Image Retrieval for Arguments Using Stance-Aware Query Expansion

Johannes Kiesel
Bauhaus-Universität Weimar
johannes.kiesel@uni-weimar.de

Nico Reichenbach
Leipzig University
nico.reichenbach@posteo.de

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

Martin Potthast
Leipzig University
martin.potthast@uni-leipzig.de

Abstract

Many forms of argumentation employ images as persuasive means, but research in argument mining has been focused on verbal argumentation so far. This paper shows how to integrate images into argument mining research, specifically into argument retrieval. By exploiting the sophisticated image representations of keyword-based image search, we propose to use semantic query expansion for both the pro and the con stance to retrieve "argumentative images" for the respective stance. Our results indicate that even simple expansions provide a strong baseline, reaching a precision@10 of 0.49 for images being (1) on-topic, (2) argumentative, and (3) on-stance. An in-depth analysis reveals a high topic dependence of the retrieval performance and shows the need to further investigate on images providing contextual information.

1 Introduction

Images are a prominent form of non-verbal communication. Images convey messages in diverse ways including scribbles, concept drawings, (produced) photos, data visualizations, or internet memes. Moreover, images can serve as an inspiration or an overview of the topic at hand. They play a key role in public discourse (Woods and Hahner, 2019) and in expressing personal opinion (Heiskanen, 2017), e.g., in the form of political memes. Spread "virally," they can gain a large followership on social media and influence political decision-making (Watt, 2015) or serve as a form of evidence for troublesome events. For example, about 150,000 images are uploaded to Facebook every minute.[1] All of these uses often form an integral part of argumentative writing or speechmaking. Composing an argumentative text or speech hence does not only include the retrieval and arrangement of written arguments, but also that of relevant "argumentative images" to go along with them.

Figure 1: Mock-up of an image search engine for arguments, featuring a query box, checkboxes to filter, and results categorized as pro (left) and con (right) images.

However, though the retrieval of written arguments has become a well-established line of research, the paper at hand is the first of its kind to define and tackle the task of retrieving "argumentative images." Inspired by existing argument search interfaces, Figure 1 illustrates how an image search engine for arguments might look.

By proposing stance-aware query expansion, this paper shows how to harness existing keyword-based image retrieval technology for the task of retrieving images for arguments. The key assumption of this expansion is that adding terms to the user's query that indicate a stance (i.e., pro or con the query's underlying issue), then images retrieved for the expanded queries support that stance. Figure 1 illustrates the desired effect: the shown images stem from Google's image search after having expanded the query `nuclear energy` with either the term `good` (images on the left-hand side) or the term `anti` (right-hand side).

[1] https://www.domo.com/learn/data-never-sleeps-8

In what follows, Section 3 introduces the task of "argumentative image" retrieval, Section 4 outlines our approach to this task—stance-aware query expansion—including three methodological variants, and the Sections 5 and 6 detail important elements of our experimental setting as well as evaluation results for the three proposed methods.

2 Related Work

Al-Khatib, 2019 defines an argumentation strategy as "a set of principles that guides the selection and arrangement of arguments (plus contextual information) in an argumentative discourse." While research in computational argumentation has so far focused on verbal argumentation, real-world discourse often integrates images to great effect, ranging from memes emerging on the spur of of a moment to figures summarizing months of research. Discourse analyses thus often include images (e.g., Frohmann, 1992; Farkas and Bene, 2020).

Images for Arguments Images can provide contextual information and express, underline, or popularize an opinion (Dove, 2012), thereby taking the form of subjective statements (Dunaway, 2018). Some images express both a premise and a conclusion, making them full arguments (Roque, 2012; Grancea, 2017). Other images may provide contextual information only and need to be combined with a conclusion to form an argument. In this regard, a recent SemEval task distinguished a total of 22 persuasion techniques in memes alone (Dimitrov et al., 2021). Moreover, argument quality dimensions like acceptability, credibility, emotional appeal, and sufficiency (Wachsmuth et al., 2017a) all apply to arguments that include images as well. And as a kind of visual argumentation scheme (a "stereotypical pattern of human reasoning"; Walton et al., 2008), some images are frequently adapted to different topics (Heiskanen, 2017). Social groups even create their own symbolisms and use them to express opinions (e.g., fringe web communities; Zannettou et al., 2018). The potentially high emotional impact of images to a vast audience (Adler-Nissen et al., 2020) can cause changes in the social discourse and eventually politics (Woods and Hahner, 2019). Examples include photos of a drowned refugee child (Barnard and Shoumali, 2015; D'Orazio, 2015), police violence (Berger, 2010), or the recent departure of Western military forces from Afghanistan.

Image search Keyword-based image search analyzing the content of images or videos has been studied for decades (Aigrain et al., 1996), pre-dated only by approaches relying on metadata and similarity measures (Chang and Fu, 1980). In a recent survey, Latif et al. (2019) 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 (Wu, 2020; Google, 2021). Also related to the retrieval of argumentative images is that of "emotional images", which relies on image features like color and composition (Wang and He, 2008; Solli and Lenz, 2011). Argumentation goes hand in hand with emotions, so that emotional features may be promising for retrieving images for arguments in the future. For lack of labeled data (argumentativeness plus emotionality), we start by exploiting keyword-based web search to retrieve images for arguments from the web, as did the earliest image search approaches (e.g., Yanai, 2001).

