Universität Leipzig
Fakultät für Mathematik und Informatik
Degree Programme B.Sc. Informatik

Query Expansion Approaches for Image Retrieval in Argumentative Contexts

Bachelor's Thesis

Nico Reichenbach
Born May 29, 1996 in Torgau

Matriculation Number 3718793

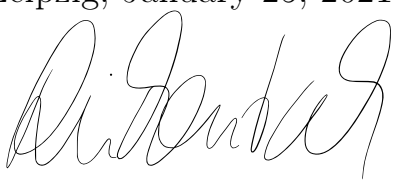
1. Referee: Junior-Prof. Dr. Martin Potthast

Submission date: January 25, 2021

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Leipzig, January 25, 2021



. . . . . .

Nico Reichenbach

Contents

Chapter 1

Introduction

> A picture is worth a thousand words.
>
> — *saying, unknown author*

As this well-known saying vividly states, images are significant components of human communication. Whether vacation snaps or internet memes, images are created, presented and perceived by individuals. Therefore they can be considered as subjective statements. Because visual impressions induce emotions, images can support someone's argumentation.

Imagine you're debating the advantages and disadvantages of nuclear power. If you want to convince your opponent of its risks, it might support you argumentation if you show her images of nuclear reactor disasters or of people who suffer from radiation. In this example, we see that images can help to express and underscore an opinion on a controversial subject. This indicates that they have an argumentative potential and can favor a particular point of view.

As the 2016 presidential election in the United States emphasized, memes and images shared in social media play an important role in public discourse (Woods and Hahner, 2019: p. 1). Thousands of images were shared in social media to express a supporting or an opposing stance on a candidate or his/her political positions. Their widespread use in social media during the election led to the discussion on their impact on the electoral behavior as well as their contribution to an increased interest in the election (Heiskanen, 2017: p. 1).

Because internet memes and images, as well as charts and tables, can be used to make opinions clear, they gain an argumentative character when used in this way. Moreover, images shared online have a participative character since users can modify and distribute them to popularize their stance on a controversial topic (Heiskanen, 2017: p. 4). The 2020 U.S. election shows, images can also contribute to polarizing political debates. So, regarding them when analyzing discourses might help to better understand social processes.

Whether gathering images for research purposes or using them to support the own argumentation, the examples demonstrate that the information need for images that take a stance on a topic exists. When it comes to finding such images, conventional search engines like Google Images only provide little support. Therefore, this work aims to satisfy this information need.

Current research in the area of argument search enables users to find textual arguments on controversial topics. Wachsmuth et al. (2017), for example, presented a framework for indexing, classifying, and retrieving arguments. Using this framework, they created an argument search index by crawling different debate portals. As a proof of concept, they implemented an online argument search engine called *args* based on this index.

Since related work concerning argument search primarily focuses on the analysis and retrieval of text documents, the objective of this thesis is to apply argument search to images in order to close the existing research gap. Because the framework of Wachsmuth et al. (2017) has an extensible structure and its index is one of the largest argument resources currently available, the following paper aims to extend their search engine *args*.

Driven by the primary assumption that polarizing images occur on websites expressing a stance on a topic, the idea is to expand the user query, which is a controversial topic in this case, with stance-expressing terms to find argumentative images. In the course of this work, we present an information retrieval system that implements this idea.

The following chapter outlines related research areas that are relevant to this work. After developing criteria that images must satisfy to be considered as argumentative, chapter 3 introduces a retrieval pipeline for finding such images. With the primary assumption in mind, this retrieval pipeline has a query expansion component. Targeted expansions of the search phrase might allow the system not only to find positive and negative images concerning controversial topics but also distinguish between them. Instead of building a new image index from scratch, the system uses an existing online image search with a high retrieval performance. For this purpose, we will develop a search engine scraper serving as image index.

Chapter 4 then presents three different strategies to gather terms that are positively or negatively related to the query topic. These are the query expansions used to search for argumentative images.

To investigate whether the primary assumption is applicable in general, we will conduct a user study in chapter 5. By comparing the three different query expansion approaches, this chapter will evaluate the retrieval performance of our system. We will also gather arising questions concerning argumentative image search by highlighting this information retrieval domain from different perspectives.

The ambition of this thesis is to develop an extensible and flexible system on which future research can be built up.

Chapter 2

Related Work

2.1 Argument Search

Whether negotiating compromises in conflicts of interest or debating socially important issues, people exchange arguments to justify their own point of view in discussions and to convince others of it. Therefore, argumentation is an essential part of human communication.

As public discussions increasingly take place on the Internet, social media and debate portals gain importance. For this reason, automatic processing of potential arguments in natural language text is becoming more relevant as well (Peldszus and Stede, 2013). Argument search systems aim to support decision making processes as well as giving an overview of debates and searching for arguments to support ones argumentation. This includes both the retrieval of arguments in large document collections and the analysis of their components.

One contribution to this research area is provided by Stab et al. (2018a) with their argument retrieval system ArgumenText[1]. The focus of their work is on the recognition of arguments and their stances in heterogeneous document collections. The data basis of ArgumenText is 400 million text documents from the English part of CommonCrawl[2]. In contrast to the argument search index of Wachsmuth et al. (2017), which relies on the structure of previously crawled debate portals to so identify arguments and their corresponding stance, Stab et al. (2018b) train a neural network to do so. The network learns from manually annotated sentences. ArgumenText's retrieval is implemented with ElasticSearch[3], which ranks documents using BM25. The retrieved arguments are, as well as in *args*, presented in a pro vs. con view.

Walker et al. (2012) presented the Internet Argument Corpus (IAC). This

[1] https://www.argumentsearch.com/
[2] https://commoncrawl.org/
[3] https://www.elastic.co/

corpus contains almost 400.000 posts from an online debate portal. For a subset of this corpus, they manually annotated these posts with argumentative markers like the degree of agreement or the emotionality of argumentation.

In this paper, we develop an image retrieval system that aims to extend Wachsmuth et al. (2017) by an argumentative image search. Because of its extensible structure and its large index, it provides a reliable basis for experimenting with a new approach in the research area of argument retrieval. Therefore, we highlight the authors' work in more detail below.

2.1.1 *args* Argument Search

Since online search engines often do not provide sufficient support when it comes to searching for arguments on controversial topics, Wachsmuth et al. (2017) developed an extensible framework to enable collaborative research on argument retrieval.

The framework's foundation is a common argument model that includes the following components:

- **Argument:** An argument consists of a conclusion and several premises. The premises have a stance towards the conclusion.
- **Argument Context:** The context is the argument's metadata (e.g. origin).

Because of the technologies used, such as Apache UIMA and Lucene, both the framework's indexing and retrieval process are extensible. By using the framework, the authors compiled an initial argument search index by crawling debates from different debate portals. By taking the portal-specific particularities into account, they were able to map the arguments from the debate portals to their argument model. Thus, each argument in this index has a conclusion, one or more premises, and a stance. Later in this work, we will use this index for one of our query expansion approaches as it is a very reliable and a large resource for argumentative texts including stance classification.

As a proof of concept, Wachsmuth et al. (2017) used the framework to develop an argument search engine called *args*[4] that uses their argument search index. *args* allows users to search for arguments on controversial topics presented in a pro vs. con view. The retrieved arguments are ranked using BM25F.

[4]https://www.args.me/index.html

2.2 Images in Discourses

Images are used as a form of argument in many contexts. To elaborate the domain of the image retrieval system we develop in this work more precisely, this section examines the argumentative character of images.

In September 2015, three photographs, taken by the photojournalist Nilüfer Demir near the Turkish city of Bodrum, spread virally in social media, such as Facebook and Twitter. The first image shows a small child lying on the shore with its face to the ground. In the others, a Turkish gendarme first stands in front of the child and then carries its body. The child, the three-year-old boy Alan Kurdi[5] and his family fled from Daesh in Syria trying to reach the Greek coast. He, his mother, and his brother drowned when their boat capsized. The father was the family's only survivor (Barnard and Shoumali, 2015).

