Touché 2022 Best of Labs: Neural Image Retrieval for Argumentation

Tobias Schreieder(✉) and Jan Braker

Leipzig University, 04109 Leipzig, Germany
{fp83rusi,jb64vyso}@studserv.uni-leipzig.de

Abstract. Given a text query on a controversial topic, the task of Image Retrieval for Argumentation is to rank images according to how well they can be used to support a discussion on the topic. This paper provides a detailed investigation of the challenges of this task by means of a novel and modular retrieval pipeline. All findings relate to our work from last year's CLEF Touché'22 lab and a reproducibility study based on it. There, we demonstrate the unified retrieval pipeline NeurArgs and provide improved stance models. This work presents the approaches of the two papers, regarding the problems identified and the solutions provided. Herewith, we achieve an effectiveness improvement in argumentative image detection of up to 0.832 precision@10. However, despite this success, our study also revealed a previously unknown negative result: when it comes to stance detection, none of the tested stance models can convincingly beat a random baseline. Therefore, we conduct a thorough error analysis to understand the inherent challenges of image stance detection and provide insight into potential new approaches to this task.

Keywords: Argumentation · Image retrieval · Image stance detection

1 Introduction

Several years ago, social media discussions changed from being mainly text-focused towards including more images or videos. Specific platforms that focus on images, like Instagram, then became increasingly popular and are still today. In discussions on social media, people thus also often include images to illustrate their stance and arguments on the topic in question, or to support written arguments. Whether images can be "argumentative," i.e., whether they can represent arguments in their own right, is controversial [5]. However, their usefulness for argumentation is obvious: Kjeldsen [7] notes that images can underpin and support arguments, clarify facts, and convey facts more effectively than words.

Although retrieval systems for textual arguments exist [19], none exist yet specifically for image retrieval for argumentation. A search engine dedicated to the retrieval of images that are relevant to controversial topics can be useful for finding images to support one's stance on social media or elsewhere, and to get a "visual" overview of the landscape of opinions at-a-glance for personal deliberation. The first shared task to present pioneering approaches to this research

A. Arampatzis et al. (Eds.): CLEF 2023, LNCS 14163, pp. 186–197, 2023.
https://doi.org/10.1007/978-3-031-42448-9_16

question was the CLEF Touché lab "Image Retrieval for Arguments" in 2022. There, three different teams presented different approaches to solving the task, which were evaluated independently and uniformly with Tira. This software tries to solve the problem of scientific reproducibility, especially for shared tasks [12].

To pave the way for more effective image retrieval systems for arguments, in this paper we first briefly present our approaches from the CLEF Touché'22 lab, which are referred to subsequently as Aramis [2]. Based on this, weaknesses of the approaches are made visible and the improved retrieval pipeline NeurArgs, as well as other compatible stance models, which we introduced with Carnot et al. [4], are discussed. Inspired by the three-stage evaluation of image retrieval for arguments proposed by Kiesel et al. [6], we propose NeurArgs, a modular retrieval system with three AI models to unify approaches: a topic model to identify images relevant to a query, an argument model to identify images suitable for argumentation, and a stance model to sort images into pros and cons. By employing the system to combine the approaches submitted to the CLEF Touché'22 lab, we improve over the lab's best score by 0.064 in the lab's precision metric, reaching a score of 0.832. However, none of the 8 stance models we evaluated convincingly improves stance detection over a random baseline.

The paper is structured as follows: Sect. 2 reviews related work. Section 3 provides a brief overview of the Touché22-Image-Retrieval-for-Arguments dataset, and Sect. 4 shows the development of the Aramis models published in 2022 to NeurArgs and details the different models that we employ in our analyses. Section 5 presents the results of the reproduced and newly developed models (all code linked there), which successfully reproduces the state-of-the-art but also unveils our main negative result in the comparison with naive baselines. Section 6 then provides a qualitative analysis of the challenges for stance detection to aid researchers in overcoming our current result.

2 Related Work

Several former works exist on argument retrieval from text collections. The first systems were *args.me* [19], *ArgumenText* [18], and *IBM debater* [9]. For their evaluation, Potthast et al. [11] suggest employing the retrieved arguments' query relevance as well as rhetorical, logical, and dialectical quality in Cranfield style experiments. However, more detailed aspects of argument quality are discussed in the literature [19] and could be used to evaluate argument search engines.