Argument search Based on argument mining from texts (cf. Peldszus and Stede, 2013), argument search engines aim to support decision-making and persuasion. Conceptually, a query to an argument search engine may either name an issue without a stance (Stab et al., 2018) (e.g., `nuclear energy`), or represent a conclusion for which supporting and attacking premises are to be retrieved (e.g., `nuclear energy mitigates climate change`). The first collection of arguments from the web is the Internet Argument Corpus (Walker et al., 2012), containing 400,000 posts from an online debate portal. The first argument search engine, args.me, indexes a similar dataset of arguments (Wachsmuth et al., 2017b). Not relying on retrieval in collections of arguments, ArgumenText (Stab et al., 2018) first searches for documents relevant to a user's query in generic web crawls, and then mines arguments on the fly within retrieved documents. Regarding the evaluation of argument search engines, judging the topic relevance of a retrieved text alone is insufficient, it must also be argumentative (Potthast et al., 2019; Bondarenko et al., 2021). Research on argumentation has identified many further quality criteria for arguments (Wachsmuth et al., 2017a), yet few have been investigated for argument retrieval, and hardly anything has been said on the argumentative quality of images.

3 Image Retrieval for Arguments

Similar to textual argument retrieval, we define the task of retrieving images for arguments as follows:

> Given a keyword query suggesting an issue or a claim for a topic, retrieve as two ranked lists those and only those images that can assist someone in (1) supporting and (2) attacking it.

This definition presumes that a given user intends to search for images suitable to assist in persuading others, an intent that encompasses a diversity of real-life scenarios, such as sending a pertinent image to a friend, using it as a cover for a news article or blog post, or on a slide in a presentation. Furthermore, it encompasses deliberation scenarios, such as forming an own opinion, and creating a collage for a school project. We expect search engines for argumentative images to offer facets as shown in Figure 1 to meet such needs.

To assess the relevance of an image to a given query, a three-fold judgment is required:

Topic relevance The image content is related to the query topic. This criterion corresponds to the notion of relevance in keyword-based image retrieval (Shanbehzadeh et al., 2000).

Argumentativeness The image can be used to support a stance regarding the query topic. This criterion corresponds to the notion of a context-dependent claim in textual argument mining (Aharoni et al., 2014).

Stance relevance The image can be used to support the predicted stance within the query topic. This criterion corresponds to the categorization into pros and cons in standard argument search (Wachsmuth et al., 2017b).

Since stance relevance entails argumentativeness, which in turn entails topic relevance, we refer to these three as "levels" of relevance.

Though previous work focused on stance relevance only (e.g., Stab et al., 2018), an analysis on all three levels provides more insight into the errors made and is especially warranted for "argumentative images." Figure 2 illustrates the different levels for the example query `nuclear energy`, showing images that fail (a) topic-relevance, (b) argumentativeness, and (c) stance-relevance (provided the image is categorized as a con). Though also
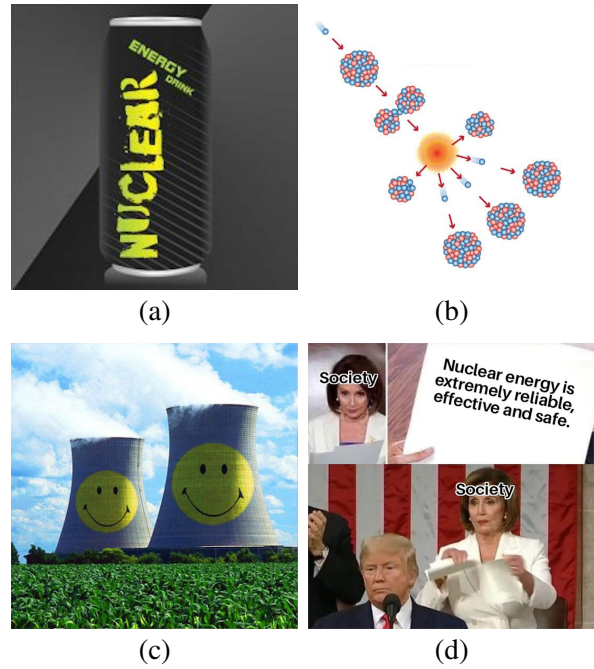

Figure 2: Example results for the query `nuclear energy` that are (a) off-topic, (b) not argumentative, (c) supportive, and (d) with ambiguous stance.

many textual arguments appeal to emotion rather than logic, images are especially suited for such an appeal. However, as the emotions invoked can depend on both the viewer and context, it can be surprisingly unclear whether an image is argumentative or to which stance it is relevant. Consider Figure 2d. The meme image mixes the question of nuclear energy with the emotions towards a certain political moment (US president Trump's 2020 State of the Union Address, where, at its conclusion, the Speaker of the House of Representatives Pelosi tore up its official copy as a symbolic comment on its contents, Stewart, 2020). Depending on the own emotions towards the shown politicians and statement, one can read the image as pro nuclear energy ("society" not caring for facts) or against ("society" ripping up lies). In the case of such images with ambiguous stance, the definition of stance relevance above suggests listing the image both as a pro and a con. In the future, however, additional considerations of argumentative quality (especially in terms of clarity) might suggest to omit such images completely, or to show them separately.

