# On Stance Detection in Image Retrieval for Argumentation

Miriam Louise Carnot
Leipzig University and ScaDS.AI

Lorenz Heinemann
Leipzig University

Jan Braker
Leipzig University

Tobias Schreieder
Leipzig University

Johannes Kiesel
Bauhaus-Universität Weimar

Maik Fröbe
Martin-Luther-Universität Halle Wittenberg

Martin Potthast
Leipzig University and ScaDS.AI

Benno Stein
Bauhaus-Universität Weimar

## 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. An important subtask therein is to determine the stance of the retrieved images, i.e., whether an image supports the pro or con side of the topic. In this paper, we conduct a comprehensive reproducibility study of the state of the art as represented by the CLEF'22 Touché lab and an in-house extension of it. Based on the submitted approaches, we developed a unified and modular retrieval process and reimplemented the submitted approaches according to this process. Through this unified reproduction (which also includes models not previously considered), we achieve an effectiveness improvement in argumentative image detection of up to 0.832 precision@10. However, despite this reproduction success, our study also revealed a previously unknown negative result: for stance detection, none of the reproduced or new approaches can convincingly beat a random baseline. To understand the apparent challenges inherent to image stance detection, we conduct a thorough error analysis and provide insight into potential new ways to approach this task.

## CCS CONCEPTS

• **Information systems → Image search**.

## KEYWORDS

argumentation, image retrieval, image stance detection

**ACM Reference Format:**
Miriam Louise Carnot, Lorenz Heinemann, Jan Braker, Tobias Schreieder, Johannes Kiesel, Maik Fröbe, Martin Potthast, and Benno Stein. 2023. On Stance Detection in Image Retrieval for Argumentation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591917


*SIGIR '23, July 23–27, 2023, Taipei, Taiwan*

ACM ISBN 978-1-4503-9408-6/ 23/07…$15.00
https://doi.org/10.1145/3539618.3591917

## 1 INTRODUCTION

With smartphones and increasing connection speeds, social media discussions evolved from being mainly text-focused towards including more and more images or videos. Specific platforms that focus on images, like Instagram, strongly gained in popularity 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 [8] notes that images can underpin and support arguments, clarify facts, and convey them more effectively than words. For example, pictorial health warnings on cigarette packages serve to emphasize and illustrate textual warnings, making the latter more effective [6].

Although retrieval systems for textual arguments have been developed [22], none specifically support image retrieval for argumentation as of yet. 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. While recent works introduced image retrieval for argumentation [7] and a first shared task was conducted in 2022 [2] at the CLEF Touché lab, the presented pioneering approaches achieved unsatisfactory overall performance.

To pave the way for more effective image retrieval systems for arguments, we conduct a detailed investigation into the current state-of-the-art. As part of this investigation, we slightly improve the state-of-the-art. Inspired by the three-stage evaluation of image retrieval for arguments proposed by Kiesel et al. [7], we propose 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 modular 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, stance detection remains extremely challenging: none of the 11 stance models we evaluated convincingly improves over a random baseline. The code for this reproducibility study is available online.[1]

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

The paper is structured as follows: Section 2 reviews related work. Section 3 provides a brief overview of the Touché22 "Image Retrieval for Arguments" dataset and Section 4 introduces our new modular system and details the different models that we employ in our analyses. Section 5 presents the results of our reproduction study, 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 them.

## 2 RELATED WORK

Several previous works exist on retrieving arguments from text collections. The first systems were *args.me* [1, 22], *ArgumenText* [21], and *IBM debater* [10]. For their evaluation, Potthast et al. [14] suggest to employ the retrieved arguments' query relevance as well as rhetorical, logical, and dialectical quality in Cranfield style experiments. However, also more fine-grained aspects of argument quality have been discussed in the literature [22] and could be used for the evaluation of argument search engines.

Approaches for image retrieval have been investigated for many years, mostly in content-based image retrieval. In content-based retrieval, the query is itself an image and relevant results are similar other images. Therefore, the content of the images needs to be analyzed. Smeulders et al. [18] 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. [16]. They used image feature vectors to establish a connection between images and terms. The works of Meharban and Priya [11] and Latif et al. [9] give a more recent overview of approaches and features for web image search. For example, Shao et al. [17] 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 especially promising when retrieving images for arguments due to colors evoking specific emotions [20]—which are part of the persuasive power of images. A relatively new approach in image retrieval is to employ optical character recognition software like Tesseract [19] 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. In this work we focus on image search using a textual query and leave the task of finding supportive or attacking images for a user-provided image to future work.

The retrieval of images for arguments has been sparsely explored so far. The pioneering work by Kiesel et al. [7] 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, as our comparative evaluation shows, not able to improve upon a random classifier with regard to stance detection. The same authors then organized a shared task at the CLEF 2022 Touché lab [2]. We employ the lab's data and the two most effective participating approaches in our system comparison. They are summarized in the respective sections (3 and 4).