In his research report, D'Orazio (2015) addresses the spreading of these photographs in social media, observing a shift in the debate on immigration. By analyzing Twitter posts from the beginning of 2015 to September 2015, he recognizes both a considerable rise of the words *refugees* and *migrants* used in the tweets and that the word *refugees* occurs significantly more frequent than the word *migrants*[6] after the photographs spread globally. Not only social media but also newspapers, as well as NGOs, adopted the image. Heads of states and governments expressed their dismay (Eisenreich, 2015) and even announced political actions (Watt, 2015). The reception of these photographs reveals the power of images to call emotions and their ability to increase the range of discourses or even shift them (Adler-Nissen et al., 2020).

Social media became an essential space for debates because they empower individuals to create and distribute media content (Adjei, 2016). With its more than 2.7 billion active users (Clement, 2020a), Facebook, for example, changes the way how public discourses take place (Abdo, 2018). 351 million images are uploaded daily to Facebook on average (Statista, 2014).

But not only social media debates illustrate the efficacy of images in discourses. Photographs of police violence against peaceful civil rights protesters supported the call for civil rights reforms in the US (Berger, 2010).

This example, together with the imapct of Nilüfer Demir's photographs, highlight that images play an integral role in public discourses. Therefore, social sciences regard images in analyses of discourses and political communication (Maasen et al., 2006; Farkas and Bene, 2020). By considering them as part of linguistic and communicative practice, their recognition in this area helps to reveal relations between verbal, social and political action (Frohmann,

[5]Initally reported as Aylan Kurdi

[6]Migrants decide to leave their homes whereas refugees are forced to leave their homes due to armed conflicts, persecution, or other reasons (UNHCR, 2016).

1992). This underlines both the demand for a search engine that is able to retrieve argumentative images and the importance of research in this area.

A discussion is the exchange of arguments contributing to a decision-making process. The discussion's subject is a claim that the debaters consider to be true or false (Stelzner, 1978). Arguments consist of one or more statements, called premises. The argument's premises refer to the discussion's subject reasoning a conclusion. The debaters justify their stance on the subject with the conclusion. Because images are able to point to subjects of discussions as well (Fegter, 2011), it is reasonable to investigate how they correlate with arguments.

By understanding argumentation only as an exchange of verbal arguments, we exclude images from the wide range of argumentation. To describe the argumentative character of images, it is necessary to extend the definition of arguments. One approach might be to simply consider visual arguments as arguments expressed visually. But if so, what elements of images map the components of arguments? We now outline a conception that we can reasonably use in the following chapters of this work.

As already stated, arguments consist of premises and a conclusion. Many images do not match this structure as they only express one statement. But if images take a stance on a discussion's subject and justify it, they satisfy the definition of an argument (Roque, 2012; Grancea, 2017). Some authors argue that images can have features of arguments but they intend to convince someone from a stance rather than justifying a claim (Roque, 2012). However, we can state that images gain an argumentative character when they refer to a controversially debated topic and they can be used in this context to justify a claim or convince someone of it.

In this section, we used examples to clarify the argumentative character of images. We also outlined a conception of what properties images must have in order to be considered as argumentative. With these insights in mind, in the following chapter we present an information retrieval system with the ambition to find argumentative images.

Chapter 3

Argumentative Image Search

Whether images of demonstrations where protesters hold up banners, caricatures, or internet memes that sarcastically take up a topic, images induce associations and emotions in the viewer. Since they are created and interpreted by individuals, images are subjective statements. They do not reflect an objective reality, but instead "serve as active rhetorical agents" (Dunaway, 2018: p. 1). As part of debates, images can express, underline, or popularize an opinion or even shift the debate on a topic, as we saw in section 2.2. Used in this way, they gain an argumentative character (Dove, 2012).

Be it for research purposes when conducting a discourse analysis or to search for images that support the own argumentation, there is a need to find argumentative images. Though web search, as well as online image search, reached a high level of maturity, they have a lack of retrieving this kind of images. Finding images that express or support an opinion on controversial topics often requires users to narrow their search manually in order to obtain relevant results.

Therefore, the following research process aims to design and implement an argumentative image retrieval system. So, this work applies argument search to images. To do so, we extend the online argument search engine *args* because of its extensible structure. *args* allows users to search for textual arguments for controversial topics offering a pro vs. con view in which the search results are grouped by stance in two columns (Wachsmuth et al., 2017). The extension aims to group the retrieved images by stance as well.

The information need we address with this system is images that viewers perceive as positive or negative concerning a controversial topic. So, the query is a controversial topic as textual input. Relevant results for this query are argumentative images expressing a stance on the queried topic.

Rather than acquiring images to build a new image index from the ground, we use an existing online image search engine with a high retrieval performance,

namely Google Image Search, for this work. Query expansion techniques presented in chapter 4 deliver words, that the extension appends to the query topic. Thus, to focus the search on argumentative images, it expands the user's query by positive and negative terms concerning the queried topic. It then searches Google Images with the positively expanded query on the one hand and the negatively expanded query on the other hand. Eventually, it scrapes the results and presents them in two columns.

After clarifying the concept of *argumentative images*, we present the system's retrieval pipeline along with its components that enables the *args* search engine to search for such images.

3.1 Argumentative Images

Section 2.2 illustrated the argumentative potential of images. But with this knowledge, how can we deduce an objective for the search results of the system that we develop in this chapter?

In Section 2.2, we stated that images have an argumentative character if the image makes a claim on a topic and justifies it or it intends to convince someone of a stance on a topic.

So in summary, the relevance of images our system aims to retrieve is not only determined by the queried topic alone but also by the stance the images take on the topic.

Based on these considerations, we elaborated the following relevance criteria in order to characterize relevant images that satisfy the information need:

1. **Topic Precision**: The image must match the topic.
2. **Argumentative Precision**: Viewers perceive a stance of the images regarding the topic.
3. **Stance Precision**: The viewers' perceived stance fits the stance the image is grouped in.

When assessing images by these criteria, criterion (1) seems to be objective. Images either display something related to the topic or they do not. The criteria (2) and (3) may rather depend on the viewer's attitude towards the topic. There might be pictures where it is unclear whether they actually take a stance on the topic. But if they do, is it a *positive* or *negative* point of view, or *both*? Moreover, the stance that viewers perceive might also depend on the image's usage and context. In chapter 5, we will discuss these question in more detail.

Figure 3.1 lists possible search results for the query *nuclear power* to explain the relevance criteria. The image in the upper left corner obviously does not match the topic, so it violates criterion (1). Though the image in the upper right is related to the topic, it is not argumentative since it only displays nuclear fission schematically. It therefore violates the relevance criterion (2). The images at the bottom are both argumentative because they show demonstrators protesting for (left image) and against (right image) nuclear power. If the our system would group the left image to the contra side and the right to the pro side, both images would satisfy all relevance criteria.

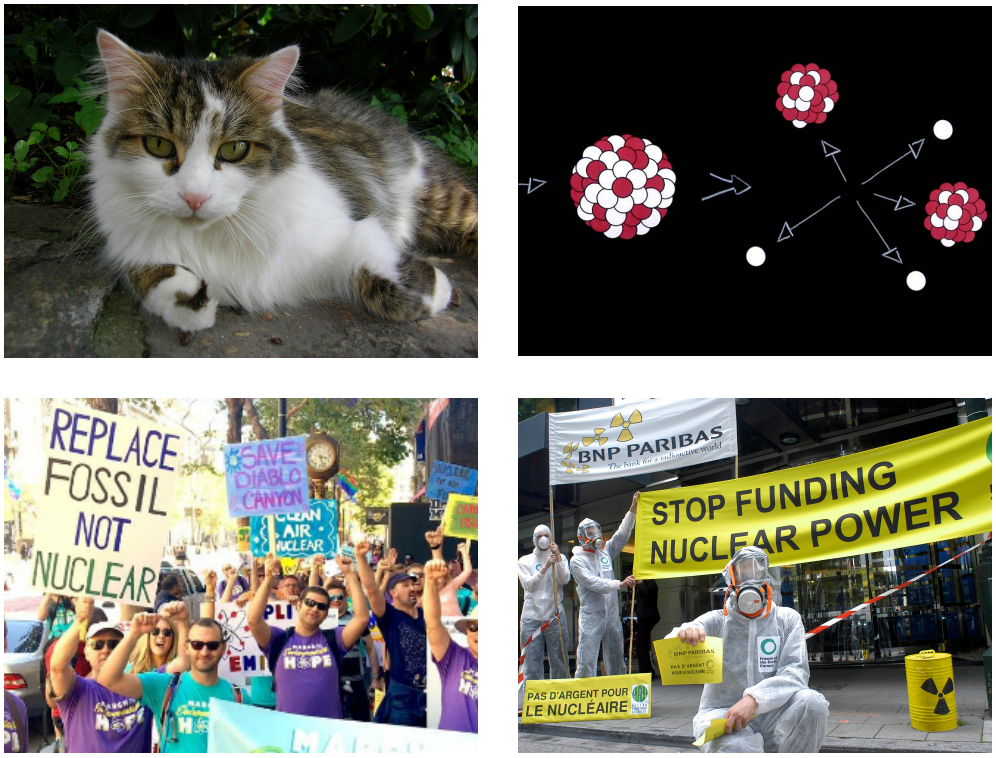


Figure 3.1: Exemplary search results for the topic *nuclear power*. The upper images are not argumentative whereas the images at the bottom are argumentative.