4 Stance-Aware Query Expansion

The basic hypothesis of our approach is that the task of image retrieval for arguments can be tackled effectively by the structure shown in Figure 3a,
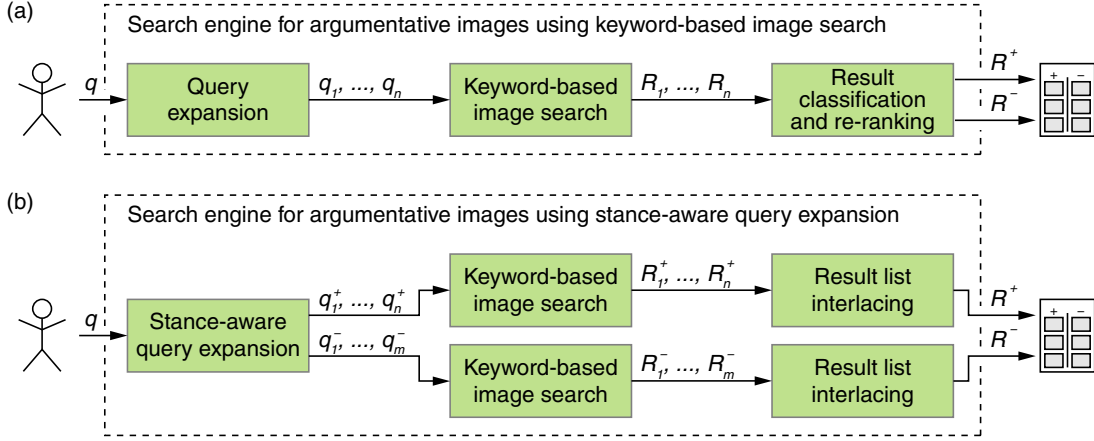
Figure 3: (a) Generic structure of an image search engine for arguments using a keyword-based image search engine and (b) the more specific structure of stance-aware query expansion employed in this paper. In the generic structure, the user's query q is expanded to n queries $q_1, \ldots, q_n$, a result list R_i is retrieved for each q_i, which are classified and re-ranked to create the lists of pro (R^+) and con images (R^-). Structure (b) corresponds to a particular case without classification and re-ranking, where different sets of expanded queries, $q_1^+, \ldots, q_n^+$ and $q_1^-, \ldots, q_m^-$, are used to independently create R^+ and R^-.

which issues for a user's query several expanded queries to a keyword-based image search engine and then fuses the result lists to a list of pro and con images. However, we assume that the semantic capabilities of modern keyword-based image search engines can be harnessed further to already provide a classification of the images, thus separating pro and con images as well as omitting neutral ones.

In particular we propose a stance-aware query expansion as depicted in Figure 3b. It generates focused queries for both relevant stances (superscript + for pro and − for con), which are processed independently of each other, presuming that a sufficient diversity of expanded stance-aware queries recalls relevant images for each stance on the top ranks. In that case, the development of a post-retrieval image stance classifier can be omitted. The result lists of each stance are interlaced, i.e., composed by taking the first result of each list for a stance, then the second of each, and so on.

We devise three methods: (1) appending always the same stance-indicating terms to the user's query, (2) appending sentiment-indicating terms that co-occur with the query's terms, and (3) appending topic-specific stance-indicating terms obtained from a text argument search engine.

(1) Good-Anti Conceivably the single most basic method is to expand the user's query with one term per stance. After some manual experiments, we opted for `good` as a pro term and `anti` as a con term. Another option for the latter was `bad`, but

for some topics this term is more associated with "doing it poorly" than with "being against it."

(2) Positive-Negative This method exploits the fact that stance is often reflected through expressions of sentiment, for example, as used in counter-argument retrieval (Wachsmuth et al., 2018). For each stance, we generate up to five queries by appending the top positive (for pro) or negative terms (for con) of the 8000 entries in the MPQA subjectivity lexicon (Wilson et al., 2005) as ranked by their co-occurrence with the query according to the Leipzig Corpora Collection's English corpus (120 million sentences, Goldhahn et al., 2012).