## 3 RE-USING THE TOUCHÉ'22 DATASET

For our investigation into the state-of-the-art in image retrieval for arguments we employ the dataset of the corresponding Touché'22 shared task [2], which was located at the CLEF 2022 conference. The data is freely accessible online.[2] The dataset contains 23,841 images for 50 controversial topics (as queries). The topics include, for example, "can alternative energy effectively replace fossil fuels?" "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 (Section 5) we employ the queries, the image pixel values and recognized text, the corresponding web page's title and HTML source code, and the rank in the result list. 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 Figure 1, are taken from this dataset.

## 4 UNIFYING ARGUMENT IMAGE RETRIEVAL

Inspired by the three-stage evaluation of image retrieval for arguments by Kiesel et al. [7], we propose a modular retrieval system with three AI models. According to Kiesel et al. [7], an image is considered relevant if it is topic-relevant, argumentative, and stance-relevant. Topic relevance evaluates whether a retrieved image fits the topic, i.e., the query. Argumentativeness evaluates whether a retrieved image can be seen as a statement on some topic. Stance relevance evaluates whether a retrieved image fits a previously specified stance (pro/support or con/attack) on the topic.

The best-performing system at the CLEF Touché lab was developed by Team Boromir [4]. For retrieval, the query is pre-processed. Both the pre-processed web page texts and the processed image texts are used with a boost for the image text. For stance detection, they fine-tuned a BERT model which shall be explained later on. We used their system as inspiration but decided on a different architecture. Figure 1 illustrates our new modular system: a topic model to rank images by their relevance to a query (Section 4.1), an argument model to rank images by their suitability for argumentation (Section 4.2), and a stance model to sort images into pros and cons (Section 4.3). Images are ranked by the sum of the topic model's and argument model's scores and presented as two ranked lists induced by the stance model's classification. Of the three models, only the argument model is query-independent. The argument model can thus be employed within the indexing process (also called "offline").

The modular architecture allows for detailed investigations of system performance by replacing single models at a time. In this work, we focus on the stance model, as we identified stance detection as the most challenging subtask for now. However, similar investigations regarding the other models are equally possible in the future. The following sections first introduce both the models in general and the specific models used in our analysis (Section 5). Table 1 provides an overview of the features each model employs.

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

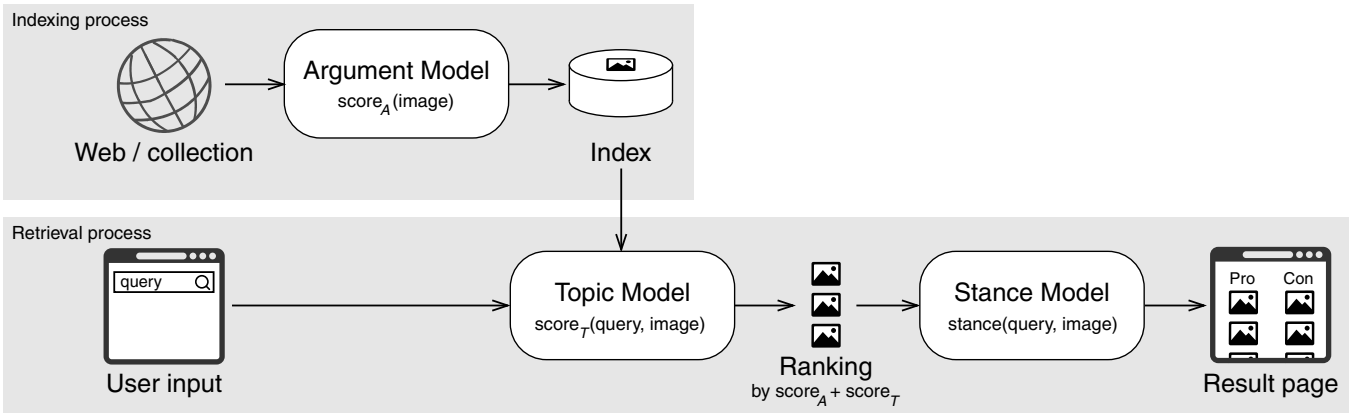



**Figure 1: Schematic of the unified image retrieval system for arguments. In the indexing process, images from the web or a collection are, together with the web pages they appeared on, scored by the argument model for argumentativeness (score$_A$) and indexed. In the retrieval process, the user issues 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.**

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

| Model | Query | Image features | | | | |
|---|---|---|---|---|---|---|
| | Text | Image file | | Web page | | SERP |
| | | Pixels | Text | Title | Source | Rank |
| Topic model | ✓ | | ✓ | | ✓ | |
| Argument model | | ✓ | ✓ | | | |
| *Stance models* | | | | | | |
| Oracle | | | | | | |
| Both-sides baseline | | ✓ | ✓ | | | |
| Random baseline | | | | | | |
| Crawl query stance | ✓ | | | | | ✓ |
| CLIP query stance | ✓ | ✓ | | | | |
| BERT title sentiment | | | | ✓ | | |
| AFINN text sentiment | | | | | ✓ | |
| Aramis Formula | ✓ | ✓ | ✓ | | ✓ | |
| Aramis Neural | ✓ | ✓ | ✓ | | ✓ | |
| Neural text+image 3class | ✓ | ✓ | ✓ | | | |
| Neural text+image 2x2class | ✓ | ✓ | ✓ | | | |
| Neural text 3class | ✓ | | ✓ | ✓ | | |
| Neural text+page 3class | ✓ | | ✓ | ✓ | ✓ | |

## 4.1 The Topic Model

In our modular retrieval system, the topic model ranks images by their relevance to the user's query by assigning a score to each image in the index (cf. Figure 1). As the score depends on the query, the topic model must be part of the retrieval process.