3.2 Retrieval Pipeline

The image retrieval system is created with the assumption in mind that images expressing a stance on a topic occur on websites that favor that same point of view. As mentioned above, the extension uses Google Images as image index. In order to find images that are positive or negative towards a topic, it extends

the user query by terms that are characteristic of the respective stance. So, the retrieval process relies on query expansion. Accordingly, there is one query for positive images and one for negative images. The extension searches Google Images with these two queries and then scraps the search results from Google's result page. The retrieved images are eventually displayed in a pro vs. con view. Figure 3.2 shows the result page of the system[1]. As you can see, it inherits the user interface of *args*.

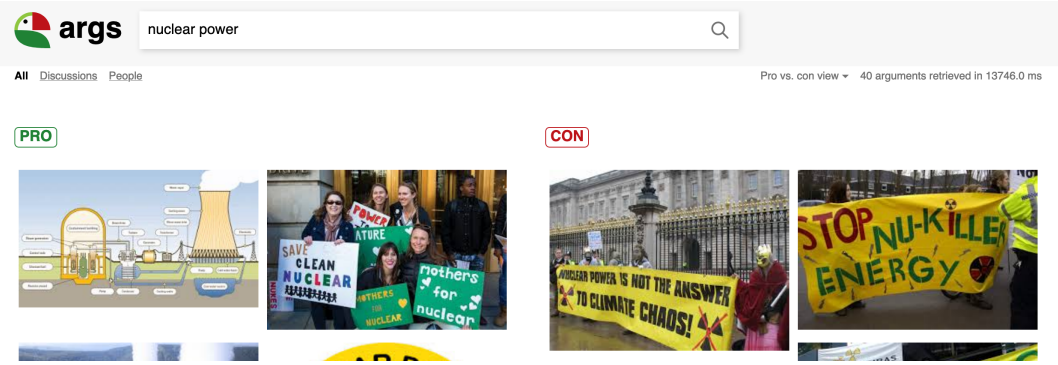


Figure 3.2: Result page of the *args* extension for the query *nuclear power*. As in *args*, the results are displayed in a pro vs. con view.

This section presents all the components of the extension's retrieval pipeline, which can be seen in figure 3.3. All its components are designed as microservices and communicate via REST with each other. This modular structure makes the system extensible, so it can be reused to experiment with new approaches.

3.2.1 Query Expansion

There is only sparse technical documentation on how Google Images works. The article *Google Image best practices*[2] outlines best practices for images' visibility in the Google Image Search from which we can infer how Google might retrieve images. Among others, it states that visual content should be relevant to the website's topic and relevant text should surround the embedded image. Furthermore, images' alt text should contain only a few concise and descriptive keywords. So we can conclude, among computer vision algorithms, Google uses meta information to retrieve images.

[1] A demo of the system can be tested at https://images.args.me/index.html. Note that the search results may differ from those in figure 3.2.

[2] https://support.google.com/webmasters/answer/114016?hl=en, accessed 10-18-2020

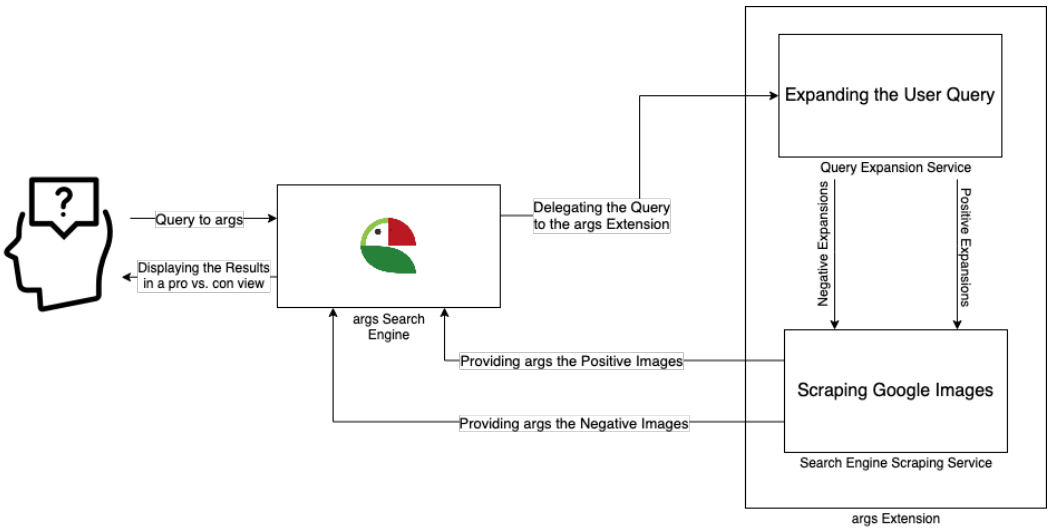


Figure 3.3: The retrieval pipeline of the argumentative image retrieval system interacting with the *args* search engine.

The distribution hypothesis is a very common theory in theoretical linguistics assuming that semantically similar words have a similar linguistic distribution (Boleda, 2020). The distribution hypothesis in other words: words with similar meanings occur in similar contexts. Analogously, we assume that argumentative images occur on websites with argumentative text concerning a topic. This is this paper's primary assumption.

This assumption, together with the information on how Google might retrieve images, leads to the idea to expand the base query by terms used frequently in text expressing a stance on the query topic.

Chapter 4 presents different methods to obtain such concise argumentative terms being associated with a point of view on the topic.

For retrieving argumentative images concerning the query topic, we search Google Images with the expanded search phrase. So, on the one hand, we have a query with the positively expanded search phrase to retrieve images with a positive stance. On the other hand, we have a query with the negatively expanded search phrase to retrieve images with a negative stance.

To implement this, we add a microservice to the extension's retrieval pipeline that returns query expansions for a given topic. Figure 3.4 presents the microservice's endpoint along with its response.

The microservice's structure is generic. Therefore, it can work with various methods to obtain query expansions. Chapter 4 presents three query expansion heuristics.

```
GET /?query=<base-query>&size=<number-of-expansions>
```

```
{
  "baseQuery": <baseQuery>,
  "method": <queryExpansionMethod>,
  "positiveTerms": [<positiveTerm>, ...],
  "negativeTerms": [<negativeTerm>, ...]
}
```

Figure 3.4: The expansion service's endpoint together with its response in JSON format.

3.2.2 Retrieval Model

Based on the fact that we use Google Images as image index, we now make considerations about the system's retrieval model.

The *Probability Ranking Principle* states that the relevance of documents to a given query is dependent on a certain probability. Search results are ranked according to this probability to increase retrieval performance (He, 2009). Retrieval models are heuristics that aim to implement the Probability Ranking Principle by approximating this probability (Stein et al., 2017).

As stated in the previous section, Google Images is a black box. There is nearly no documentation on how Google retrieves and ranks its search results. Further, we cannot adjust the retrieval model, so in this work, we inherit Google Images' retrieval model and ranking.

