we evaluated all documents across all runs for each team that appeared in the top-5 pooling. A more comprehensive discussion including all teams' approaches is covered in the extended lab overview [15].

Team *Captian Levi* (submitted the relevance-wise most effective run) first retrieved 2,000 documents using Pyserini's BM25 [49] ($k_1 = 1.2$ and $b = 0.68$) by combining the top-1000 results for the original query (topic title) with the results for modified queries, where they (1) only removed stopwords (using the NLTK [9] stopword list), (2) replaced comparative adjectives with synonyms and antonyms found in WordNet [54], (3) added extra terms using pseudo-relevance feedback, and (4) used queries expanded with the docT5query model [57] provided by the Touché organizers. Queries and corpus were also processed by using stopwords and punctuation removal and lemmatization (with the WordNet lemmatizer). The initially retrieved results were re-ranked using monoT5 and duoT5 [65]. Additionally ColBERT [43] also was used for initial ranking. The team Captain Levi submitted in total 5 runs that differ in strategies of modifying queries, initial ranking models, and final re-ranking models. Finally, the stance was detected using the pre-trained RoBERTA-Large-MNLI language model [50] without fine-tuning in two steps: by first detecting if the document has a stance and after that for documents that were not classified as 'neutral' or 'no stance' detecting

Table 4. Results for Task 2 Argument Retrieval for Comparative Questions. The left part (a) shows the evaluation results of a team's best run according to the results' relevance, while the middle part (b) shows the best runs according to the results' quality, and the right part (c) shows the stance detection results (the teams' ordering is the same as in the part (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) Best relevance score per team

	nDCG@5	
Team	**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

	nDCG@5	
Team	**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

$\mathbf{F_1}$ **macro**
–
0.261
0.220
–
0.191
0.158
0.235
0.106

which comparison object the document favors. This stance detector achieved the highest macro-averaged F_1 score across all teams.

Team *Aldo Nadi* (submitted the quality-wise most effective run) re-ranked passages that were initially retrieved with BM25F [71] (default Lucene implementation with $k_1 = 1.2$ and $b = 0.75$) on two fields: the text of the original passages, and the passages expanded with docT5query. All texts were processed with the Porter stemmer [61], removing stopwords using different lists such as Snowball [62], a default Lucene stopword list, a custom list containing the 400 most frequent terms in the retrieval collection excluding the comparison objects contained in the 50 search topics, etc. Queries were expanded using a relevance feedback method that is based on the Rocchio Algorithm [72]. For the final ranking, the team experimented with re-ranking (up to top-1000 documents from the initial ranking) based on the argument quality by multiplying the relevance and the quality scores and Reciprocal Ranking Fusion [20]. The quality scores were predicted using the IBM Project Debater API [7]. Aldo Nadi submitted 5 runs, which vary by different combinations of the proposed methods, e.g., testing different stopword lists, using the quality-based re-ranking or fusion, etc. The team did not detect the stance.

Team *Katana* submitted in total 3 runs that all used different variants of ColBERT [43]: (1) pre-trained on MS MARCO [56] by the University on Glasgow,[21] (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

[21] http://www.dcs.gla.ac.uk/~craigm/colbert.dnn.zip

processing. For the stance detection Katana used a pre-trained XGBoost-based classifier that is part of Comparative Argumentation Machine [77, 60].

Team *Captain Tempesta* used linguistic properties of text such as non-informative symbol frequency (hashtags, emojis, etc.), the difference between the short word (less or equal than 4 characters) frequency and the long word (more than 4 characters) frequency, and adjective and comparative adjective frequencies. Based on these properties for each document in the retrieval corpus, the quality score was computed as a weighted sum (weights were assigned manually). At a query time, the relevance score of Lucene BM25 ($k_1 = 1.2$ and $b = 0.75$) was multiplied with the quality score; the final ranking was created by sorting documents by the descending final scores. Search queries were created by removing stopwords (Lucene default list) from topic titles and lowercasing query terms except for the brand names,[22] query terms were stemmed using Lovins stemmer [51]. The team's 5 submitted runs differ in the weights manually assigned for the different quality properties. They did not detect the stance.