For our analysis, the topic model we use combines the respective features of the two best-performing approaches in the Touché'22 shared task, Boromir [4] and Aramis [3]. Specifically, we employ textual matching of the query and text from the image's context (web page) and from the image itself. The recognized text on the image and the query are preprocessed using standard lowercasing and stopword and punctuation removal. The text from the HTML source code of the web page on which the image was originally found gets extracted and is also being preprocessed. 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. Section 5), we did not investigate further models but focused our attention on different stance models instead.

## 4.2 The Argument Model

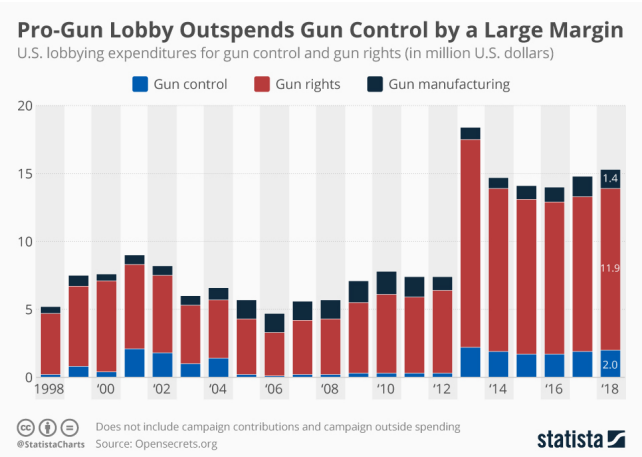
In our modular retrieval system, the argument model ranks images by their suitability for argumentation by assigning a score to each image in the index (cf. Figure 1). Conceptually, an image that shows either critical or supportive attitudes should receive a high argument score. Unlike for the topic model, this score does not depend on the query and the argument model can thus be part of the indexing process. Therefore, the model's score for each image is calculated at indexing time and indexed alongside the image, and directly used in the retrieval function.

For our analysis, the argument model we use employs the query-independent features that are also employed by the Aramis approach for the Touché'22 shared task [3]. Furthermore, we employ the same neural network classifier as Aramis for calculating the argumentativeness score from the features but train it on the official ground truth ratings that the organizers released after the shared task instead of the self-created ratings that Aramis used.[3] We detail those features below for completeness.

The first set of features our specific argument model employs 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. Braker et al. [3] argue that red and green are often used to express a

[3]To avoid test set leakage we employ cross-validation as described in Section 5.

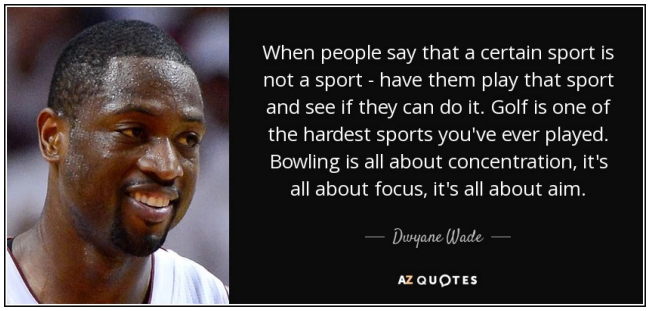

(a)




(b)

**Figure 2: (a) Example image for the usage of the colors red and green to express opinions. (b) Example image for the usage of short texts (axis labels, legend) in diagrams.**


(a)




(b)

**Figure 3: (a) Example image for the usage of visual content between lines of text in a meme. (b) Example image for the usage of a photo next to text in a picture-quote.**

stance, with blue and yellow then used for comparison. Figure 2 (a) shows one example image for illustration. Note that the argument model does not predict the stance, but only argumentativeness.

Other features used for the neural network are the image type (graphic or photography) and diagram-likeness. We adopt the simple heuristic of Aramis for image type classification [3]: If the ten most common colors make up more than 30% of the image it is classified as graphic (cartoon, clip art, ...), otherwise as photography. For diagram-likeness, we also employ the heuristic of Aramis which is to use the percentage of the image covered by short texts. A horizontal kernel is used to remove larger texts in horizontal orientation, leaving vertical texts and short pieces as used in diagrams (cf. Figure 2 (b)). Both Braker et al. [3] and Brummerloh et al. [4] argue that diagrams usually have an argumentative character.

The final set of features from Aramis concerns the use of text in general: text length, sentiment, the area percentage of the image occupied by text, and the position of the text in an 8x8 grid [3]. The sentiment score is again used to identify opinions. The usage of text position is used as a hint to identify both memes, which are often argumentative and often use vertical space between the lines of texts to show visual content(cf. Figure 3 (a)), and one widespread form of picture quotes, which often have text either to the left or right of a photo (cf. Figure 3 (b)). The text is extracted using Tesseract OCR[4] 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. A convolutional layer is used to process the 8x8 text grid.