Nevertheless, Google provides an advanced search feature[3] which we use to influence the image retrieval, however. It offers the possibility to determine a group of words that the websites must contain by putting it into double quotes. Besides, we can define single words, of which the website must include at least one by inserting an *OR* between all of these words. Applied to the retrieval task, this means that the query topic, the base query, is the phrase the website must contain. The query expansions are the terms, of which the website must include at least one.

Figure 3.5 shows an exemplary query to Google Images.

```
(<expansion_1> OR ... OR <expansion_n>) "<base-query>"
```

Figure 3.5: An exemplary expanded query to Google Images that uses Google's advanced search feature.

[3] https://www.google.com/advanced_image_search

3.2.3 Image Index

With its market share of over 80%, Google is the most popular online search engine (Clement, 2020b). Due to its popularity and its average precision of 0.97 for popular topics in image search (Uyar and Karapinar, 2017), the extension uses Google Images as its image index. This section presents a microservice that takes a query topic and query expansions as input to scrape Google Images according to the extension's retrieval model.

Web Scraping

To automatically acquire and process Google's image search results it is necessary to harvest the data directly from Google's website and convert it into a structured format.

Web scraping addresses this task. It is the process of extracting certain desired information of HTML pages and structuring it in a normalized format (Zhao, 2017). From semantic web approaches (Malik and Rizvi, 2011) over data and text mining applications (Sung-min Kim and Young-guk Ha, 2016) to information retrieval tasks (Kurniawati and Triawan, 2017), web scraping is an essential technology to make use of online data.

For the use case of scraping Google Images the desired information is mainly the search results' image URLs. In order to present the results subsequently, also the alt text, the image origin, its rank and its thumbnail are useful to scrape. Figure 3.6 shows how our extension structures the images scraped from Google in JSON format[4].

```
{
  "queryString": <query>,
  "results": [
    {
      "imageUrl": <imageURL>,
      "thumbnailURL": <thumbnailURL>,
      "origin": <originURL>,
      "rank": <rank>
      "altText": <altText>
    },...
  ]
}
```

Figure 3.6: The structure of the scraped images in JSON format.

[4]https://www.json.org/json-en.html

Techniques

There are several techniques to find the information sought in the HTML document. 3 commonly used methods are:

- describing the HTML tags that contain the desired information with **Regular expressions**
- accessing the **HTML DOM** to get the particular HTML tags
- selecting the corresponding HTML tags by using the **XPath** query language[5]

In comparision, the HTML DOM method ist the most time efficient, whereas the XPath method has the smallest memory footprint (Gunawan et al., 2019).

To access the websites containing the desired information, either lightweight command line browsers like cURL[6] or full-featured browsers controlled remotely by frameworks like Selenium[7] are suitable.

Difficulties of Web Scraping

Programs that automatically access websites can send a large number of requests in a brief period of time. This may consume lots of resources with the effect of slowing down the affected server. As a consequence, the operating costs for running the servers increase (Ormiston and Eloff, 2006). When it comes to search engines, scraping applications may affect how a page is ranked in the results[8].

For these reasons many websites try to prevent automated access by complicating it with various obstacles.

- **Request Rate Limitation**
  If the number of requests exceeds a limit in a given amount of time the server blocks the client's requests and answers with HTTP status code 429 Too Many Requests[9].
- **IP blocking**
  In addition to request rate limitation websites can block the client's IP address if it sends too many requests at a time.
  Instead of sending all requests with the same IP address, distributing the

[5]https://www.w3.org/TR/1999/REC-xpath-19991116/
[6]https://curl.haxx.se/
[7]https://www.selenium.dev/
[8]https://support.google.com/webmasters/answer/66357?hl=en, accessed 10-18-2020
[9]https://tools.ietf.org/html/rfc6585

requests by rotating the IP addresses with a proxy switcher munges the client's identity.

- **CAPTCHAs**
  Completely **A**utomated **P**ublic **T**uring tests to tell **C**omputers and **H**umans **A**part are another widely used technique to distinguish whether a computer or a human is accessing a website. Clients sending many requests in a short time need to solve problems like recognizing numbers in a distorted image. Besides graphical CAPTCHAs, also audio and video CAPTCHAs are common.
  Without using additional algorithms that first recognize the problem and then the respective pattern in the image or the audio it is not possible to solve the CAPTCHA (Xu et al., 2020; Moy et al., 2004).

- **Testing the User Agent**
  The user-agent request header in HTTP identifies the client software sending the request[10]. As a result, websites can block supposedly suspicious user agents like cURL.
  Changing the user agent to a commonly used browser like Firefox or Chrome helps to avoid this problem.

- **Frequently Changing Website Structure**
  Another issue scraping applications face is a frequently changing website structure. As stated above, scrapers find the sought information in the corresponding HTML tags. If the DOM structure changes, for example because the relevant HTML tags get different IDs or classes, it is not possible to retrieve the information anymore.
  Such changes require an adjustment of the scraper's source code.

Scraping Service for Google Images

Equipped with this knowledge about web scraping, this section builds a REST service for scraping Google Images.

It has one endpoint, shown in figure 3.7. Remebering figure 3.5, the parameter q is the word or the group of words, which the website must include. Parameter *include* prepresents the terms of which the website must contain at least one.

To scrape the results, a first attempt was to use cURL as user agent to access Google's website. It failed because Google blocks requests sent by cURL

[10]https://developer.mozilla.org/de/docs/Web/HTTP/Headers/User-Agent

```
GET /?q=<query-topic>&include=<list-of-expansions>
```

Parameters:

- *q* - the query topic
- *incluce* - list of query expansions

Figure 3.7: The expansion service's endpoint along with a short description of its parameters.

referring to its terms of services. We then decided to use Google Chrome controlled remotely by Selenium WebDriver[11]. This framework provides an API to send commands to browsers in order to control them. Because the initialization of WebDriver instances takes a while, the scraping service initializes a configurable number of instances on startup. These are then used to perform the query. Before an instance is ready to perform a query, the instance's cache and all cookies are deleted to avoid personalized search results. The processing of a query is performed according to the following steps:

1. access Google Images' website
2. entering the query as described in figure 3.5
3. scraping the respective information out of the HTML tags using XPath
4. shaping the information to the structure as described 3.6

The scraping service returns the image results as JSON structured as in figure 3.6.

With docker-compose[12], we can achieve scalability easily. Each of the scraping service's components runs in a docker container[13]. Docker-compose then allows us to scale up the chrome instances, so the service can process more requests in parallel.

[11]https://www.selenium.dev/documentation/en/webdriver/
[12]https://docs.docker.com/compose/
[13]https://www.docker.com/resources/what-container

Chapter 4

Argumentative Query Expansion

This chapter develops query expansion approaches to expand a user query in the form of a controversial topic in order to find argumentative images.

In addition to their use as design elements to attract the visitor's attention and to loosen up the text, websites embed images to arouse interest and to present additional information. As our brain processes images quickly compared to text (Trafton, 2014), they find use on nearly every website to transport emotions and to convince the visitor of the website's content.

In the context of websites pronouncing an opinion on a topic, it is reasonable to assume that the stance not only reflects in websites' text but also in the embedded images. The assumption put in other words, websites that embed images promoting a stance on a controversial topic also contain text favoring the same stance. As mentioned in section 3.2.1 this assumption is an analogy to the distribution hypothesis.

Textual arguments in general intend to inform readers that the conclusion made is justified or intend to convince them of the conclusion itself (Azar, 1999). When it comes to controverial topics, arguments often intend to convince the audience from a stance on the topic. So in this case, the stance on the topic is the conclusion.

Chapter 3 presented an extension to the online argument search engine *args* aiming to retrieve argumentative images to a given controversial topic. According to the primary assumption that argumentative images occur on websites with argumentative text, the extension expands the query topic by words that are expressive for a stance on this topic. So, the task is now to find terms that are significant for a stance on a topic. These can be sentiment-expressing words, that are frequently used regardless of the topic (e.g. *excellent* vs *poor*), but also words that rather find use in a specific domain to express a stance (e.g. *explosion* vs *co2-neutral* in context of the topic *nuclear power*).

In the following, we present three heuristics addressing this task.