Team *Olivier Armstrong* submitted one run. They first identified the compared objects, aspects, and predicates in queries (topic titles) using a RoBERTa-based classifier fine-tuned on the provided stance dataset. After removing stopwords, queries were expanded with synonyms found with WordNet. Then 100 documents were retrieved using Elasticsearch BM25 ($k_1 = 1.2$ and $b = 0.75$) as initial ranking. Using the DistilBERT-based classifier [75] fine-tuned by Alhamzeh et al. [4] (Touché 2021 participant), Olivier Armstrong identified premises and claims in the retrieved documents. Before the final ranking the following scores were calculated for each candidate document: (1) arg-BM25 score by querying the new re-indexed corpus (only premises and claims are kept) using the original queries, (2) argument support score, i.e., the ratio of premises and claims in the document, (3) similarity score, i.e., the averaged cosine similarity between the original query and every argumentative sentence in the document, both represented using the SBERT embeddings [68]. The final score was obtained by summing up the normalized individual scores. The final ranking included 25 documents sorted by the descending score. For the stance detection, the team used an LSTM-based neural network with one hidden layer that was pre-trained on the provided stance dataset.

Team *Puss in Boots* was our baseline retrieval model that used a BM25 implementation in Pyserini [49] 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.

Team *Grimjack* submitted 5 runs using query expansion and query reformulation to increase recall followed by a re-ranking step to improve precision and balance the stance distribution. For the first result they simply retrieved 100 passages ranked with the query likelihood with Dirichlet smoothing ($\mu = 1000$) using the original, unmodified queries (topic titles). Another approach re-ranks the top-10 of the initially retrieved passages using (1) argumentative axioms [14, 8] that are based on premises and claims in documents that were identified us-

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

ing TARGER [19], (2) newly proposed comparative axioms that "prefer" more comparative objects or earlier occurrence of comparative objects premises and claims, and (3) argument quality axioms that rank higher documents with a higher argument quality score; the quality scores were calculated using the IBM Project Debater API [7]. Next result ranking is based on the previous one, where the document positions are changed based on the predicted stance such as the 'pro first object' document is followed by the 'pro send object' followed by 'neutral' stance; the steps are then repeated. The document stance was predicted using the IBM Project Debater API [7]. The last two results used T0++ [76] to expand queries, e.g., by combining original queries with newly generated, where T0++ received topic descriptions as input, to assess the argument quality, and to detect the stance in zero-shot settings. The runs differed in whether the re-ranking that balanced the stance classes distribution was used.

Team *Asuna* proposed a three-step approach that consisted of preprocessing, search, and re-ranking. For each document (text passage) in the retrieval corpus the following 3 components were computed: one-sentence extractive summary using LexRank [28], premises and claims were identified with TARGER [19], and spam scores were found in the Waterloo Spam Rankings dataset [21].[23] Initial retrieval of top-40 documents was performed with a Pyserini [49] implementation of BM25F with default parameters ($k_1 = 0.9$ and $b = 0.4$). Queries (topic titles) were lemmatized and stopwords were removed using NLTK and extended with the most frequent terms coming from the topics modeled using LDA [10] generated for the initially retrieved documents. The extended queries were used to again retrieve top-40 passages with BM25F. Finally, team Asuna re-ranked the initially retrieved documents with the Random Forests classifier [37] fed with the following features: BM25F score, number of times the document was retrieved for different queries (original, three LDA topics from documents, and one LDA topic from the task topics' descriptions), number of tokens in documents, number of sentences in documents, number of premises in documents, number of claims in documents, spam-scores, predicted argument quality scores, and predicted stances. 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 [33]. The stance was also predicted using DistilBERT fine-tuned on the provided stance dataset [11].

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. To this end, the goal of the third task was to retrieve images that indicate an agreement or disagreement to some stance on a given topic as two separate lists similar to textual argument search.

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

6.1 Task Definition

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.

6.2 Data Description

Topics. Task 3 employs 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[24] and manual checks. The dataset contains various resources for each image, including the associated page for which it was retrieved as HTML page and as detailed web archive,[25] and information on how the image was ranked by Google. The full dataset is 368 GB large. To kickstart machine learning approaches, we provided 334 relevance judgments from [44].

6.3 Participant Approaches

In total, 3 teams submitted 12 runs to this task. The teams pursued quite different approaches. However, all participants employed OCR (Tesseract[26]) 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). Team Boromir used the ranking information for internal evaluation.