## 4.3 The Stance Model

In our modular retrieval system, the stance model sorts the ranked images into pros and cons (cf. Figure 1). To this end, stance models label each image for a topic as pro, con, both, or neither (cf. Kiesel et al. [7]). Images labeled as neither pro nor con are discarded, whereas the others are placed on the result page in the respective column in decreasing score order. Note that, according to the Touché task definition, an image can be both pro and con, in which case it

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

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 results of the Touché'22 lab show that none of the participating models achieved a high precision for stance detection specifically [2]. We thus focused our investigation on the stance detection subtask and compared 14 approaches, including two baseline approaches and the oracle. Specifically, we compare the best approach of Toucheé'22 (Boromir) with the following 13 approaches:

**Oracle.** This theoretic approach 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 indeed 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%.

**Both-sides baseline.** This baseline classifies each image as both pro and con, which results in two identical result lists.

**Random baseline.** This baseline classifies each image as either pro or con with equal probability.

**Crawl query stance.** This approach labels each image based on which result list it was originally found while crawling. If the image occurred in the top 100 of the result list where the query was extended with "good," it is labeled as pro, and likewise for "anti" and con. This method thus corresponds to the stance detection part of the stance-aware query expansion approach by Kiesel et al. [7].

**CLIP query stance.** This approach is a modification of the crawl query stance approach that employs CLIP [15] instead of the crawled result lists. It uses CLIP to compute the image's similarity to (1) the query extended with "good" for pro and (2) the query extended with "anti" for con. It selects the stance with the higher similarity.

**BERT title sentiment.** This approach employs the stance detection model of the best-performing approach at the Touché'22 shared task, Boromir [4]: a sentiment detection BERT-model trained on the Large Movie Review Dataset.[5] The model is used to classify the sentiment of the title of the image's original web page. Images, where the title is classified as positive, are labeled as pro, and those where it is classified as negative are labeled as con. Brummerloh et al. [4] argue that the sentiment of the title usually indicates the sentiment of the entire article and thus also for the images on that web page which are used to underline this opinion.

**AFINN text sentiment.** This approach employs the alternate stance detection model of Boromir [4], which reached a lower score in the Touché'22 shared task: sentiment detection using the AFINN dictionary which was created by Nielsen [12]. For each word in the web page's text, its score in the AFINN dictionary is looked up and then summed up. Single scores range from -5 for very negative to 5 for very positive. If the sum is negative this approach labels the image as con and if the sum is positive as pro.

**Aramis Formula.** This approach uses the heuristic formula developed by team Aramis that is based on thirteen different features [3]. They are calculated from, amongst others, the query, the image text, the HTML text around the image, the interrelation and sentiments of the mentioned texts, and the colors in the image. The weights for each feature were set manually by the Aramis group of the Touché lab based on their assumptions.

[5]https://ai.stanford.edu/~amaas/data/sentiment/

**Aramis Neural.** This neural network, also developed by team Aramis [3], uses the same features as the Aramis Formula to classify images as either pro, neutral or con. The neutral images are not further used in the results.

**Neural text+image 3class.** This approach employs a feedforward neural network classifier using the image resized to 256x256 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.** This approach employs the same architecture as neural text+image 3class 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.

**Neural text 3class.** This approach is the same as neural text+image 3class, but instead of using the image pixels, it employs the title of the image's original web page as input. Like neural text+image, it also employs the query and the recognized text of the images.

**Neural text+page 3class.** This approach is the same as neural text 3class but also uses the HTML text in a window around the image.

## 5 REPRODUCING AND EXTENDING THE TOUCHÉ'22 BENCHMARK

Table 2 presents the result of our detailed and extended reproduction. For consistency with the existing evaluation and fair comparison, we only use the 6607 images of the Touché'22 dataset for which ratings exist 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.

Besides comparing more approaches, our evaluation also goes deeper than the original one of Bondarenko et al. [2] in that it shows results also for pro and con separately, and employs NDCG@10 in addition to precision@10 as used in the Touché lab. The Touché lab only used precision@10, arguing that this metric resembles closest the setting of a user looking at a single page of result images. However, rank-based metrics might be more appropriate in other situations. Nevertheless, as Table 2 reveals, the scores for precision@10 and NDCG@10 are very similar, with one exception: the both-sides baseline would rank considerably lower when using NDCG@10 than when using precision@10.

### 5.1 Topic and Argument Retrieval

We first detail the results for the retrieval of topic-relevant and argumentative images, i.e., without stance detection. This setup corresponds to omitting the stance model in Figure 1. The topic and the argument models are used for the retrieval of topic-relevant and argumentative images and assign a score to the images. These 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 $\text{score}_T$ and the argument model's $\text{score}_A$ than the simple sum, but none lead to significant improvements.