The first one follows a statistical strategy based on information theory and the Kullback-Leibler divergence. Textual arguments retrieved with the *args* index are the data basis to gather such words. The second heuristic returns words from a sentiment dictionary. By using sentence co-occurrence information, it ranks the words in the dictionary. The last method generically appends fix words to the user query.

4.1 Kullback-Leibler Heuristic

Our first method uses the *args* index to find stance-colored words that we can expand the user query with. With more than 290,000 arguments indexed (Wachsmuth et al., 2017), it is a promising resource to find words that express a stance on a topic.

args collects argumentative texts from several online debate portals and therefore allows searching for arguments by topic. Querying *args*' index with a topic will return two sets. One set contains arguments in favor of the topic. The other one contains those against the topic.

If opinions on a subject are exchanged, the language the discussants use to express an argument for a point of view is colored by it. For example, when debating nuclear power, supporters may state that nuclear power is a *co2-neutral* energy source. Opponents, on the other hand, might point out the risks of *radiation* and possible *accidents*. Since our data basis consists of documents that aim to convince other people from a stance on a topic, also the terminology in these documents is biased by the stance. It is therefore consequently to expect that the documents in the *args* index contain words that carry the respective stance that the arguments promotes. We also expect that the use of these terms differs statistically in the document sets.

To summarize, the data basis is natural language text that is classified by stance. By comparing the occurrences of the words in the pro documents with those in the contra documents the method takes advantage of this information to find words that are specific for a point of view on a topic. We develop this method with the expectation that we find noticeable statistical peculiarities that indicate these words.

4.1.1 Acquiring and Preprocessing the Data

args provides a REST-API[1] for retrieving arguments on a topic. The API returns documents with stance information. Because the documents consist of

[1] https://www.args.me/api-en.html

English sentences, it is necessary to segment the sentences into single words (tokens).

While investigating the relationship between the words and the topic, we are interested in the concepts the words represent, not in the concrete inflected realization of them. So, in addition to the tokenization, it is necessary to cast the inflected forms to their dictionary form, or in other words, their lemma. This process is therefore called *lemmatization*.

In this work, we use the Java library *OpenNLP*[2] to deal with these preprocessing tasks, because it contains both a tokenization and a lemmatization API. It also provides pre-trained models[3] for the English language.

After the preprocessing is done, we count the occurrences of each token in both sets to get the corresponding type's frequency in the sets in order to obtain a vocabulary that we can subsequently perform text statistical analysis with.

4.1.2 Preliminary Considerations

So far, we do not know what particular types we are looking for in the arguments. Neither do we know how they relate to the topic. As already stated, the expectation is that stance-colored and polarizing words occur in the respective documents arguing for or against the controversial query topic. The documents in the *args* index are already classified by stance. Therefore, it is helpful to split the vocabulary created in the preprocessing step into two parts. Illustrating the idea with set theory, the splitting results in two overlapping vocabularies:

- A - vocabulary of pro arguments
- B - vocabulary of contra arguments

The vocabulary A contains all types (in particular their lemma) of the pro arguments along with their occurrences in the pro arguments. The vocabulary B contains the types of the contra arguments. Figure 2 visualizes this conception as a Venn diagram.

4.1.3 Kullback-Leibler divergence

To recapitulate, the assumption we follow with this heuristic is that stance-colored words are distributed differently in A and B.

[2]https://opennlp.apache.org/
[3]http://opennlp.sourceforge.net/models-1.5/

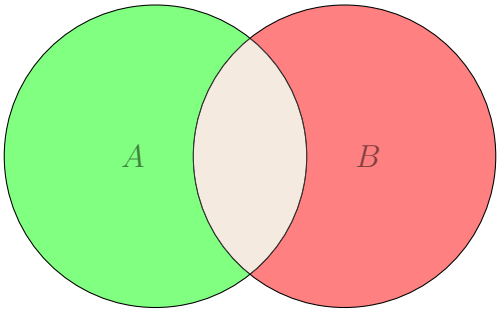


Figure 4.1: The pro vocabulary (A) and contra vocabulary (B) as Venn Diagramm.

In order to determine which terms are particularly significant for one of the two vocabularies, we now compare the probability distributions of A and B. The Kullback-Leibler divergence (D_{KL}), also known as relative entropy (Kullback and Leibler, 1951), determines how two distributions are different from each other. It is therefore a suitable measure for this query expansion approach.

The entropy of a random variable X is the expected information content defined as

$$\begin{aligned} H = \mathbb{E}X &= \sum_{x \in X} P(X = x) * I(x) \\ &= \sum_{x \in X} P(X = x) * log(1 - P(X = x)) \\ &= \sum_{x \in X} P(X = x) * log(\frac{1}{P(X = x)}) \end{aligned}$$

where $I(x)$ is the information content of the event's probability (Shannon, 1948). Let A, B be two random variables. Then, the Kullback-Leibler divergence from A to B is:

$$\begin{aligned} D_{KL}(A||B) &= \sum_{x \in X} P(A = x) * (\ log(P(A = x)) - log(P(B = x))\ ) \\ &= \sum_{x \in X} P(A = x) * log(\frac{P(A = x)}{P(B = x)}) \end{aligned}$$

Note that D_{KL} is not symmetric.

Afgani et al. (2008) presented an algorithm to detect statistical anomalies using the D_{KL}. It estimates the D_{KL} between a reference and an actual distribution and selects those where the divergence exceeds a threshold.

Inspired by this algorithm, the idea is to rank the types in each vocabulary according to their contribution to the divergence. In this context, the

distribution $P(X = x)$ is the relative term frequency of type x in vocabulary X:

$$P(x = X) = \frac{freq_X(x)}{\sum_{x \in X} freq_X(x)}$$

The more a type contributes to the divergence, the more its distribution differs in the two vocabularies A and B. The contributions to the D_{KL} are the sum's addends:

$$\delta_x(A||B) = P(A = x) * log \frac{P(A = x)}{P(B = x)}$$

So, the top-ranked types occur often in one vocabulary but rarely in the other. We interpret those with the greatest δ to be the most decisive for the set of arguments. These are words that our image retrieval system presented in the previous chapter 3 expands the user query with.

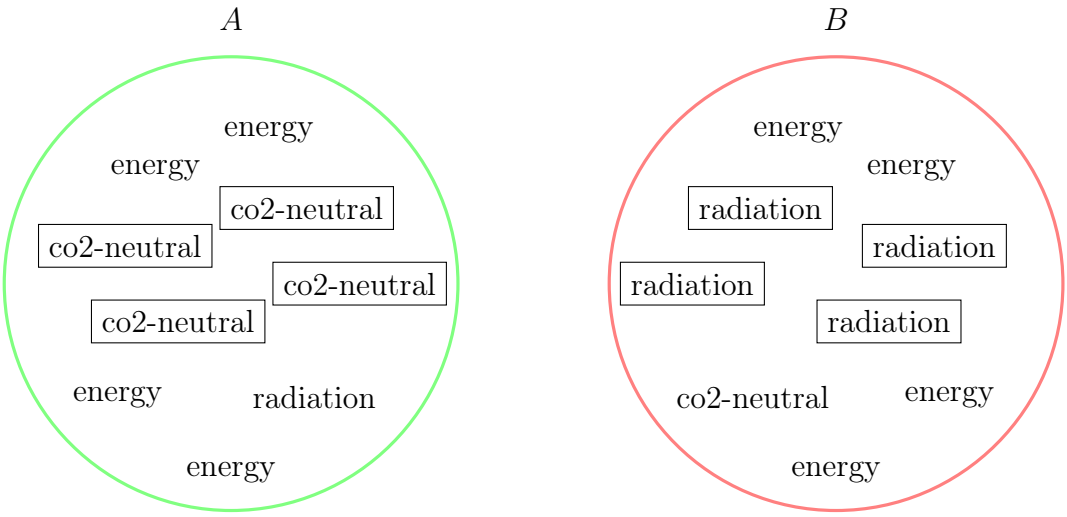


Figure 4.2: Two vocabularies with the types of exemplary pro arguments (A) and exemplary contra arguments (B). The D_{KL} method will choose the framed words.