(3) Pros-Cons This method employs argument search engines to identify terms typical for certain topic-stance-combinations. E.g., in arguments retrieved for `nuclear energy`, "CO2 neutrality" occurs more often in pro arguments than in con ones, whereas "radiation" occurs more often in con arguments than in pro ones. Based on work in anomaly detection (Afgani et al., 2008), this method calculates the specificity of a term t to a stance s, $\delta(t, s)$, as their contribution to the Kullback-Leibler divergence of the term distributions between the two stances:

$$\delta(t, s) = \mathrm{P}(T{=}t|S{=}s) \cdot \log \frac{\mathrm{P}(T{=}t|S{=}s)}{\mathrm{P}(T{=}t|S{\neq}s)},$$

where $P(T = t|S = s)$ is the probability of observing t given the stance s. This probability is

estimated by word frequencies[2] in all arguments that an argument search engine—`args.me`[3] in our case (Wachsmuth et al., 2017b)—retrieves for the query and s. For preprocessing, we lemmatize both arguments and the query. Furthermore, we filter out all arguments from the website *debate.org*, as we found that its debate structure encourages to reference opposing points in arguments to counter them, which diminishes the δ of the respective terms. The method generates up to five queries for each stance by appending the top terms as ranked by δ.

5 Crowdsourcing Relevance Judgements

In retrieval tasks, human relevance judgments of retrieval results for a fixed set of topics allows for evaluating the effectiveness of competing retrieval models. Known as the Cranfield paradigm or TREC-style evaluation (Voorhees, 2001), it is also employed in textual argument retrieval within the Touché shared tasks (Bondarenko et al., 2020). All our expansion methods retrieve results for the same set of queries, which are then pooled and manually judged with respect to their relevance to a given query's information need.

To evaluate the methods presented in Section 4, we employ a sample of 20 controversial questions from the Touché 2020 Task 1 test set (Bondarenko et al., 2020), from which we derive one query each.[4] We calculate the number of relevant images in the methods' ten top-ranked images (precision@10) for both rankings of pro and con images (thus actually for 20 images), which is straightforward to interpret for multiple relevance levels. To ensure state-of-the-art keyword-based retrieval and a large image index, all methods retrieve images using the Google image search.

Figure 4 shows the annotation interface for gathering relevance judgments. As described in Section 3, we distinguish relevance on three levels: topic, argumentativeness, and stance. The first three options, the image being able to support the pro, contra, or both stances, indicate the image stance(s) and that it is argumentative and on topic. The fourth option, the image not being able to support a stance, indicates that the image is on-topic

Figure 4: Annotation interface, showing (from top to bottom) the current topic, question, and image, as well as the generic annotation options and comment box.

only. The final option indicates that the image has no stance and is neither argumentative nor on-topic.

Using the annotation interface of Figure 4, we collected 2988 annotations on 993 images[5] and 20 topics from 12 layperson annotators: three annotators per image and topic. The images' order was randomized for each annotator within a topic to avoid order biases. The annotator agreement is fair to moderate (Fleiss' κ of 0.39; Fleiss, 1981; Figure 7 shows κ per topic). We compute a single label for each image and topic pair as per majority vote (two out of three), treating an annotation *both* as a vote for both *pro* and *con*, and assigning the ground-truth label *both* if there are at least two votes for both *pro* and *con*.[6] To ensure consistency in the face of difficulties reported by the annotators, we reviewed the ground-truth and changed the label for 223 images. The vast majority of changes (85%) were due to vagueness in our instructions: Annotators labeled images that can introduce the topic or question but have no argumentative value otherwise (see Figure 5 for examples) often as being able to support *both* stances though they should have labeled them as non-argumentative (*neither*).

---

[2] To avoid zero probabilities, we employ standard add-one smoothing (Jurafsky and Martin, 2009).

[3] API: `https://www.args.me/api-en.html`

[4] From the original 49 topics, 13 were omitted for which one or more methods found no expansion terms, sampling 20 at random from the remainder to meet our labeling budget. Questions, queries, retrieved images, and annotations are available at `https://doi.org/10.5281/zenodo.5202934`.

[5] Three images were retrieved for two topics.

[6] E.g., if votes are *off-topic*, *neither*, *pro*, the result would be *neither* as two voted for on-topic but only one for a stance.

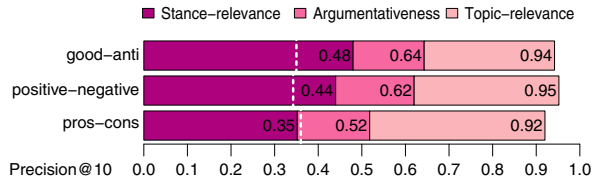Figure 5: Examples of images annotated as being able to support *both* stances but corrected to *neither*.

Figure 6: Precision@10 achieved by the three query expansion methods at three relevance levels, averaged across 20 topics and each topic's individual pro and con rankings. The dashed lines indicate the expected precision@10 of a random image stance classifier dividing the retrieved images into the pro and con rankings. The bars are overlaid, not stacked.

6 Evaluation

This section reports on a comparative evaluation of the three query expansions methods by their retrieval effectiveness, followed by a topic-wise error analysis. Last, we also carried out an investigation into the counterintuitive case of ambiguous images that can support both stances.