6.4 Task Evaluation

We employed crowdsourcing on Amazon Mechanical Turk[27] to evaluate the topical relevance, argumentativeness, and stance of the 6,607 images that the approaches retrieved, 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 employed MACE [38] to identify images with high disagreement (confidence ≤ 0.55) and re-judged them ourselves (2,056 images).

[24] https://www.phash.org/
[25] Archived using https://github.com/webis-de/scriptor
[26] https://github.com/tesseract-ocr/tesseract
[27] https://www.mturk.com

Table 5. Results for 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.

		Precision@10		
Team	**Run**	**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

6.5 Task Results

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 same run performed best across all three measures. A more comprehensive discussion including all teams' approaches is covered in the extended lab overview [15].

We provided one tough baseline for comparison, called *Minsc*, which ranks images according to the ranking from our original Google queries that included the filter words "good" (for pro) and "anti" (for con). Indeed, only team Boromir was able to beat this tough baseline. Remarkably, they did so especially for on-topic relevance, which is the closest to classical information retrieval.

Team *Aramis* focused on image features. They tested the use of hand-crafted formula vs. fully-connected neural network classifiers for both argumentativeness and stance detection. Features were based on OCR, image color, image category (graphic vs. photo; diagram-likeness), and query–text similarity. In our evaluation, the hand-crafted formula performed better that the neural approaches, maybe due to differences in the annotation procedure of the training set. However, the performance drop was not large, with their worst runs still achieving a Precision@10 of 0.664 (-0.037), 0.609 (-0.025), and 0.344 (-0.037).

Team *Boromir* indexed both image text (boosted 5-fold) and web page text, using stopword lists, min-frequency filtering, and lemmatization. They clustered images and manually assigned retrieval boosts per cluster to favor more argumentative images, especially diagrams. They employed textual sentiment detection for stance detection, using either a dictionary (AFINN) or a BERT classifier. Their approach performed best and convincingly improved over the baseline. In our evaluation, the BERT classifier improves over the dictionary and the image clustering had negative effects, as it seems 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 (-0.056), 0.728 (-0.040), and 0.411 (-0.014).

Team *Jester* focused on emotion-based image retrieval per facial image recognition,[28] image text, and the associated web page's text that is close to the image in the HTML source code. They assign positive leaning images to the pro-stance

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

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 (-0.025), 0.618 (-0.029), and 0.336 (-0.014). Thus emotion features seem helpful but insufficient when taken alone.

7 Conclusion

In this paper, we report on the third year of the Touché lab at CLEF 2022 and its three shared tasks: (1) argument retrieval for controversial questions, (2) argument retrieval for comparative questions, and (3) image retrieval for arguments. In the third Touché edition, the units of retrieval were different to the previous editions, including relevant argumentative sentences, passages, and images as well as their stance detection (our previous tasks focused on the retrieval of entire documents). From 58 registered teams, 23 participated in the tasks and submitted at least one valid run. Along with various query processing, query reformulation and expansion methods, and sparse retrieval models, the approaches had an increased focus on Transformer models and diverse re-ranking techniques. Not only the quality of documents and arguments was estimated, but also the predicted stance was considered for creating a final ranking. All evaluation resources developed at Touché are shared freely, including search queries (topics), the assembled manual relevance and argument quality judgments (qrels), and the ranked result lists submitted by the participants (runs). A comprehensive survey of developed approaches is included in the extended lab overview [15].

We plan to continue our activities for establishing a collaborative platform for researchers in the area of argument retrieval by providing submission and evaluation tools as well as by organizing collaborative events such as workshops, fostering the accumulation of knowledge and the development of new approaches in the field. For the next iteration of the Touché lab, we plan to expand current test collections with manual judgments, to extend evaluation with other argument quality dimensions and deeper document pooling.

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" (grant WA 4591/3-1), all part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999), and the German Ministry for Science and Education (BMBF) through the project "Shared Tasks as an Innovative Approach to Implement AI and Big Data-based Applications within Universities (SharKI)" (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.