To illustrate the idea, suppose you have two vocabularies A and B as shown in figure 4.2. A contains the tokens occurring in the pro arguments, B the tokens in the contra arguments for the topic *nuclear power*. As you can see, the type *energy* is distributed equally whereas the occurrences of *co2-neutral* and *radiation* differ.

Table 4.1 illustrates, that this difference is also noticeable in the contribution to D_{KL}. The type *energy* does not contribute to the divergence of the two vocabularies in A and B at all since it occurs in the same number. *Co2-neutral*

has the highest contribution to the divergence from A to B, whereas *radiation* decreases it. Conversely, *radiation* has the highest contribution to the D_{KL} from B to A, and *co2-neutral* decreases the distance. The interpretation of this is that *co2-neutral* is particularly characteristic for the vocabulary A as *radiation* is for B. Thus, the method extends the query nuclear power with the word *co2-neutral* for pro images and *radiation* for contra images.

Table 4.1: The contributions to D_{KL} for the types in vacablulary A and B. This heuristic will choose the terms with the greatest δ.

Type	$\delta(A\|\|B)$	$\delta(B\|\|A)$
energy	0.00	0.00
co2-neutral	0.27	−0.07
radiation	−0.07	0.27

Add-One Smoothing

The δ is calculated by dividing the relative frequencies in the vocabularies with each other and logarithmizing the quotient. As a consequence, neither the divisor nor the divider must be zero. However, if there are no occurrences of a type in either vocabulary, one of the two will be zero. To alleviate this problem, we smooth the relative frequencies as follows:

$$P(x = A) = \frac{freq_A(x) + 1}{\sum_{x \in A} freq_A(x) + \sum_{x \in A \cup B} 1}$$

All in all, the algorithm includes the following steps to generate n query expansions for topic t and for stance s:

1. retrieve the arguments from *args* for topic t
2. tokenize the arguments
3. generate vocabulary A for pro arguments and B for contra argunents
4. lemmatize the tokens in A and B
5. calculate $\delta_x(A \,||\, B)$ and $\delta_x(B \,||\, A)$ for each token in the arguments
6. generate list A' with all tokens ranked by $\delta_x(A \,||\, B)$

7. generate list B' with all tokens ranked by $\delta_x(B \mid\mid A)$
8. if s=PRO: return top n ranked words of A'

   else: return top n ranked words of B'

As mentioned in section 2.1, the *args* index contains arguments from several debate portals. Taking a further look at the portal crawled by *args*, the debate portal *debate.org* stands out with its specific debate structure. In one to five rounds two discussants post their arguments and can rebut the opponent's argument[4]. Due to this debate structure, arguments on this portal are characterized by frequent direct and indirect quotes from the counter-arguments. The debates are conversation-like, so the arguments depend on each other. As a result, words and their frequencies in the pro and contra arguments may overlap. The Kullback-Leibler heuristic we just presented is based on selecting words from the arguments based on statistical peculiarities. Therefore, we exclude this debate portal as a data source in this work.

4.2 Sentiment Dictionary Heuristic

In the last section, we developed a method that uses the Kullback-Leibler divergence to search for sentiment-specific terms in arguments from the *args* index. It aims to find words in textual arguments that express a stance depending on the query topic. In addition to such topic-specific words, more general and topic-independent words can express a point of view or an attitude concerning a topic as well.

Therefore, this section presents a heuristic to expand the query with words having a polarity independent from the topic. It acquires query expansion terms by selecting words from a sentiment dictionary.

Wilson et al. (2005) compiled the MPQA subjectivity lexicon. Along with POS tag information, this sentiment dictionary includes the a priori prior polarity of more than 8000 sentiment-expressing terms.

To select words from this lexicon, we use their sentence co-occurrences with the query topic as ranking. If there are significant co-occurrences of the word and the topic, it is common to use it to comment on that topic.

The Leipzig Corpora Collection (LCC) provides 136 monolingual corpora with text-statistical information (Goldhahn et al., 2012) including sentence co-occurrence information. For this work, we use one if its English corpora with more than 120M sentences to find frequent sentence co-occurrences. This

[4]https://www.debate.org/help/faq/

corpus is available as an SQL database. The following tables of this database are relevant in this context:

- *words* - list of all words the corpus contains
  - w_id - unique identifier
  - word - word form (token)
  - freq - number of occurrences
- *co_s* - sentence co-occurrences
  - w1_id - id of a word
  - w2_id - id of a word
  - freq - number of sentences w_1 and w_2 occur in commonly
- *inv_w* - inverse list
  - w_id - id of a word
  - s_id - id of one sentence the word occurs in

If the search query consists of only one word, the sentence co-occurrences can be determined with the SQL statement in figure 4.3.

```
SELECT freq FROM co_s WHERE
  w1_id = (SELECT w_id FROM words WHERE word = <word1>) AND
  w2_id = (SELECT w_id FROM words WHERE word = <word2>)
```

Figure 4.3: SQL statement to query the LCC corpus for a search phrase consisting of only one word.

But since search phrases may consist of more than one word (e.g. *nuclear power*), the information in the *co_s* table is not sufficient. For this purpose, the table *inv_w* must be used instead. The co-occurrences are then determined with the SQL statement as in figure 4.4.

Both statements return the number of sentences in which the search phrase and the word of the sentiment term occur commonly in. Applying these SQL statements to all words in the sentiment dictionary will create a ranking. Eventually, this method selects those words from MPQA with the highest number of the co-occurrences.

The heuristic presented in this section selects words from the MPQA subjectivity lexicon according to their sentence co-occurrences with the query topic in order to extend the search query with these words. In summary, the following steps are performed to get n query expansion for topic t and for stance s:

```
SELECT COUNT(DISTINCT word_1.s_id)
    FROM inv_w word_1
    JOIN inv_w word_2 ON word_1.s_id = word_2.s_id
    ...
    JOIN inv_w word_n on word_1.s_id = word_n.s_id
    WHERE word_1.w_id =
      (SELECT w_id FROM words WHERE word = <word_1>)
    AND word_2.w_id =
      (SELECT w_id FROM words WHERE word = <word_2>)
    ...
    AND word_n.w_id =
      (SELECT w_id FROM words WHERE word = <word_n>)
```

Figure 4.4: SQL statement to query the LCC corpus for a search phrase consisting of n words.

1. tokenize the search query t
2. for each word in the sentiment dictionary:
   - determine number of sentence co-occurrences with the tokens of t by using the SQL statement in figure 4.4
3. if s=PRO: generate a list L containing all positive dictionary entries ranked by their sentence co-occurrences with their query topic

   else: generate a list L containing all negative dictionary entries ranked by their sentence co-occurrences with their query topic
4. return top n ranked words of L

4.3 Good-Anti Heuristic

Apart from the two heuristics just presented, in this work we also try searching for argumentative images using a rather simple method.

As mentioned in the last chapter, Google Images has a very high retrieval performance. Among other things, this is because Google itself also expands the user query. For this reason, the approach here is to append the word *good* to search queries for obtaining positive images and the word *anti* to search for negative images, because these are very polarizing words in terms of expressing agreement or rejection. Since Google Images has an average precision of > 0.9, we expect this heuristic to achieve a constantly high precision across all query topics. Figure 4.5 shows the two search queries for the topic *nuclear power*.

```
good "nuclear power"
anti "nuclear power"
```

Figure 4.5: Two search queries for the topic *nuclear power* the words *good* and *anti* appended.

To evaluate whether the three methods are suitable for finding argumentative images, the next chapter analyzes image search results for preselected topics in a user study.

Chapter 5

Evaluation of Retrieval Performance

First, Chapter 3 introduced an extension for the *args* search engine. Its goal is to find argumentative images on controversial topics. In doing so, it expands the query with sentiment and stance expressing terms and then searches Google Images with the expanded query. So, there is one query for positive and one query for negative images. Chapter 4 presented three heuristics that generate such query expansions. This approach is inspired by the assumption that images favoring a stance on a topic are embedded on websites that contain text arguing for the same point of view.

In this chapter, we will evaluate whether the approach of finding argumentative images with query expansion works. Furthermore, we will compare the retrieval perfomance of the presented heuristics. The chapter's objective is also to gather questions concerning argumentative image search in general that arise during the evaluation.