**Table 2: The table shows the precision@10 and NDCG@10 scores on condensed lists for all 50 topics sorted by precision@10 for 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 topic model and the argument model as described in section 4, except for Best of Touché'22 and the Oracle.**

| Stance Model | Precision@10 | | | | | | | | | NDCG@10 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Topic-relevance | | | Argumentativeness | | | Stance-relevance | | | Topic-relevance | | | Argumentativeness | | | Stance-relevance | | |
| | Pro | Con | Both | Pro | Con | Both | Pro | Con | Both | 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 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.929 | 0.964 |
| Neural text+image 2x2class | 0.924 | 0.822 | 0.873 | 0.830 | 0.766 | 0.798 | 0.660 | **0.310** | **0.485** | 0.928 | 0.847 | 0.887 | 0.831 | 0.789 | 0.810 | 0.657 | **0.341** | **0.499** |
| BERT title sentiment | 0.892 | 0.872 | 0.882 | 0.806 | 0.802 | 0.804 | 0.674 | 0.250 | 0.462 | 0.909 | 0.885 | 0.897 | 0.813 | 0.814 | 0.814 | 0.673 | 0.266 | 0.470 |
| CLIP query stance | **0.932** | **0.932** | **0.932** | 0.836 | 0.824 | 0.830 | 0.662 | 0.256 | 0.459 | **0.937** | **0.934** | **0.935** | **0.843** | 0.830 | **0.836** | 0.667 | 0.267 | 0.467 |
| Aramis Formula | 0.920 | 0.814 | 0.867 | **0.838** | 0.742 | 0.790 | **0.690** | 0.216 | 0.453 | 0.920 | 0.837 | 0.878 | 0.835 | 0.757 | 0.796 | **0.685** | 0.239 | 0.462 |
| Both-sides baseline | 0.926 | 0.926 | 0.926 | 0.832 | **0.832** | **0.832** | 0.662 | 0.232 | 0.447 | 0.928 | 0.928 | 0.928 | 0.831 | **0.831** | 0.831 | 0.658 | 0.246 | 0.452 |
| Neural text+image 3class | 0.924 | 0.866 | 0.895 | 0.830 | 0.800 | 0.815 | 0.660 | 0.226 | 0.443 | 0.928 | 0.878 | 0.903 | 0.831 | 0.805 | 0.818 | 0.657 | 0.234 | 0.446 |
| Random baseline | 0.894 | 0.888 | 0.891 | 0.816 | 0.812 | 0.814 | 0.664 | 0.222 | 0.443 | 0.908 | 0.895 | 0.901 | 0.823 | 0.815 | 0.819 | 0.654 | 0.239 | 0.447 |
| Aramis Neural | 0.694 | 0.676 | 0.685 | 0.668 | 0.640 | 0.654 | 0.588 | 0.278 | 0.433 | 0.733 | 0.708 | 0.721 | 0.703 | 0.668 | 0.686 | 0.602 | 0.303 | 0.453 |
| Best of Touché'22 (Boromir) | 0.884 | 0.872 | 0.878 | 0.782 | 0.754 | 0.768 | 0.594 | 0.256 | 0.425 | 0.895 | 0.877 | 0.886 | 0.787 | 0.746 | 0.767 | 0.609 | 0.260 | 0.435 |
| Crawl query stance | 0.830 | 0.728 | 0.779 | 0.744 | 0.694 | 0.719 | 0.610 | 0.214 | 0.412 | 0.842 | 0.761 | 0.801 | 0.761 | 0.720 | 0.740 | 0.612 | 0.227 | 0.420 |
| AFINN text sentiment | 0.766 | 0.908 | 0.837 | 0.708 | 0.814 | 0.761 | 0.564 | 0.222 | 0.393 | 0.797 | 0.904 | 0.851 | 0.735 | 0.809 | 0.772 | 0.587 | 0.241 | 0.414 |
| Neural text+page 3class | 0.644 | 0.616 | 0.630 | 0.598 | 0.560 | 0.579 | 0.504 | 0.154 | 0.329 | 0.691 | 0.675 | 0.683 | 0.649 | 0.611 | 0.630 | 0.541 | 0.176 | 0.358 |
| Neural text 3class | 0.668 | 0.668 | 0.668 | 0.602 | 0.602 | 0.602 | 0.458 | 0.190 | 0.324 | 0.704 | 0.704 | 0.704 | 0.632 | 0.632 | 0.632 | 0.469 | 0.219 | 0.344 |

As seen in Table 2, with a precision@10 of 92.6% for topic-relevance and 83.2% for argumentativeness, the both-sides baseline outperforms all methods that competed in the CLEF 2022 Touché lab. For reference, the most effective method from the lab, developed by team 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, the CLIP query stance model retrieves always 20 images and reaches nearly the same performance as the both-sides baseline, even a slightly better one in terms of NDCG@10. 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 AFINN text sentiment. 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%, shown by the oracle. This is because not every topic has ten images on each side in the evaluation data, which is particularly common on the con side. Missing images are treated in the same way as incorrect images. We find that the neural text+image 2x2class model, which uses the image and associated text as input, is the most effective with a precision@10 of 48.5%. The BERT title sentiment model using only the title of the web page where the image appeared on and the CLIP query stance approach reach places two and three with a precision@10 of 46.2% and 45.9%, respectively. On the pro side, the Aramis Formula model performed best, exceeding 69%. The results range from 45.8% 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 15.4% and 31.0%. The best theoretically possible result is 80.2% (oracle).