Approaches for image retrieval have been explored for many years, mostly in content-based image retrieval. There, the query is itself an image and relevant results are similar to other images. Therefore, the content of the images needs to be analyzed. Smeulders et al. [15] provide an overview of the conducted research in the field in the early years. One of the important early projects regarding content-based image retrieval was presented by Rui et al. [13]. They used image feature vectors to establish a connection between images and terms. The works of Latif et al. [8] and Meharban and Priya [10] give a more recent overview of approaches and features for web image search. For example, Shao et al. [14]

propose to reduce the number of colors of images to a few representative ones in order to search more effectively for images containing a certain color-base. Color features seem to be highly promising when retrieving images for arguments due to colors evoking specific emotions [17]. A relatively new approach in image retrieval is to employ optical character recognition software like Tesseract [16] to extract the text from the images and then to extract standard features from the text for indexing. This approach seems especially promising for meme images and other images containing written arguments.

The retrieval of images for arguments has been sparsely explored so far. The pioneering work by Kiesel et al. [6] attempted this task by simply extending the search query with different terms to get different results for each stance. In their most effective approach, the query was either extended with the word "good" (for the pro stance) or the word "anti" (for the con stance). This method achieved good results overall, but was not able to improve upon a random classifier with regard to stance detection. The same authors then organized a shared task at the CLEF Touché'22 lab [1]. We employ the lab's data and the most effective participating approaches in our system comparison (Sects. 3 and 5).

3 The Touché'22 Dataset

For our research into image retrieval for argumentation, we employ the dataset of the corresponding Touché'22 shared task [1], which was located at the CLEF 2022 conference. The data is freely accessible online.[1] The dataset contains 23,841 images for 50 controversial topics. The topics include, for example, "is golf a sport?" or "should education be free?" The images were crawled using regular image search engine queries related to the 50 topics. In addition to the image itself, the dataset contains, for example, a screenshot of the web page it appeared on, the text from that web page, or the image's rank in the regular search engine's result list. For our analysis (Sect. 5) we employ the queries, the image pixel values and recognized text, and the corresponding web page's HTML source code. The dataset also contains three relevance ratings (on-topic, pro, con) for each of the 6607 images that the participants retrieved for the 2022 lab. The images shown in this paper, except the schematic of our modular system in Fig. 1, are taken from this dataset. The Aramis group evaluated 20 topics, resulting in 9559 evaluated images. A comparison between the two evaluations doesn't show much of a difference, so we decided to use the ratings from the Touché'22 lab.

4 Development of the NeurArgs Approach

Inspired by the three-stage evaluation of image retrieval for arguments by Kiesel et al. [6], we propose the retrieval system NeurArgs with three AI models. Each

[1] https://touche.webis.de/data.html#touche22-image-retrieval-for-arguments.

stage has its own model, which is illustrated in Fig. 1. An image is considered relevant if it fits the topic (topic-relevance), provides a statement on some topic (argumentativeness) and fits a previously specified stance (pro/support or con/attack) on the topic (stance-relevance). In this paper, we want to showcase the development from the Aramis approaches to the new NeurArgs framework.

Using a modular architecture allows us to investigate each stage separately. The previous work shows, that stance detection is the most challenging subtask for now. Therefore, we compare different stance models based on the unified topic and argument model of NeurArgs. The following sections first introduce both the models in general and the specific models used in our analysis (Sect. 5). Table 1 provides an overview of the features each model employs.

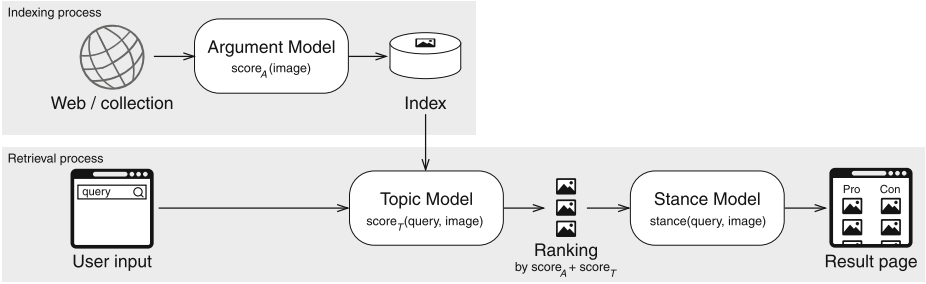


Fig. 1. Schematic of NeurArgs: Images from the web or a collection are, together with the web pages they appeared on, scored by the argument model for argumentativeness (score$_A$) and indexed. In the retrieval process, the user issues a query, which is used to score the images for topicality (score$_T$), rank images by the sum of the two scores, and classify their stance to sort them into two result lists (Pro vs. Con) for display.