Bibliography

[1] Aigrain, P., Zhang, H., Petkovic, D.: Content-based representation and retrieval of visual media: A state-of-the-art review. Multimedia Tools and Applications **3**(3), 179–202 (1996), https://doi.org/10.1007/BF00393937

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

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

[4] Alhamzeh, A., Bouhaouel, M., Egyed-Zsigmond, E., Mitrovic, J.: Distilbert-based argumentation retrieval for answering comparative questions. In: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 2936, pp. 2319–2330, CEUR-WS.org (2021), URL http://ceur-ws.org/Vol-2936/paper-209.pdf

[5] Alshomary, M., Düsterhus, N., Wachsmuth, H.: Extractive snippet generation for arguments. In: Proceedings of the 43nd International ACM Conference on Research and Development in Information Retrieval, SIGIR 2020, pp. 1969–1972, ACM (2020), URL https://doi.org/10.1145/3397271.3401186

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

[7] Bar-Haim, R., Kantor, Y., Venezian, E., Katz, Y., Slonim, N.: 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, pp. 267–274, Association for Computational Linguistics (2021), URL https://doi.org/10.18653/v1/2021.emnlp-demo.31

[8] Bevendorff, J., Bondarenko, A., Fröbe, M., Günther, S., Völske, M., Stein, B., Hagen, M.: Webis at TREC 2020: Health Misinformation track. In: Voorhees, E., Ellis, A. (eds.) Proceedings of the 29th International Text Retrieval Conference, TREC 2020, NIST (Nov 2020)

[9] Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly (2009), ISBN 978-0-596-51649-9, URL http://www.oreilly.de/catalog/9780596516499/index.html

[10] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003), URL http://jmlr.org/papers/v3/blei03a.html
[11] Bondarenko, A., Ajjour, Y., Dittmar, V., Homann, N., Braslavski, P., Hagen, M.: Towards understanding and answering comparative questions. In: Proceedings of the 15th ACM International Conference on Web Search and Data Mining, WSDM 2022, ACM (2022), https://doi.org/10.1145/3488560.3498534
[12] Bondarenko, A., Braslavski, P., Völske, M., Aly, R., Fröbe, M., Panchenko, A., Biemann, C., Stein, B., Hagen, M.: Comparative web search questions. In: Proceedings of the 13th ACM International Conference on Web Search and Data Mining, WSDM 2020, pp. 52–60, ACM (2020), URL https://dl.acm.org/doi/abs/10.1145/3336191.3371848
[13] Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2020: Argument retrieval. In: Working Notes Papers of the CLEF 2020 Evaluation Labs, CEUR Workshop Proceedings, vol. 2696 (2020), URL http://ceur-ws.org/Vol-2696/
[14] Bondarenko, A., Fröbe, M., Kasturia, V., Völske, M., Stein, B., Hagen, M.: Webis at TREC 2019: Decision track. In: Voorhees, E., Ellis, A. (eds.) Proceedings of the 28th International Text Retrieval Conference, TREC 2019, NIST (Nov 2019)
[15] Bondarenko, A., Fröbe, M., Kiesel, J., Syed, S., Gurcke, T., Beloucif, M., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2022: Argument retrieval. In: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org (2022)
[16] Bondarenko, A., Gienapp, L., Fröbe, M., Beloucif, M., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2021: Argument retrieval. In: Working Notes Papers of the CLEF 2021 Evaluation Labs, CEUR Workshop Proceedings, vol. 2936 (2021), URL http://ceur-ws.org/Vol-2936/
[17] Chang, N., Fu, K.: Query-by-pictorial-example. IEEE Transactions on Software Engineering **6**(6), 519–524 (1980), https://doi.org/10.1109/TSE.1980.230801
[18] Chekalina, V., Bondarenko, A., Biemann, C., Beloucif, M., Logacheva, V., Panchenko, A.: 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, pp. 302–311, Association for Computational Linguistics (2021), URL https://www.aclweb.org/anthology/2021.eacl-demos.36/
[19] Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., Panchenko, A.: TARGER: Neural argument mining at your fingertips. In: Proceedings of the 57th Annual Meeting of the Association