6.1 Overall Retrieval Effectiveness

Figure 6 shows the precision@10 for the three query expansion methods introduced in Section 4 with respect to the three levels of relevance assessed. Nearly all retrieved images can be expected to be relevant to the topic (92% to 95%). This being testimony to the effectiveness of Google's image search, it also indicates that the query expansion methods do not impair the keyword-based image search.

The basic good-anti heuristic performs best, achieving an argumentative precision@10 of 0.64 and outperforming the pros-cons method by 0.12. An inspection of the pros-cons method's expanded terms reveals that they lack the stance-specificity of the other methods' terms. For example, the expansions money and person (pros-cons) are less specific than the expansions cheap and harm (positive-negative), as only the latter terms intrinsically convey a preference or stance.

All methods lose 16 to 18 percentage points

when being judged for stance-relevance, showing that 25% (good-anti) to 33% (pros-cons) of the retrieved images that are argumentative do support the opposite stance as intended. While for two of the three methods their stance precision@10 is better than a random stance assignment (dashed lines in Figure 6), it is far from perfect. Since Google image search is not public, it is challenging to pinpoint the source of the stance errors. A possible explanation may be that expansion terms found on an image's web page do not refer to being pro or con a given topic, but are used in a different context or even to convey the opposite stance.

6.2 Topic-wise Error Analysis

We analyzed the retrieval effectiveness on a per-topic basis to learn which are the most challenging ones for retrieving argumentative images. It turns out that some topics are less well-suited to a keyword-based image retrieval (or to our query expansion methods). Moreover, some arguments are not suited to being illustrated or are for other reasons not found through image search. Figure 7 shows the retrieval effectiveness as precision@10 and the image stance distribution per topic.

As the average precision@10 scores show, topic precision is high overall, except for the query standardized tests education (for issue: "Do standardized tests improve education?"). On closer inspection, we found that many off-topic images were either on education or standardized tests, possibly hinting at a lack of images that combine both.

Average argumentativeness precision@10, however, varies a great deal between 0.18 and 0.87. As an inspection of topics with low precision reveals, many of the retrieved images are well-suited to introducing the topic, but not as argument support. For example, most images for body cameras police ("Should body cameras be mandatory for police?") show a police officer wearing a body camera. Similarly, many of the images for performance-enhancing drugs in sports ("Should performance-enhancing drugs be accepted in sports?") show sports equipment and syringes. For queries related to commercial products, like e-cigarettes ("Is vaping with e-cigarettes safe?"), bottled water ("Should bottled water be banned?"), or school uniforms ("Should students have to wear school uniforms?"), many of the images are product photos from shopping or review pages and merely display the prod-
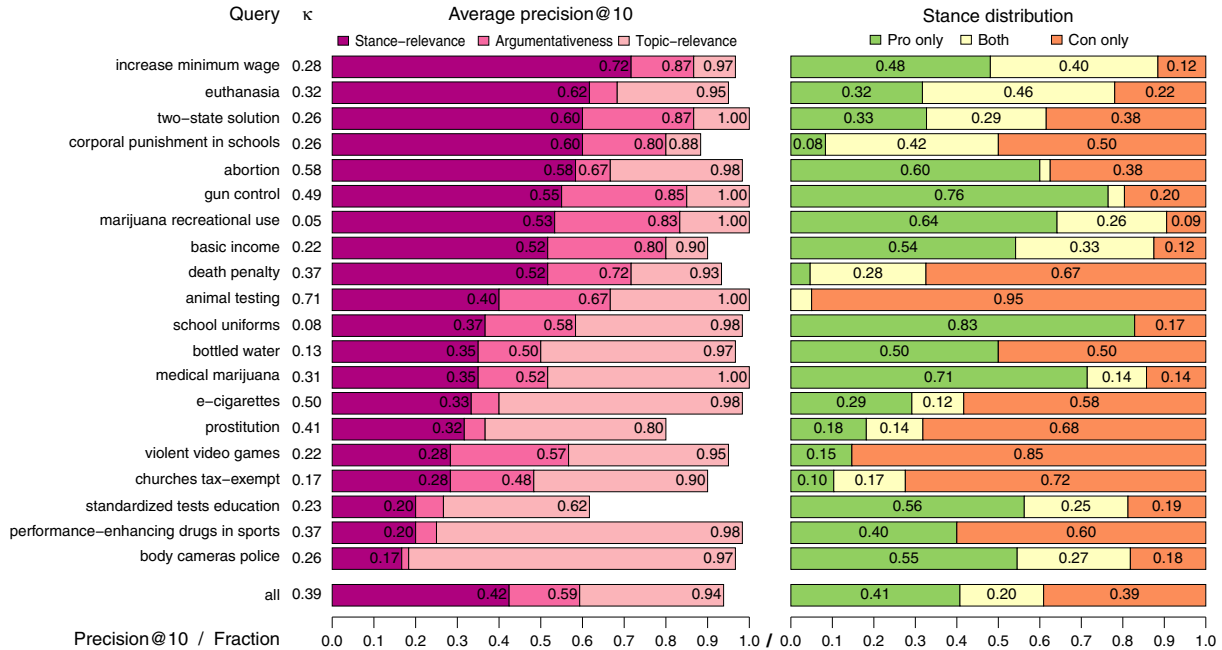
Query	κ	Average precision@10			Stance distribution		
		Stance-relevance	Argumentativeness	Topic-relevance	Pro only	Both	Con only
increase minimum wage	0.28	0.72	0.87	0.97	0.48	0.40	0.12
euthanasia	0.32	0.62		0.95	0.32	0.46	0.22
two–state solution	0.26	0.60	0.87	1.00	0.33	0.29	0.38
corporal punishment in schools	0.26	0.60	0.80	0.88	0.08	0.42	0.50
abortion	0.58	0.58	0.67	0.98	0.60		0.38
gun control	0.49	0.55	0.85	1.00	0.76		0.20
marijuana recreational use	0.05	0.53	0.83	1.00	0.64	0.26	0.09
basic income	0.22	0.52	0.80	0.90	0.54	0.33	0.12
death penalty	0.37	0.52	0.72	0.93		0.28	0.67
animal testing	0.71	0.40	0.67	1.00			0.95
school uniforms	0.08	0.37	0.58	0.98	0.83		0.17
bottled water	0.13	0.35	0.50	0.97	0.50		0.50
medical marijuana	0.31	0.35	0.52	1.00	0.71	0.14	0.14
e–cigarettes	0.50	0.33		0.98	0.29	0.12	0.58
prostitution	0.41	0.32		0.80	0.18	0.14	0.68
violent video games	0.22	0.28	0.57	0.95	0.15		0.85
churches tax–exempt	0.17	0.28	0.48	0.90	0.10	0.17	0.72
standardized tests education	0.23	0.20		0.62	0.56	0.25	0.19
performance–enhancing drugs in sports	0.37	0.20		0.98	0.40		0.60
body cameras police	0.26	0.17		0.97	0.55	0.27	0.18
all	0.39	0.42	0.59	0.94	0.41	0.20	0.39