4.1 The Topic Model

The topic model of NeurArgs ranks images by their relevance to the user's query by assigning a score to each image in the index (cf. Fig. 1). As the score depends on the query, the topic model must be part of the retrieval process. As a first naive approach, Aramis proposed a DirichletLM model based on the HTML page where the image is located and the preprocessed query as input [2]. The Touché'22 evaluation showed that this approach achieved poorer topic-relevance precision@10 compared to the baseline [1]. Therefore, we adapt the Elasticsearch BM25 retrieval from the best-performing system of Boromir [3].

The NeurArgs topic model combines our introduced pipeline and extends them in several places. Specifically, we employ textual matching of the query, text from the image's context (web page) and text from the image itself. The query and the recognized text on the image are preprocessed using standard stopword and punctuation removal, and lowercasing. The text from the HTML source code of the image's web page gets extracted and is also being preprocessed.

Table 1. Input features employed by the respective models detailed in Sect. 4: search query (topic), image pixels, recognized text (via OCR), and HTML source code of the web page on which the image was originally found.

Model	Query	Image features		
	Text	Image file		HTML
		Pixels	Text	
Topic model	✓		✓	✓
Argument model		✓	✓	
Stance models				
Oracle				
NeurArgs baseline		✓	✓	
Random baseline				
Aramis Formula	✓	✓	✓	✓
Aramis Neural	✓	✓	✓	✓
Neural text+image 3class	✓	✓	✓	
Neural text+image 2x2class	✓	✓	✓	

The part of this text that can be found close to the image is indexed using Elasticsearch's BM25. Additionally, the recognized text on the image is used for retrieval boosting. As this topic model already considerably improved over the best approach in Touché'22 (cf. Sect. 5), we did not investigate further models but focused our attention on different stance models instead.

4.2 The Argument Model

Our NeurArgs argument model ranks images by their suitability for argumentation by assigning a score to each image in the index (cf. Fig. 1). Conceptually, an image that shows either critical or supportive attitudes should receive a high argument score. Unlike for the topic model, this score does not depend on the query. Therefore, the model's score for each image is indexed alongside the image, and directly used in the retrieval function. The argument model employs the query-independent features that are previously employed by the Aramis approach for the Touché'22 lab [2]. Furthermore, we employ the same neural network classifier for NeurArgs as Aramis for calculating the argumentativeness score from the features. We detail those features below for completeness.

The first set of features are color properties, with the intent to capture the overall mood of the image. We calculate the average and dominant color of the image as RGB values, as well as the area share of red, green, blue, and yellow. Other features used for the neural network are the image type (graphic or photography) and diagram-likeness. We adopt the simple common color heuristic of Aramis for image type classification, and the approach to diagram-likeness based on horizontal kernels [2]. Additionally, general text features are used: text

length, sentiment, the area percentage of the image occupied by text, and the position of the text in a 8×8 grid [2]. Here, the text position is used as a hint to identify memes and image quotes. The text is extracted using Tesseract OCR[2] after converting the image to gray scale and adjusting Tesseract's configuration for maximum text recognition. Afterward, only words that occur in a standard English dictionary are kept to improve detection precision.

4.3 The Stance Model

The stance model sorts the ranked images into pros and cons (cf. Fig. 1). To this end, stance models label each image for a topic as pro, con, both, or neither (cf. Kiesel et al. [6]). Only images labeled as pro or con are placed on the result page in the respective column in decreasing score order. Note that, according to the Touché task, an image can be both pro and con, in which case it is considered a relevant image if placed in either one or both result lists. As the score depends on the query, the stance model must be part of the retrieval process.