for Computational Linguistics, ACL 2019, pp. 195–200, ACL (2019), URL https://doi.org/10.18653/v1/p19-3031
[20] Cormack, G.V., Clarke, C.L.A., Büttcher, S.: 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, pp. 758–759, ACM (2009), https://doi.org/10.1145/1571941.1572114, URL https://doi.org/10.1145/1571941.1572114
[21] Cormack, G.V., Smucker, M.D., Clarke, C.L.A.: Efficient and effective spam filtering and re-ranking for large web datasets. Inf. Retr. **14**(5), 441–465 (2011), https://doi.org/10.1007/s10791-011-9162-z, URL https://doi.org/10.1007/s10791-011-9162-z
[22] Devlin, J., Chang, M., Lee, K., Toutanova, K.: 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, pp. 4171–4186, ACL (2019), URL https://doi.org/10.18653/v1/n19-1423
[23] Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., Da San Martino, G.: SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In: 15th International Workshop on Semantic Evaluation (SemEval'2021), pp. 70–98, Association for Computational Linguistics, Online (Aug 2021), https://doi.org/10.18653/v1/2021.semeval-1.7, URL https://aclanthology.org/2021.semeval-1.7
[24] Dove, I.J.: On images as evidence and arguments. In: van Eemeren, F.H., Garssen, B. (eds.) Topical Themes in Argumentation Theory: Twenty Exploratory Studies, pp. 223–238, Argumentation Library, Springer Netherlands, Dordrecht (2012), https://doi.org/10.1007/978-94-007-4041-9_15
[25] Dumani, L., Neumann, P.J., Schenkel, R.: A framework for argument retrieval - ranking argument clusters by frequency and specificity. In: Proceedings of the 42nd European Conference on IR Research (ECIR 2020), Lecture Notes in Computer Science, vol. 12035, pp. 431–445, Springer (2020), URL https://doi.org/10.1007/978-3-030-45439-5_29
[26] Dumani, L., Schenkel, R.: Quality aware ranking of arguments. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 335–344, CIKM '20, Association for Computing Machinery (2020), URL https://doi.org/10.1007/978-3-030-45439-5_29
[27] Dunaway, F.: Images, emotions, politics. Modern American History **1**(3), 369–376 (Nov 2018), ISSN 2515-0456, 2397-1851, https://doi.org/10.1017/mah.2018.17
[28] Erkan, G., Radev, D.R.: LexRank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004), https://doi.org/10.1613/jair.1523, URL https://doi.org/10.1613/jair.1523

[29] Fröbe, M., Bevendorff, J., Gienapp, L., Völske, M., Stein, B., Potthast, M., Hagen, M.: 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, pp. 2398–2404, ACM (2021), URL https://dl.acm.org/doi/10.1145/3404835.3463246
[30] Fröbe, M., Bevendorff, J., Reimer, J., Potthast, M., Hagen, M.: 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), URL https://dl.acm.org/doi/10.1145/3397271.3401212
[31] Fröbe, M., Bittner, J., Potthast, M., Hagen, M.: 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), Lecture Notes in Computer Science, vol. 12036, pp. 12–19, Springer (2020), URL https://link.springer.com/chapter/10.1007%2F978-3-030-45442-5_2
[32] Gienapp, L., Stein, B., Hagen, M., Potthast, M.: Efficient pairwise annotation of argument quality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5772–5781, Association for Computational Linguistics, Online (Jul 2020), https://doi.org/10.18653/v1/2020.acl-main.511, URL https://aclanthology.org/2020.acl-main.511
[33] Gienapp, L., Stein, B., Hagen, M., Potthast, M.: Efficient pairwise annotation of argument quality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 5772–5781, Association for Computational Linguistics (2020), URL https://www.aclweb.org/anthology/2020.acl-main.511/
[34] Google: Google images best practices. Google Developers (2021), URL https://support.google.com/webmasters/answer/114016
[35] Grancea, I.: Types of visual arguments. Argumentum. Journal of the Seminar of Discursive Logic, Argumentation Theory and Rhetoric **15**(2), 16–34 (2017)
[36] Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., Slonim, N.: 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, pp. 7805–7813, AAAI Press (2020), URL https://ojs.aaai.org/index.php/AAAI/article/view/6285
[37] Ho, T.K.: Random decision forests. In: Third International Conference on Document Analysis and Recognition, ICDAR 1995, August 14 - 15, 1995, Montreal, Canada. Volume I, pp. 278–282, IEEE Computer Society (1995), URL https://doi.org/10.1109/ICDAR.1995.598994

[38] Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., Hovy, E.: 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), pp. 1120–1130, Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), URL https://aclanthology.org/N13-1132

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

[40] Jindal, N., Liu, B.: 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, pp. 1331–1336, AAAI Press (2006), URL http://www.aaai.org/Library/AAAI/2006/aaai06-209.php

[41] Kaszkiel, M., Zobel, J.: Passage Retrieval Revisited. In: Belkin, N.J., Narasimhalu, A.D., Willett, P., Hersh, W.R., Can, F., Voorhees, E.M. (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, pp. 178–185, ACM (1997), https://doi.org/10.1145/258525.258561, URL https://doi.org/10.1145/258525.258561