Precision@10 / Fraction

Figure 7: Topic-wise analysis depicting Fleiss' κ, precision@10 per relevance level averaged across expansion methods and stances, and the stance distribution for the retrieved images as judged by our annotators.

uct. We assume that in these cases the expansion methods cannot sufficiently counter the search engine optimization (SEO) done by companies working in that domain, suggesting a pre-filtering by web document and/or image genre.

Similarly, average stance precision@10 varies a lot between topics, ranging from 0.17 to 0.72. This variation is partly due to the large differences in the stance of the retrieved images, as illustrated in the right plot of Figure 7. For some topics, argumentative precision and stance precision differ by as much as 0.30 (`gun control` and `marijuana recreational use`). Upon inspection, we noticed that the errors in the stance assignment dominantly occur in a single direction per topic. Though our setup intends to retrieve an equal amount of pros and cons per topic, the right plot of Figure 7 shows that the result set is often skewed in one direction. In the extreme, 95% of the retrieved argumentative images for `animal testing` are con only—which is plausible, as illustrating the benefits of animal testing is much more complex than showing animals being treated poorly. On the other hand, an unexpectedly high fraction of images are ambiguous and can be used to support both stances. For example, this is the case for nearly half of the retrieved argumentative images for `euthanasia`. Figure 8c shows one example for the topic.