To do so, we describe, conduct and evaluate a user study in the course of this chapter.

5.1 Structure of the User Study

The study is conducted with 12 subjects. Because of the scope of this work, the participants were not randomly selected but came from the same milieu and the same age group.

For 20 sampled topics, positive and negative images are gathered with the our image retrieval system using the three heuristics from chapter 4. For each topic and method, the participants judge the 10 top-ranked search results. The images are displayed in a web application where the subjects are asked to assign a stance to the images.

After explaining the research questions and the selected topics, in this section, we describe the implementation of the study.

5.1.1 Research Questions

Section 3.1 defined three relevance criteria for argumentative images:

1. **Topic Precision**: The image must match the topic.
2. **Argumentative Precision**: The image is argumentative concerning the topic.
3. **Stance Precision**: The viewers' perceived stance matches the stance the extension grouped it in.

According to these criteria, the participants assess images in order to get an overview of how well the approach of our system performs. So, research question (1) is about finding out how many relevant images our system retrieves by using the presented heuristics. To achieve this, the participants can choose one of the following annotation classes to asses an image:

- *Pro* - The image favors a positive stance on the topic.
- *Con* - The image favors a positive stance on the topic.
- *Both* - The image might favor both a positive and a negative stance on the topic.
- *Neither* - The image matches the topic but does not favor any stance.
- *Off Topic* - The image does not match the topic.

Section 5.3 determines the heuristics' precision for each relevance criterion.

Since the participants rate the images based on their subjective perceptions, measuring how much the participants agree in their judgments might help to better understand the subjective character of argumentative images and whether their assessment by stance is unambiguous. The user agreement is therefore research question (2). By highlighting the users' judgments from different perspectives, section 5.2 provides an overview to reveal specifics in argumentative image search and gathers arising questions.

5.1.2 Topics

For this evaluation, we will reuse already elaborated topics from the Touché Argument Retrieval Lab. The Touché Lab was organized in the context of CLEF2020 and aims to foster research related to argument retrieval (Bondarenko et al., 2020). It includes two argument retrieval tasks, which were addressed by 17 teams using different approaches. The retrieval tasks are listed in figure 5.1 below.

1. "Given a focused collection of arguments and some socially important and controversial topic, retrieve arguments that could help an individual forming an opinion on the topic, or arguments that support/challenge their existing stance."

2. "Given a generic web crawl and a comparative question relating to a personal decision, retrieve documents with arguments that could help an individual to arrive at a conclusion regarding their decision."

Figure 5.1: The retrieval tasks of the Touché Lab (Bondarenko et al., 2020).

We chose the 49 topics of the first retrieval task[1], as it fits best to our retrieval task of finding argumentative images on controversial topics. In order to have reasonable search phrases, we first tagged the topics before we gather the images with our heuristics. The tags are listed in table A.2[2].

The Good-Anti heuristic generically appends the words *good* and *anti* to the search phrase, so it provides query expansions for every topic. The Kullback-Leibler and the Sentiment-Dictionary heuristics did not provide query expansions for all topics. For 13 of the 49 topics, at least one heuristic found no pro or contra expansions. From the remaining 36 topics, we randomly sampled 20 of them (listed in A.1) to limit the time effort for the subjects.

5.1.3 Study Implementation

The group is split into four subgroups with three people. So, each group annotates five of the 20 topics so that every image is assessed by 3 persons. Both the 10 top-ranked pro and contra images that our system retrieved by using the three heuristics will be annotated by the participants. There is only one vote per participant per image. Figure 5.2 summarizes the composition of the images.

[1] https://webis.de/events/touche-20/topics-task-1.xml

[2] Due to their large scale, you find the tables A.1 to A.12 in the appendix A.

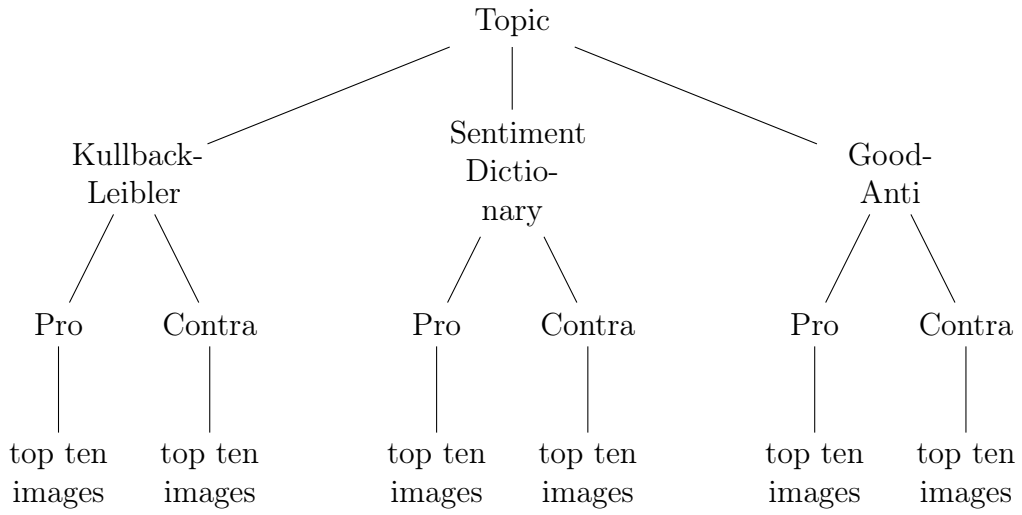


Figure 5.2: The composition of the retrieved images that the participants will annotate.

The images are displayed to the participants grouped by topic. In order to obtain the participants' subjective perception of the images, the only information they have, is the topic that the image belongs to. Furthermore, the images' order is randomized in terms of stance, rank, and heuristic with the Fisher-Yates shuffle[3] (Fisher and Yates, 1963) to avoid any other bias. Moreover, duplicates are removed.

Mapping the Annotation Classes

In order to answer the research questions from 5.1, the user annotations of the images have to be assigned to the annotation classes. This is the users' perceived stance of the image. We decided for the following mapping:

- *Pro*:
  - more than 2 votes for *Pro* **or**
  - 1 vote for *Pro*, 1 vote for *Both* and 0 votes for *Con*
- *Con*:
  - more than 2 votes for *Con* **or**
  - 1 vote for *Con*, 1 vote for *Both* and 0 votes for *Pro*

[3]The Fisher-Yates shuffle is a commonly used algorithm to randomly order a sequence of objects.

- *Both*:
  - more than 2 votes for *Both* **or**
  - 1 vote for *Pro* and 1 vote for *Con*
- *Neither*:
  - more than 2 votes for *Neither*
- *Off Topic*:
  - more than 2 votes for *Off Topic*

If none of these conditions are satisfied, we consider the image to be completely irrelevant. This is equivalent to the annotation class *Off Topic* since none of the relevance criteria are met if the image does not fit the topic.

Annotation Interface

To conduct the study, we developed an annotation tool that allows participants to judge the images. The tool is a static web application developed with the Javascript frontend framework Vue.js[4]. For the infrastructure of the tool, we use two AWS S3 buckets[5]. One bucket is responsible for hosting the website, while the other bucket stores user information, the image annotations of the participants, and the images. Since this is a static website, the business logic runs on the client. Figure 5.3 shows the schematic structure of the infrastructure.

The user interface, shown in Figure 5.4, has a simple and clear design. When users enter the tool, they need to enter their user ID first. Then, they are directed to the annotation interface, where the current image for assessment is displayed. Also the image's topic is shown to the users. The annotation classes are listed below the image. To submit an annotation, an annotation class must be selected.