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

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

[44] Kiesel, J., Reichenbach, N., Stein, B., Potthast, M.: Image retrieval for arguments using stance-aware query expansion. In: Proceedings of the 8th Workshop on Argument Mining, ArgMining 2021 at EMNLP, pp. 36–45, ACL (2021)

[45] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: 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

[46] Latif, A., Rasheed, A., Sajid, U., Ahmed, J., Ali, N., Ratyal, N.I., Zafar, B., Dar, S.H., Sajid, M., Khalil, T.: Content-based image retrieval and

feature extraction: A comprehensive review. Mathematical Problems in Engineering **2019**, 21 (2019), https://doi.org/10.1155/2019/9658350
[47] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning, pp. 1188–1196, PMLR (2014)
[48] Levy, R., Bogin, B., Gretz, S., Aharonov, R., Slonim, N.: Towards an argumentative content search engine using weak supervision. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, pp. 2066–2081, Association for Computational Linguistics (2018), URL https://www.aclweb.org/anthology/C18-1176/
[49] Lin, J., Ma, X., Lin, S., Yang, J., Pradeep, R., Nogueira, R.: 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, pp. 2356–2362, ACM (2021), https://doi.org/10.1145/3404835.3463238, URL https://doi.org/10.1145/3404835.3463238
[50] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), URL http://arxiv.org/abs/1907.11692
[51] Lovins, J.B.: Development of a stemming algorithm. Mech. Transl. Comput. Linguistics **11**(1-2), 22–31 (1968), URL http://www.mt-archive.info/MT-1968-Lovins.pdf
[52] Ma, N., Mazumder, S., Wang, H., Liu, B.: 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, pp. 5782–5788, Association for Computational Linguistics (2020), URL https://www.aclweb.org/anthology/2020.acl-main.512/
[53] Mikolov, T., Chen, K., Corrado, G., Dean, J.: 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
[54] Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM **38**(11), 39–41 (1995)
[55] Nadamoto, A., Tanaka, K.: A comparative web browser (CWB) for browsing and comparing web pages. In: Proceedings of the 12th International World Wide Web Conference, WWW 2003, pp. 727–735, ACM (2003), URL https://doi.org/10.1145/775152.775254
[56] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: 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, CEUR Workshop Proceedings, vol. 1773, CEUR-WS.org (2016), URL http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[57] Nogueira, R., Lin, J., Epistemic, A.: From doc2query to doctttttquery. Online preprint (2019), URL https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery-v2.pdf
[58] Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (1999), URL http://ilpubs.stanford.edu:8090/422/
[59] Palotti, J.R.M., Scells, H., Zuccon, G.: 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, pp. 1325–1328, ACM (2019), URL https://doi.org/10.1145/3331184.3331399
[60] Panchenko, A., Bondarenko, A., Franzek, M., Hagen, M., Biemann, C.: Categorizing comparative sentences. In: Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, pp. 136–145, Association for Computational Linguistics (2019), URL https://doi.org/10.18653/v1/w19-4516
[61] Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980), https://doi.org/10.1108/eb046814, URL https://doi.org/10.1108/eb046814
[62] Porter, M.F.: Snowball: A language for stemming algorithms (2001), URL http://snowball.tartarus.org/texts/introduction.html
[63] Potthast, M., Gienapp, L., Euchner, F., Heilenkötter, N., Weidmann, N., Wachsmuth, H., Stein, B., Hagen, M.: Argument search: Assessing argument relevance. In: Proceedings of the 42nd International Conference on Research and Development in Information Retrieval, SIGIR 2019, pp. 1117–1120, ACM (2019), URL https://doi.org/10.1145/3331184.3331327
[64] Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA integrated research architecture. In: Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, The Information Retrieval Series, vol. 41, pp. 123–160, Springer (2019), URL https://doi.org/10.1007/978-3-030-22948-1_5
[65] Pradeep, R., Nogueira, R., Lin, J.: 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
[66] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
[67] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020), URL http://jmlr.org/papers/v21/20-074.html
[68] Reimers, N., Gurevych, I.: 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, pp. 3980–3990, Association for Computational

Linguistics (2019), https://doi.org/10.18653/v1/D19-1410, URL https://doi.org/10.18653/v1/D19-1410