6.3 On Images with Ambiguous Stance

When 46% of the retrieved images can provide support for both stances, as for `euthanasia`, should an image search engine for arguments separate pro and con images in its search results as in Figure 1? To investigate this question, we inspected all 111 images that were judged ambiguous. Figure 8 exemplifies the image categories we identified. We found that most of the 111 images provide contextual information (78 images, 70%) in the form of geographic comparisons as in Figure 8a, statistics (especially polls), or forecasts. Similar categories include sources on the topic (10 images, 9%) and definitions (11 images, 10%). Images from these three categories (89% total) can strongly support an argumentation, even though different pieces of information they entail might lend support to different stances. Due to the benefits of such images, it thus makes sense to display them separately in a search interface, either on-demand or up front. As outliers, two other images contrast key points for pro and con, supporting mostly deliberation tasks. The remaining images were ambiguous in the way hypothesized before in that their stance changes depending on how one interprets them. Images of this kind will be a challenge to image retrieval for arguments, since their interpretation requires common sense and/or domain knowledge.
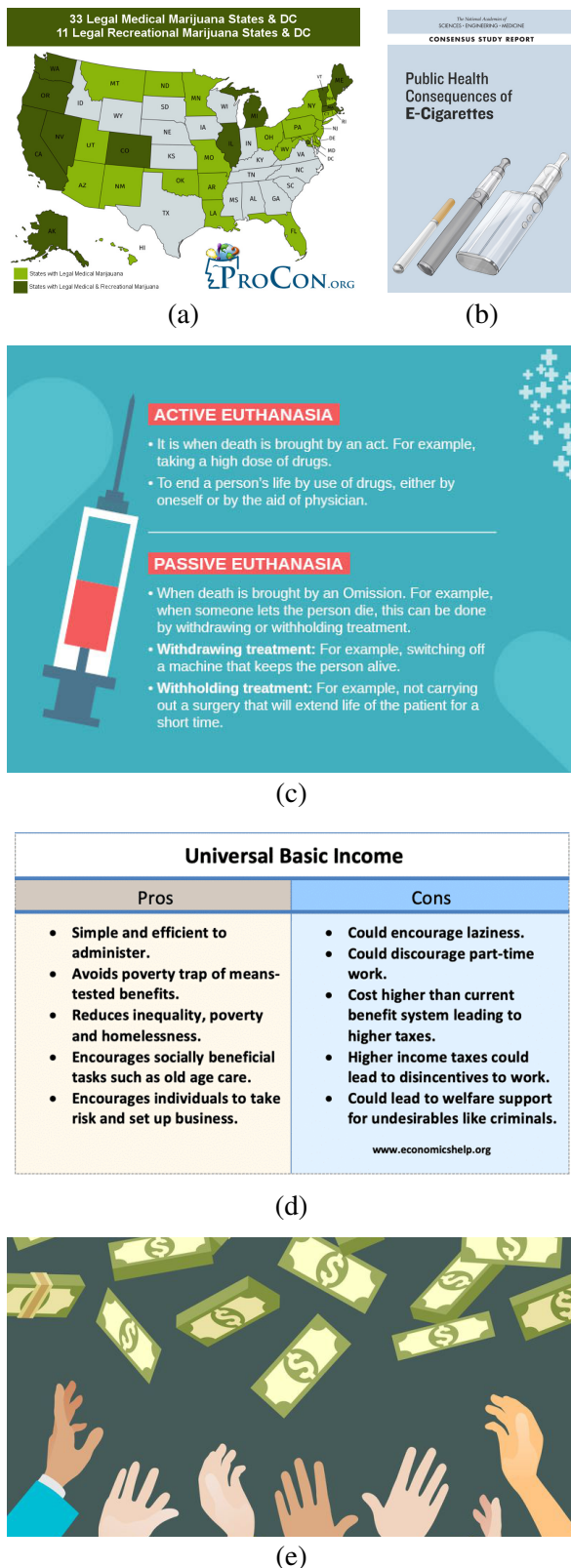
(a)

(b)

(c)

(d)

(e)

Figure 8: Example images representing different categories of images that are able to support both stances: (a) contextual information: 70% of 111 analyzed images, (b) sources: 9%, (c) definitions: 10%, (d) key point tables: 2%, and (e) ambiguous messages: 9%.

7 Conclusion

This paper introduces the task of image retrieval for arguments, proposes with stance-aware query expansion a family of methods to tackle the task that exploits existing keyword-based image search technology, and carries out a first empirical analysis for the task using human annotations on 993 images and 20 topics. In the experiments, the basic good-anti heuristic, which expands a user query with the same two terms (`good` for pro images, `anti` for con ones), outperforms more sophisticated query expansion approaches that employ terms specific to the query's topic. An error analysis unveils a high topic-dependence of the retrieval effectiveness. Finally, the paper investigates ambiguous images labeled as supporting both the pro and con stances. In cases where many neutral images are found, we suggest displaying them separately.

Relying on Google's image search engine ensures a state-of-the-art retrieval model at the price of exact replicability. Since commercial image search engines are subject to frequent changes it must be investigated how reproducible our results are. Developing an open image retrieval system for arguments would enable laboratory evaluation. It would also allow for creating image representations tailored to argumentation, but requires creating and labeling an extensive image collection that covers a manifold of controversial topics.

There is a lot of room to improve precision. Filtering images by analyzing their web pages may avoid ones that merely show a product or otherwise fail to make a point. Further analyses of the web pages with argument mining technologies could provide for a better stance classification of the images. Also images search engine feature to retrieve similar images may be helpful in this regard. Moreover, classifiers to identify persuasion techniques or emotions in images may prove beneficial to identify images for arguments. Assessing the quality of argumentative images could improve the imagined search engine's evaluation and utility to users. To foster research, we run a corresponding shared task to be held as part of the Touché 2022 lab.[7]

This paper seeks to extend argument mining in general and argument retrieval in particular to considering images. Other argument mining tasks (e.g., relation extraction in articles) could similarly be extended to account for the widespread use of images in argumentative texts or speechmaking.

---

[7] https://webis.de/events?q=touche#touche-2022

References

Rebecca Adler-Nissen, Katrine Emilie Andersen, and Lene Hansen. 2020. Images, emotions, and international politics: the death of Alan Kurdi. *Review of International Studies*, 46(1):75–95.

M. Afgani, S. Sinanovic, and H. Haas. 2008. Anomaly detection using the kullback-leibler divergence metric. In *1st International Symposium on Applied Sciences on Biomedical and Communication Technologies (ISABEL 2008)*, pages 1–5. ISSN: 2325-5331.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *1st Workshop on Argument Mining (ArgMining 2014)*, pages 64–68. ACL.