The Touché'22 results show that none of the participating models achieved a high precision for stance detection [1]. In our reproducibility study, we focus our investigations on the stance detection subtask and compare 14 approaches, including two baseline approaches and the oracle. In this paper, we compare the stance models of Aramis and selected further developments with the results of the best approach of Touché'22 (Boromir). Boromir used a sentiment detection BERT-model to classify the image based on the sentiment of the title of the image's original web page [3]. All other models are explained and evaluated in the corresponding paper of Carnot et al. [4]. In the following, we briefly discuss the modular stance models that classify the previously generated ranking results:

Oracle is a theoretic approach that uses the ground-truth stance labels and thus provides the upper limit. As the ground truth contains only stance labels for topic-relevant and argumentative images, the oracle's scores are the overall achievable maximum for our setting. However, as the dataset contains less than 10 images for some topic and stance combinations, this score is less than 100%.

NeurArgs baseline classifies each image as both pro and con, which results in an identical result list for each stance.

Random baseline classifies images as either pro or con with equal probability.

Aramis Formula uses the heuristic formula developed by team Aramis that is based on the same features used in the argument model [2]. Additionally, the query, the interrelation, and sentiments of the mentioned texts are used as features. The weights for each feature were set manually by Aramis.

Aramis Neural is a neural network, also developed by team Aramis [2], that uses the same features as Aramis Formula to classify images as either pro, neutral, or con. The neutral images are not further used in the results.

[2] https://github.com/tesseract-ocr/tesseract.

Neural text+image 3class employs a feedforward neural network classifier using the image resized to 256×256 pixels, the query text, and the recognized text of the images as input. The network combines a BERT model with a ResNet50V2 extended by some dropout layers to prevent overfitting. It has three output neurons that represent pro, neutral, and con.

Neural text+image 2x2class employs the same architecture as the neural text+image 3class approach, but with a single output neuron. The architecture is trained twice, once for pro and once for con images. Both are entirely independent of each other. The network calculates a score for the entry which shows if the image fits the stance. It needs to be above half of the highest score of the current query to be accepted in the respective category.

5 Evaluation of the NeurArgs Approach

Table 2 shows the results of our extended analysis. For consistency with the existing evaluation, we only use images where rating already exists and refrain from annotating images ourselves. Hence, the retrieved lists are condensed, a 5-fold cross-validation is used for evaluating the machine-learning-based approaches. The code for this study is available online.[3] Besides comparing more approaches, our evaluation also goes deeper than the original one of Bondarenko et al. [1] in that it shows results also for pro and con separately. The Touché'22 lab only used precision@10, arguing that this was closest to the setting of a user looking at a single page of result images. Additionally, we calculated NDCG@10 scores, which performed very similar to precision@10 and are therefore not separately shown. The exact values can be found in the work of Carnot et al. [4].

5.1 Topic and Argument Retrieval

We first detail the results for the retrieval of topic-relevant and argumentative images. This setup corresponds to omitting the stance model in Fig. 1 from the NeurArgs retrieval. The NeurArgs topic and argument models are used for all shown stance models. Since the assignment to the classes pro or con is based on the images with the highest score, the stance model can influence the topic relevance and argumentativeness scores. At this point, we also tested different weightings for the topic model's score_T and the argument model's score_A than the simple sum, but none lead to improvements for the different approaches.

As seen in Table 2, with a precision@10 of 92.6% for topic-relevance and 83.2% for argumentativeness, the NeurArgs baseline outperforms all methods from the CLEF Touché'22 lab. For reference, the most effective method from the lab, developed by Boromir, only achieved a topic-relevance precision score of 87.8% (-4.8%) and an argumentativeness score of 76.8% (-6.4%). Note that the baseline uses the same images for both stances and thus always retrieves only 10 images total, whereas other approaches might retrieve up to 20 images. However, most

[3] https://github.com/webis-de/SIGIR-23.

Table 2. The table shows the precision@10 scores on condensed lists for all 50 topics, sorted by stance-relevance (both) for all stance detection models. For this purpose, topic-relevance, argumentativeness and stance-relevance are always evaluated in relation to the overall system for the 20 images retrieved (10 pro and 10 con). The "both" scores are the averages for the 10 pro and 10 con images. In each case, the best results were highlighted in bold. All stance models follow the NeurArgs topic model and the argument model as described in Sect. 4, except for Best of Touché'22 and the Oracle.