However, Table 2 also shows the main negative result of our reproduction: 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 both-sides 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 and NDCG@10, we found that only the oracle improves over it significantly. Worse still, a number of models—specifically those that employ only text features—were not able to outperform the random nor the both-sides baseline model. 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 Section 5 reproduced the seemingly good results of the approaches submitted to 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:

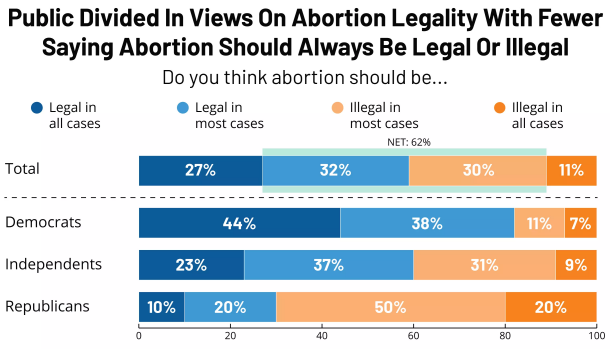



**Figure 4: Different valuations cause stance ambiguity. The image could be pro "should abortion be legal?" if one thinks very highly of Democrats, but con if one despises them.**

*Semantic gap for diagrams.* Charts and diagrams usually present information in the form of lines, bars, or others. The length or size of those geometrical forms can be interpreted by humans but does not have a deeper meaning to the system. Recognizing the stance of a diagram requires a semantic understanding of the image and the world that goes beyond current methods. To solve this problem it is thus necessary to integrate approaches that semantically interpret diagrams, like recent transformer models that generate natural language descriptions for these (e.g., [13]).

*Different valuations cause stance ambiguity.* Some images, especially diagrams, often provide several pieces of information. Therefore, different audiences might draw different or even opposite conclusions from the same image. Specifically, a person's background, socialization, and opinions influence whether they consider entities or events positive or negative—and thus whether an image related to the entity or event could serve as pro or con. For example, several diagrams in the dataset refer to opinions based on affiliation with political parties, as in Figure 4. Someone who feels being part of the Democrats / Republicans sees in the diagram that their favorite party is clearly in favor of / against legal abortion and could thus see the image as pro / con for that topic. This problem is challenging for both algorithms and annotation campaigns. To solve this problem 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, especially symbolic ones, require the viewer to have certain background knowledge to understand why they could be pro or con. For example, for the topic "should abortion be legal?" the image of a coat hook could be classified as con, as these items were used in a very dangerous abortion practice. But viewed without that knowledge, the image is not even topic-relevant. As one more example, the image in Figure 5 is pro "is human activity primarily responsible for global climate change?" for viewers who make the connection between the burning of forests and fields and damage to the climate. Again, the viewer's knowledge and opinions play a

**Figure 5: Image understanding depends on background knowledge. The image could be pro "is human activity primarily responsible for global climate change?" depending on how the viewer connects it to their background knowledge.**

crucial role. This problem provides a challenge both for algorithms and annotation campaigns. Analyzing the context in which the image is used (i.e., the web page) could provide hints on the relevant knowledge and connections.

*Regional images.* A problem that has so far not been an issue in the Touché lab is that some images, but also some topics, are only of relevance for people in some regions of the world. The Touché lab focussed on US topics and considered image relevance from a US perspective. However, the relevance of arguments (both textual and visual) for some topics can change between regions. For example the topic "should bottled water be banned?" is only relevant to countries in which many people buy bottled water. In a country in which drinking tap water is not safe, images showing related illnesses would be considered con. Also, several systems retrieve for "is a college education worth it?" mostly images like in Figure 6, which cover information for the US only. This problem provides a challenge both for algorithms and annotation campaigns. For algorithms, one could train separate models on data from different regions. For annotation campaigns, one could ensure regional diversity within the group of annotators and mark disagreements correspondingly. Moreover, the dataset should include information on which country or region an image stems from.

*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.

*Both stances in one image.* Some images explicitly relate to both stances at the same time. For example, Figure 7 shows an image that lists textual arguments for both stances on "should adults be allowed to carry a concealed handgun?" Other images with both

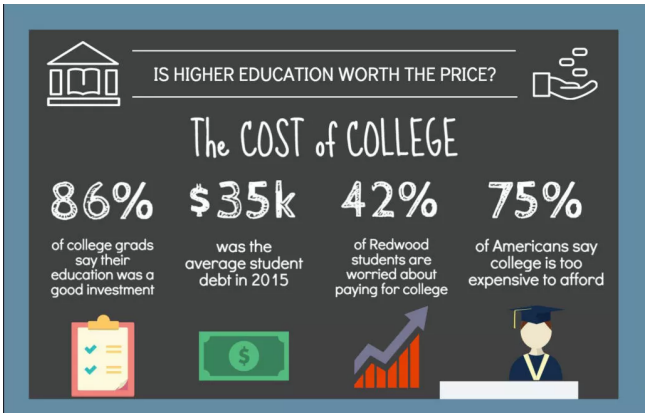



**Figure 6: Regional images. The image might be con "is a college education worth it?" but is only really relevant for people in the US.**


**Figure 7: Both stances in one image. The image explicitly contains arguments pro and con "should adults be allowed to carry a concealed handgun?"**

stances show different groups of people, some indicating a pro image and some indicating con. To solve this problem such images could be identified and classified as belonging to a special "both" category as suggested by Kiesel et al. [7].