Philippe Aigrain, Hongjiang Zhang, and Dragutin Petkovic. 1996. Content-based representation and retrieval of visual media: A state-of-the-art review. *Multimedia Tools and Applications*, 3(3):179–202.

Khalid Al-Khatib. 2019. *Computational Analysis of Argumentation Strategies*. Dissertation, Bauhaus-Universität Weimar.

Anne Barnard and Karam Shoumali. 2015. Image of drowned syrian, aylan kurdi, 3, brings migrant crisis into focus. The New York Times.

Martin A. Berger. 2010. Fixing images: Civil rights photography and the struggle over representation. *RIHA Journal*, 10.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of touché 2020: Argument retrieval. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*.

Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of touché 2021: Argument retrieval. In *Advances in Information Retrieval. 43rd European Conference on IR Research (ECIR 2021)*, volume 12036 of *Lecture Notes in Computer Science*, pages 574–582, Berlin Heidelberg New York. Springer.

Ning-San Chang and King-sun Fu. 1980. Query-by-pictorial-example. *IEEE Transactions on Software Engineering*, 6(6):519–524.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *15th International Workshop on Semantic Evaluation (SemEval'2021)*, pages 70–98, Online. Association for Computational Linguistics.

Francesco D'Orazio. 2015. Journey of an image: from a beach in Bodrum to twenty million screens across the world. In *The Iconic Image on Social Media: A Rapid Research Response to the Death of Aylan Kurdi*. Visual Social Media Lab.

Ian J. Dove. 2012. On images as evidence and arguments. In Frans H. van Eemeren and Bart Garssen, editors, *Topical Themes in Argumentation Theory: Twenty Exploratory Studies*, Argumentation Library, pages 223–238. Springer Netherlands, Dordrecht.

Finis Dunaway. 2018. Images, emotions, politics. *Modern American History*, 1(3):369–376.

Xénia Farkas and Márton Bene. 2020. Images, politicians, and social media: Patterns and effects of politicians' image-based political communication strategies on social media. *The International Journal of Press/Politics*.

J. L. Fleiss. 1981. The measurement of interrater agreement. In *Statistical methods for rates and proportions*, 2 edition, pages 212–236. John Wiley & Sons, New York.

Bernd Frohmann. 1992. The power of images: A discourse analysis of the cognitive viewpoint. *Journal of Documentation*, 48(4):365–386.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 759–765. European Language Resources Association (ELRA).

Google. 2021. Google images best practices. Google Developers.

Ioana Grancea. 2017. Types of visual arguments. *Argumentum. Journal of the Seminar of Discursive Logic, Argumentation Theory and Rhetoric*, 15(2):16–34.

Benita Heiskanen. 2017. Meme-ing electoral participation. *European journal of American studies*, 12(2).

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 2 edition. Prentice Hall series in artificial intelligence. Prentice Hall.

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*, 2019:21.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument search: Assessing argument relevance. In *42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2019)*. ACM.

Georges Roque. 2012. Visual argumentation: A further reappraisal. In Frans H. van Eemeren and Bart Garssen, editors, *Topical Themes in Argumentation Theory*, volume 22, pages 273–288. Springer Netherlands, Dordrecht. Series Title: Argumentation Library.

Jamshid Shanbehzadeh, Amir-Masoud Eftekhari-Moghadam, and Fariborz Mahmoudi. 2000. Image indexing and retrieval techniques: past, present, and next. In *Storage and Retrieval for Media Databases 2000*, volume 3972 of *SPIE Proceedings*, pages 461–470. SPIE.

Martin Solli and Reiner Lenz. 2011. Color emotions for multi-colored images. *Color Research & Application*, 36(3):210–221.

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 *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2018)*, pages 21–25. ACL.

Emily Stewart. 2020. Why nancy pelosi ripping up some papers has set the internet on fire. Vox.

Ellen M. Voorhees. 2001. The philosophy of information retrieval evaluation. In *Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. Computational argumentation quality assessment in natural language. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 176–187.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In *4th Workshop on Argument Mining (ArgMining 2017)*, pages 49–59, Copenhagen, Denmark. Association for Computational Linguistics.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 241–251. Association for Computational Linguistics.

Marilyn A Walker, Pranav Anand, Jean E Fox Tree, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 812–817, Istanbul.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Weining Wang and Qianhua He. 2008. A survey on emotional semantic image retrieval. In *International Conference on Image Processing (ICIP 2008)*, pages 117–120. IEEE.

Nicholas Watt. 2015. David Cameron says UK will take thousands more Syrian refugees. *The Guardian*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 347–354. ACL.

Heather Suzanne Woods and Leslie Ann Hahner. 2019. *Make America meme again: the rhetoric of the alt-right*, volume 45 of *Frontiers in political communication*. Peter Lang, New York.

Angela Wu. 2020. Learn more about what you see on google images. Google Blog.

Keiji Yanai. 2001. 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.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference (IMC 2018)*, pages 188–202. ACM.