Stance Model	Precision@10								
	Topic-relevance			Argumentativeness			Stance-relevance		
	Pro	Con	Both	Pro	Con	Both	Pro	Con	Both
Oracle	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.802	0.901
Neural text+image 2x2class	0.924	0.822	0.873	0.830	0.766	0.798	0.660	**0.310**	**0.485**
Aramis Formula	0.920	0.814	0.867	**0.838**	0.742	0.790	**0.690**	0.216	0.453
NeurArgs baseline	**0.926**	**0.926**	**0.926**	0.832	**0.832**	**0.832**	0.662	0.232	0.447
Neural text+image 3class	0.924	0.866	0.895	0.830	0.800	0.815	0.660	0.226	0.443
Random baseline	0.894	0.888	0.891	0.816	0.812	0.814	0.664	0.222	0.443
Aramis Neural	0.694	0.676	0.685	0.668	0.640	0.654	0.588	0.278	0.433
Best of Touché'22 (Boromir)	0.884	0.872	0.878	0.782	0.754	0.768	0.594	0.256	0.425

other models achieve almost the same performance as the NeurArgs baseline, with only slight losses in terms of topic relevance and argumentativeness. Moreover, Table 2 shows that the scores for topic-relevance and argumentativeness are very similar between images retrieved for pros and cons, with only a few exceptions like for Neural text+image 2x2class. Thus, the images retrieved for both pro and con are equally argumentative for most approaches.

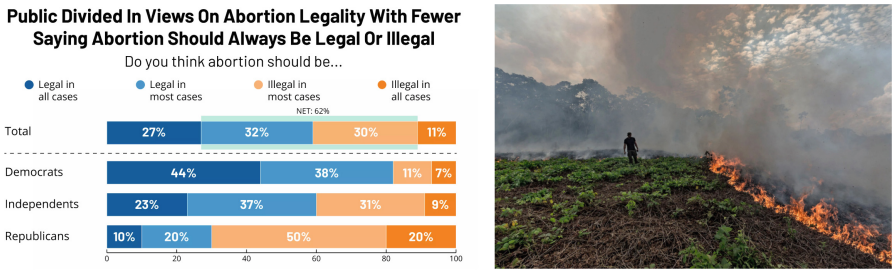
5.2 Comparison of Stance Detection Models

Table 2 shows that stance detection is a challenge in image retrieval for argumentation. The best result that possibly could have been achieved for stance-relevance precision@10 lies at 90.1% as shown by the oracle. This is because not every topic has ten images on each side in the evaluation data, and missing images will be treated in the same way as an incorrect image. We find that the neural text+image 2x2class model that uses the image and associated text as input is the most effective, with a precision@10 of 48.5%. In this comparison, however, the NeurArgs baseline, which outputs an identical list of images for pro and con, comes in third with 44.7% (-3.8%). On the pro side, other models achieved the highest precision results: the Aramis Formula model exceed 69.0%. In general, results on the pro side range from 58.8% to 69.0%. Unfortunately, none of the models were able to classify the majority of con images correctly. The precision range for the con side lies between 21.6% and 31.0%. The reason for this drop in precision on the con side can be found in the dataset. For a number of topics, there are not enough con images annotated to retrieve ten images, which makes it impossible to achieve high precision scores. Therefore, the best theoretically possible result is 80.2% (Oracle).

However, Table 2 also shows the main negative result of our study: none of the approaches can convincingly beat our baselines. With a stance-relevance (both) precision@10 of 44.3%, the random baseline is about half a percentage point below the NeurArgs baseline. When we conducted significance tests (Student's t-test with Bonferroni correction at p=0.05) to detect if our approaches improve significantly upon the baseline in terms of precision@10, we found that only the oracle improves over it significantly. Worse, a number of models, such as Aramis Neural and Best of Touché'22 (Boromir), were not able to outperform the random model or the NeurArgs baseline. Especially when considering that one of the baselines is purely random, we thus have to conclude that, so far, stance detection in image retrieval for argumentation is an unsolved problem.