*Neutral images.* Some images, especially some diagrams, contain thought-provoking impulses when considering a topic, but are not evidently pro nor con. However, they might be visually very similar to arguments that are clearly pro and con, which can be a problem during training. For example, the image in Figure 8 is very informative without clearly being pro or con "is vaping with e-cigarettes safe?" However, one can imagine visually very similar images that are clearly pro or clearly con, which provides a challenge in classifier training. To solve this problem it might be necessary to develop a classifier to detect neutral images. Such approaches likely need semantic interpretations of the images, as already suggested in the semantic gap for diagrams challenge.

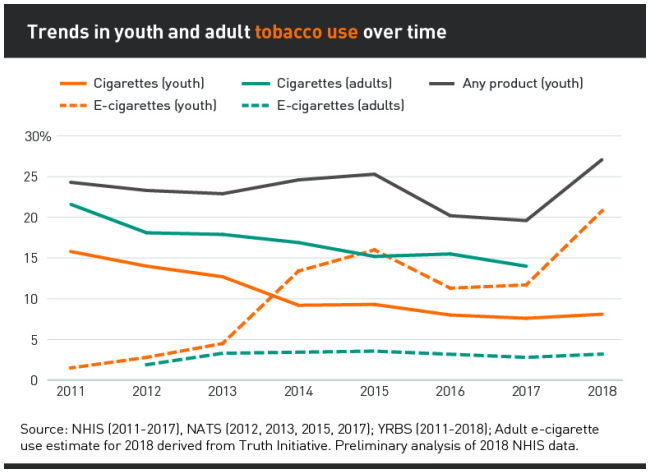



**Figure 8: Neutral images. The image is neither clearly pro nor clearly con "is vaping with e-cigarettes safe?" but one can easily think of modifications that would make it point towards a certain stance.**

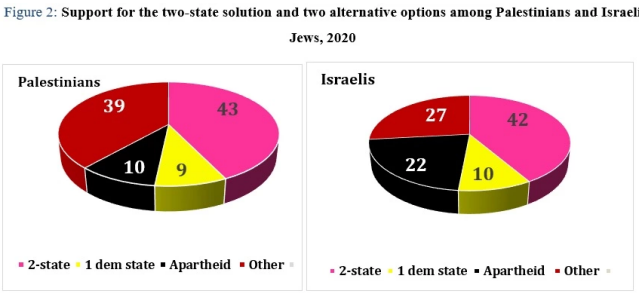



**Figure 9: More than two stances for "is a two-state solution an acceptable solution to the Israeli-Palestinian conflict?"**

*More than two stances.* For some topics, there are more than two possible stances, making a binary classifier the wrong choice. For example, Figure 9 names three different stances (and "other"!) for a solution for the Israeli-Palestinian conflict. To solve this problem one could cluster images instead of classifying them. However, a solution will likely require adapting the task, making the new task incompatible with the available data.

*Irony and Jokes.* Many images in the dataset, especially memes, make use of irony and jokes. Such kind of humor may not be understood by all people, and neither by algorithms. Figure 10 shows an image that was retrieved by one of the reproduced systems for the topic "do violent video games contribute to youth violence?" The image is a joke on the idea that video games created violent behavior, as if violence had not existed before video games. The irony is used to undermine the arguments of the opposing side. We expect irony detection for images to be very challenging. Still, it might be possible to transfer advances in textual irony detection (e.g., [23]) to visual irony detection.


**Figure 10: Irony and Jokes. The image is con "do violent video games contribute to youth violence?" only if one understands the joke about "pong" being a violent video game.**

## 7 CONCLUSION

For the task of image retrieval for argumentation, we compared 14 approaches (including the previous state-of-the-art, two baselines, and the oracle) while emphasizing the subtask of stance detection. We reproduced the setup of the Touché'22 lab at CLEF, but considerably extended the analysis. To compare different approaches, we proposed a modular image retrieval system: 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 in our study employ features of the query, the image file (e.g., pixels), the web page an image was indexed on, or the rank at which an image was found by different queries to Google. The approaches for the topic and the argument models that we combined from the reproduced submissions to the Touché'22 lab provide a new state-of-the-art for the respective parts of the task, reaching 0.932 precision@10 for topic-relevance and 0.832 precision@10 for argumentativeness.

However, the extended analysis of our reproduction also uncovered a strong negative result: none of the reproduced or new 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. For each challenge, we offer examples and propose possible approaches to address the challenge.

## ACKNOWLEDGEMENTS

This work has been funded by the EC under GA 101070014 (OpenWebSearch.eu)

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI

## REFERENCES

[1] Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me Corpus. In *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings (Lecture Notes in Computer Science, Vol. 11793)*, Christoph Benzmüller and Heiner Stuckenschmidt (Eds.). Springer, 48–59. https://doi.org/10.1007/978-3-030-30179-8_4

[2] Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022. Overview of Touché 2022: Argument Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings (Lecture Notes in Computer Science, Vol. 13390)*, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer, 311–336. https://doi.org/10.1007/978-3-031-13643-6_21

[3] Jan Braker, Lorenz Heinemann, and Tobias Schreieder. 2022. Aramis at Touché 2022: Argument Detection in Pictures using Machine Learning. *Working Notes Papers of the CLEF* (2022).