[69] Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., Gurevych, I.: Classification and clustering of arguments with contextualized word embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 567–578, Association for Computational Linguistics, Florence, Italy (Jul 2019), https://doi.org/10.18653/v1/P19-1054, URL https://aclanthology.org/P19-1054

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

[71] Robertson, S.E., Zaragoza, H., Taylor, M.J.: Simple BM25 extension to multiple weighted fields. In: Proceedings of the 13th International Conference on Information and Knowledge Management, CIKM 2004, pp. 42–49, ACM (2004), URL https://doi.org/10.1145/1031171.1031181

[72] Rocchio, J.: Relevance feedback in information retrieval. The Smart retrieval system-experiments in automatic document processing pp. 313–323 (1971)

[73] Roque, G.: Visual argumentation: A further reappraisal. In: van Eemeren, F.H., Garssen, B. (eds.) Topical Themes in Argumentation Theory, vol. 22, pp. 273–288, Springer Netherlands, Dordrecht (2012), ISBN 978-94-007-4040-2, https://doi.org/10.1007/978-94-007-4041-9_18, URL http://link.springer.com/10.1007/978-94-007-4041-9_18, series Title: Argumentation Library

[74] Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text mining: applications and theory **1**(1-20), 10–1002 (2010)

[75] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR **abs/1910.01108** (2019), URL http://arxiv.org/abs/1910.01108

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

[77] Schildwächter, M., Bondarenko, A., Zenker, J., Hagen, M., Biemann, C., Panchenko, A.: Answering comparative questions: Better than ten-blue-links? In: Proceedings of the 2019 Conference on Human

Information Interaction and Retrieval, CHIIR 2019, pp. 361–365, ACM (2019), URL https://doi.org/10.1145/3295750.3298916
[78] Solli, M., Lenz, R.: Color emotions for multi-colored images. Color Research & Application **36**(3), 210–221 (2011), https://doi.org/10.1002/col.20604
[79] Stab, C., Daxenberger, J., Stahlhut, C., Miller, T., Schiller, B., Tauchmann, C., Eger, S., Gurevych, I.: 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, pp. 21–25, Association for Computational Linguistics (2018), URL https://www.aclweb.org/anthology/N18-5005
[80] Sun, J., Wang, X., Shen, D., Zeng, H., Chen, Z.: CWS: A comparative web search system. In: Proceedings of the 15th International Conference on World Wide Web, WWW 2006, pp. 467–476, ACM (2006), URL https://doi.org/10.1145/1135777.1135846
[81] Wachsmuth, H., Naderi, N., Habernal, I., Hou, Y., Hirst, G., Gurevych, I., Stein, B.: Argumentation quality assessment: Theory vs. practice. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, pp. 250–255, Association for Computational Linguistics (2017), URL https://doi.org/10.18653/v1/P17-2039
[82] Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: Computational argumentation quality assessment in natural language. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, pp. 176–187 (2017), URL http://aclweb.org/anthology/E17-1017
[83] Wachsmuth, H., Potthast, M., Al-Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., Stein, B.: Building an argument search engine for the web. In: Proceedings of the Fourth Workshop on Argument Mining (ArgMining), pp. 49–59, Association for Computational Linguistics (2017), URL https://doi.org/10.18653/v1/w17-5106
[84] Wachsmuth, H., Stein, B., Ajjour, Y.: "PageRank" for argument relevance. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, pp. 1117–1127, Association for Computational Linguistics (2017), URL https://doi.org/10.18653/v1/e17-1105
[85] Wachsmuth, H., Syed, S., Stein, B.: Retrieval of the best counterargument without prior topic knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, pp. 241–251, Association for Computational Linguistics (2018), URL https://www.aclweb.org/anthology/P18-1023/

[86] Wang, W., He, Q.: A survey on emotional semantic image retrieval. In: International Conference on Image Processing (ICIP 2008), pp. 117–120, IEEE (2008), https://doi.org/10.1109/ICIP.2008.4711705
[87] Wu, A.: Learn more about what you see on google images. Google Blog (2020), URL https://support.google.com/webmasters/answer/114016
[88] Yanai, K.: Image collector: An image-gathering system from the world-wide web employing keyword-based search engines. In: International Conference on Multimedia and Expo, (ICME 2001), IEEE (2001), https://doi.org/10.1109/ICME.2001.1237772