6 Insights into Image Stance Detection

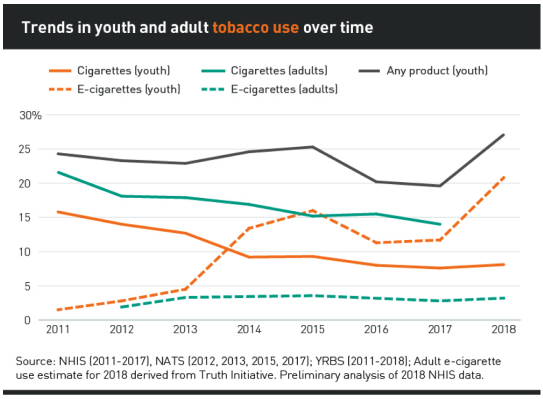
Although our analysis in Sect. 5 confirms the seemingly good results of the approaches from the Touché lab, our analysis also revealed that no approach can convincingly beat naive baselines such as random or both-sides classification in detecting the image stance. This negative result suggests that the analyzed approaches fail to account for key challenges of the stance detection task. To uncover these challenges, we performed a qualitative analysis of the images the approaches retrieved and misclassified. Specifically, we identified nine challenges:



(a) (b)

Fig. 2. (a) Different valuations cause stance ambiguity. The image could be pro "should abortion be legal?" if one supports the Democrats, but con if not. (b) Image understanding depends on background knowledge. The image could be pro "is human activity primarily responsible for global climate change?" depending on the viewer's expertise.

Different Valuations Cause Stance Ambiguity. Images or diagrams may contain several pieces of information that lead to different or opposite conclusions for different audiences. Specifically, a person's background, socialization, and opinions influence how they interpret the image stance. Figure 2a illustrates this by the political party affiliation. Someone who supports the Democrats sees in the chart that their favorite party is in favor of legal abortion (pro stance).



(a)

(b)

Fig. 3. (a) Neutral images. The image is neither clearly pro nor con "is vaping with e-cigarettes safe?". (b) Irony and Jokes. The image is con "do violent video games contribute to youth violence?" if one gets the joke about "pong" being a violent game.

Republicans, instead, might see the image as con. This problem is challenging for both algorithms and annotation campaigns. To solve it for algorithms, one could identify images with this problem and either not show them in the results or classify them based on a user-provided audience profile. For annotation campaigns, one could provide special training for annotators for such cases.

Image Understanding Depends on Background Knowledge. Some images require the viewer to have certain background knowledge to understand their stance. The image in Fig. 2b is pro "is human activity primarily responsible for global climate change?" for viewers who connect the burning of forests and the climate impact. Without that, the image is not even topic-relevant. This problem provides a challenge for algorithms and annotation campaigns. Analyzing the context of the image web page could provide hints on the relevant knowledge.

Unbalanced Image Stance Distribution. For some topics, there are much more pro images available than con images, or vice versa, which can result in biased stance detectors if one does not pay attention to such skewed data in the training process. For example, the dataset contains only very few con images for the topic "should bottled water be banned?". One solution is to balance the training dataset and remove topics with overly skewed distributions.

Neutral Images. Some images, like diagrams, contain thought-provoking impulses on a topic, but are not evidently pro or con. However, they can be visually very similar to arguments with a unique stance, which can be a problem. For example, the image in Fig. 3a is very informative without clearly being pro or con "is vaping with e-cigarettes safe?". Nevertheless, one can imagine visually very similar images that are clearly pro or con, which provides a challenge

in classifier training. To solve this, it might be necessary to develop a classifier to detect neutral images. Such approaches likely need semantic interpretations.

Irony and Jokes. Many images, especially memes, contain irony and jokes, which may not be understood by humans or algorithms. Figure 3b shows a meme that was retrieved for the topic "do violent video games contribute to youth violence?". The image is a joke on the idea that video games created violence, as if violence had not existed before. We expect irony detection for images to be very challenging. Still, it might be possible to transfer advances in textual irony detection (e.g., [20]) to visual irony detection.

Additional Problems. Besides the problems mentioned, there are additional problems such as regional images that are only relevant for people in certain regions. Further, it happens with several topics that both stances are found in one image, making a direct assignment difficult. This is exacerbated by images with more than two stances, which makes the choice of a binary classifier unsuitable. Another problem is understanding the semantics of diagrams by algorithms.