[4] Thilo Brummerloh, Miriam Louise Carnot, Shirin Lange, and Gregor Pfänder. 2022. Boromir at Touché 2022: Combining Natural Language Processing and Machine Learning Techniques for Image Retrieval for Arguments. *Working Notes Papers of the CLEF* (2022).

[5] Marc Champagne and Ahti-Veikko Pietarinen. 2019. Why Images Cannot be Arguments, but Moving Ones Might. *Argumentation* 34, 2 (June 2019), 207–236. https://doi.org/10.1007/s10503-019-09484-0

[6] Fachbereich WD 5: Wirtschaft und Verkehr, Ernährung, Landwirtschaft und Verbraucherschutz. 2017. *Wirksamkeit von bildlichen Warnhinweisen auf Zigarettenpackungen.* Technical Report WD 5 - 3000 - 024/17. Deutscher Bundestag.

[7] Johannes Kiesel, Nico Reichenbach, Benno Stein, and Martin Potthast. 2021. Image Retrieval for Arguments Using Stance-Aware Query Expansion. In *8th Workshop on Argument Mining (ArgMining 2021) at EMNLP*, Khalid Al-Khatib, Yufang Hou, and Manfred Stede (Eds.). Association for Computational Linguistics, 36–45. https://doi.org/10.18653/v1/2021.argmining-1.4

[8] Jens E. Kjeldsen. 2014. The Rhetoric of Thick Representation: How Pictures Render the Importance and Strength of an Argument Salient. *Argumentation* 29, 2 (Dec. 2014), 197–215. https://doi.org/10.1007/s10503-014-9342-2

[9] Afshan Latif, Aqsa Rasheed, Umer Sajid, Jameel Ahmed, Nouman Ali, Naeem Iqbal Ratyal, Bushra Zafar, Saadat Hanif Dar, Muhammad Sajid, and Tehmina Khalil. 2019. Content-based image retrieval and feature extraction: a comprehensive review. *Mathematical Problems in Engineering* 2019 (2019).

[10] Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, 2066–2081. https://aclanthology.org/C18-1176/

[11] M.S. Meharban and Dr.S. Priya. 2016. A Review on Image Retrieval Techniques. *Bonfring International Journal of Advances in Image Processing* 6, 2 (April 2016), 07–10. https://doi.org/10.9756/bijaip.8136

[12] Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903* (2011).

[13] Jason Obeid and Enamul Hoque. 2020. Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. In *Proceedings of the 13th International Conference on Natural Language Generation (INLG'20)*, Brian Davis, Yvette Graham, John D. Kelleher, and Yaji Sripada (Eds.). Association for Computational Linguistics, 138–147. https://aclanthology.org/2020.inlg-1.20/

[14] Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument Search: Assessing Argument Relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 1117–1120. https://doi.org/10.1145/3331184.3331327

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[16] Yong Rui, Thomas S Huang, and Sharad Mehrotra. 1997. Content-based image retrieval with relevance feedback in MARS. In *Proceedings of international conference on image processing*, Vol. 2. IEEE, 815–818.

[17] Hong Shao, Yueshu Wu, Wen-cheng Cui, and Jinxia Zhang. 2008. Image Retrieval Based on MPEG-7 Dominant Color Descriptor. In *Proceedings of the 9th International Conference for Young Computer Scientists, ICYCS 2008, Zhang Jia Jie, Hunan, China, November 18-21, 2008*. IEEE Computer Society, 753–757. https://doi.org/10.1109/ICYCS.2008.89

[18] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence* 22, 12 (2000), 1349–1380.

[19] R. Smith. 2007. An Overview of the Tesseract OCR Engine. In *9th International Conference on Document Analysis and Recognition (ICDAR'07)*. IEEE Computer Society, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

[20] Martin Solli and Reiner Lenz. 2011. Color emotions for multi-colored images. *Color Research & Application* 36, 3 (April 2011), 210–221. https://doi.org/10.1002/col.20604

[21] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. 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-HLT 2018, New Orleans, Louisiana, USA, June 2-4, 2018, Demonstrations*, Yang Liu, Tim Paek, and Manasi S. Patwardhan (Eds.). Association for Computational Linguistics, 21–25. https://doi.org/10.18653/v1/n18-5005

[22] Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, Ivan Habernal, Iryna Gurevych, Kevin D. Ashley, Claire Cardie, Nancy L. Green, Diane J. Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern R. Walker (Eds.). Association for Computational Linguistics, 49–59. https://doi.org/10.18653/v1/w17-5106

[23] Shiwei Zhang, Xiuzhen Zhang, Jeffrey Chan, and Paolo Rosso. 2019. Irony detection via sentiment-based transfer learning. *Information Processing and Management* 56, 5 (2019), 1633–1644. https://doi.org/10.1016/j.ipm.2019.04.006