7 Conclusion

For the task of image retrieval for argumentation, we compared 8 approaches (including the previous state-of-the-art, two baselines, and the oracle) while emphasizing stance detection. To compare different approaches, we proposed the modular image retrieval system NeurArgs: a topic model to identify images relevant to a query, an argument model to identify images suitable for argumentation, and a stance model to sort images into pros and cons. The approaches shown in our paper employ features of the query, the image file or the web page an image was indexed on. The NeurArgs approach for the topic and argument model, which we have derived from the experience of the Touché'22 lab, provide a new state-of-the-art for the respective parts of the task, reaching 0.926 precision@10 for topic-relevance and 0.832 precision@10 for argumentativeness.

However, the extended analysis also uncovers a strong negative result: none of the analyzed approaches can convincingly beat a random baseline (or a both-sides baseline) when it comes to stance detection. We thus conclude that stance detection in image retrieval for argumentation is so far an unsolved problem. To pave the way for future approaches, we identified nine different challenges for stance detection and provided some examples and possible solutions.

References

1. Bondarenko, A., et al.: Overview of touché 2022: argument retrieval. In: Barrón-Cedeño, A., et al. (eds.) CLEF 2022. LNCS, vol. 13390, pp. 311–336. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-13643-6_21
2. Braker, J., Heinemann, L., Schreieder, T.: Aramis at touché 2022: argument detection in pictures using machine learning. Working Notes Papers of the CLEF (2022)

3. Brummerloh, T., Carnot, M.L., Lange, S., Pfänder, G.: Boromir at touché 2022: combining natural language processing and machine learning techniques for image retrieval for arguments. Working Notes Papers of the CLEF (2022)
4. Carnot, M.L., et al.: On stance detection in image retrieval for argumentation. In: Proceedings of the SIGIR. ACM (2023)
5. Champagne, M., Pietarinen, A.V.: Why images cannot be arguments, but moving ones might. Argumentation **34**(2), 207–236 (2019)
6. Kiesel, J., Reichenbach, N., Stein, B., Potthast, M.: Image retrieval for arguments using stance-aware query expansion. In: Al-Khatib, K., Hou, Y., Stede, M. (eds.) ArgMining 2021 at EMNLP. ACL (2021)
7. Kjeldsen, J.E.: The rhetoric of thick representation: how pictures render the importance and strength of an argument salient. Argumentation **29**(2), 197–215 (2014). https://doi.org/10.1007/s10503-014-9342-2
8. Latif, A., et al.: Content-based image retrieval and feature extraction: a comprehensive review. Math. Probl. Eng. **2019** (2019)
9. Levy, R., Bogin, B., Gretz, S., Aharonov, R., Slonim, N.: Towards an argumentative content search engine using weak supervision. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) Proceedings of the COLING. ACL (2018)
10. Meharban, M., Priya, D.: A review on image retrieval techniques. Bonfring Int. J. Adv. Image Process. **6** (2016)
11. Potthast, M., et al.: Argument search: assessing argument relevance. In: Piwowarski, B., Chevalier, M., Gaussier, É., Maarek, Y., Nie, J., Scholer, F. (eds.) Proceedings of the SIGIR. ACM (2019)
12. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA integrated research architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World. TIRS, vol. 41, pp. 123–160. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22948-1_5
13. Rui, Y., Huang, T.S., Mehrotra, S.: Content-based image retrieval with relevance feedback in mars. In: Proceedings of the ICIP, vol. 2. IEEE (1997)
14. Shao, H., Wu, Y., Cui, W., Zhang, J.: Image retrieval based on MPEG-7 dominant color descriptor. In: Proceedings of the ICYCS. IEEE (2008)
15. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
16. Smith, R.: An overview of the tesseract OCR engine. In: Proceedings of the ICDAR. IEEE (2007)
17. Solli, M., Lenz, R.: Color emotions for multi-colored images. Color Res. Appl. **36** (2011)
18. Stab, C., et al.: ArgumenText: searching for arguments in heterogeneous sources. In: Liu, Y., Paek, T., Patwardhan, M.S. (eds.) Proceedings of the NAACL-HLT. ACL (2018)
19. Wachsmuth, H., et al.: Building an argument search engine for the web. In: Habernal, I., et al. (eds.) Proceedings of the ArgMining@EMNLP. ACL (2017)
20. Zhang, S., Zhang, X., Chan, J., Rosso, P.: Irony detection via sentiment-based transfer learning. Inf. Process. Manag. **56**(5) (2019)