5.2 User Agreement

As already explained in section 3.1, the perceived stance of an argumentative image may depend on the context of its use, and the viewer's perception. Therefore, the assessments made by the participants in the study are strongly subjective. So, before evaluating the system's retrieval performance,

[4]https://vuejs.org/
[5]https://aws.amazon.com/s3/

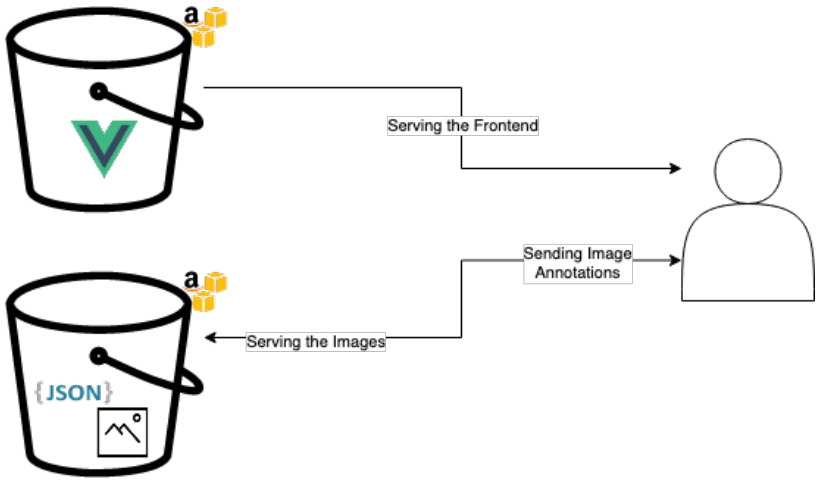


Figure 5.3: The infrastructure of the annotation tool implemented with AWS S3 buckets.

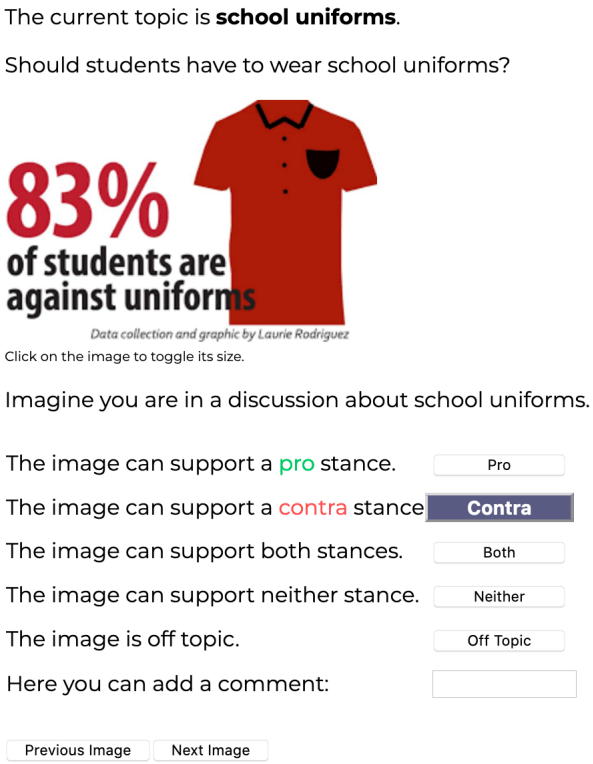


Figure 5.4: The user interface of the annotation tool.

we investigate whether the assessment of argumentative images by stance is unambiguous in order to understand their subjective character. Addressing research question (2), this section determines the user agreement concerning the relevance criteria.

Cohen's Kappa (Cohen, 1960) is a common and suitable measure for determining the user agreement of nominal scaled data. In this study, more than two annotators assess the images. Since this statistical measure can only represent the user agreement of two raters, we use Fleiss' Kappa (Fleiss, 1981), an extension of Cohen's Kappa, instead. The measure of agreement is mapped to a number κ, where

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

with

- p_0 - relative agreement
- p_e - probability of random agreement

The value of κ can be interpreted as follows (Landis and Koch, 1977):

1. $\kappa < 0$: poor agreement
2. $0 \leq \kappa \leq 0.2$: slight agreement
3. $0.2 < \kappa \leq 0.4$: fair agreement
4. $0.4 < \kappa \leq 0.6$: moderate agreement
5. $0.6 < \kappa \leq 0.8$: substantial agreement
6. $0.8 < \kappa \leq 1$: almost perfect agreement

For this review, the judgments were summarized in the following way:

- **Agreement if the topic matches**:
  $\{\textit{Off Topic}\}$ vs. $\{\textit{Pro} \cup \textit{Con} \cup \textit{Both} \cup \textit{Neither}\}$
- **Agreement whether the image is argumentative**:
  $\{\textit{Off Topic} \cup \textit{Neither}\}$ vs. $\{\textit{Pro} \cup \textit{Con} \cup \textit{Both}\}$
- **Agreement concerning the annotation class**:
  $\{\textit{Off Topic}\}$ vs. $\{\textit{Neither}\}$ vs. $\{\textit{Both}\}$ vs. $\{\textit{Pro}\}$ vs. $\{\textit{Con}\}$

Table 5.1 shows the participants' agreement on relevance criteria (1) and (2) as well as the agreement concerning the annotation class across all topics. As you can see, there is only fair agreement if the images match the topic. This is a surprising result because section 3.1 considered this criterion to be rather objective. Since the probability of random agreement is very high and the κ-statistic penalizes disagreement, this uncertainty may be due to the small group of participants. Although clearly described in a user guide handed out to the participants before the study started, another explanation for this might be that the annotators understood the question concerning criterion (1) differently. For example, if the topic is "Should students have to wear school uniforms?" and an image shows only a school uniform, there could be disagreement about whether the image fits the topic since it does not refer to the controversy of the topic.

There is also a high level of disagreement among the participants about whether an image is argumentative or not. With this low κ, it seems that the question about an image's argumentative character is rather difficult to answer and very subjective.

The agreement concerning the annotation classes is moderate. Though it is slightly higher compared to the other categories, the participants' assessments are characterized by disagreement as well. This highlights the subjective character of images, which we already emphasized at several points in this work.

Table 5.1: κ concerning relevance criteria (1) and (2) and concerning the annotation class across all topics.

User Agreement Category	κ Across All Topics
Agreement if the topic matches	0.368
Agreement whether the image is argumentative	0.216
Agreement concerning the annotation class	0.418

Table A.3 lists the user agreement regarding the annotation classes by topic. It reveals large differences in the user agreement among the different topics. This discontinuity suggests that the use of argumentative images is more appropriate for some topics than it is for others. Also, it becomes clear that assessing argumentative images is not an easy task. This also corresponds with the feedback the participants gave after this user study.

In table 5.2, we see three topics with moderate and substantial user agreement along with the percentage of pro and contra annotations. Noticing the participants judging the images in favor of a specific stance, this conspicu-

ousness raises the question of whether the annotators evaluated the images of these topics in a biased way. Future user studies could query the participants' attitudes to investigate this relationship in more detail. In addition to the participants' bias, another reason could be related to the topics themselves. It might be more difficult to create argumentative images for a certain stance on these topics.

Table 5.2: κ for three topics with moderate and substantial user agreement. The *Pro* and the *Contra* votes include the votes for *Both*.

Topic	κ	*Pro* votes	*Contra* votes
Abortion	0.595	63%	47%
Animal Testing	0.707	39%	96%
Gun Control	0.506	70%	31%

5.3 Retrieval Performance

The image retrieval system we developed in this work is based on the assumption that argumentative images occur on websites with argumentative text. It therefore uses query expansions to find images taking a stance on a controversial topic. Highlighting research question (1), this section provides an overview of the heuristic's precision concerning the relevance criteria in order to see how well the system's approach performs. To determine the precision concerning the relevance criteria, the participant's assessments are mapped according to section 5.1.3.

An information retrieval system's precision is the proportion of retrieved relevant documents among all retrieved documents (Kent et al., 1955):

$$precision = \frac{|relevant\ documents \cap retrieved\ documents|}{|retrieved\ docuemnts|}$$

5.3.1 Topic Precision

First of all, we determine our system's precision regarding relevance criterion (1). So, in this examination, relevant documents are those that have not been assigned to the annotation class *Off Topic*. Accordingly, all images with the annotation class *Pro*, *Con*, *Both* and *Neither* are relevant.

In table 5.3 you see the topic precision across all topics. All heuristics achieve an average precision of more than 0.9. Remembering section 3.2.2, this demonstrates that the use of Google's advanced image search feature works with our approach.

Tables A.4, A.5, and A.6 list the topic precision for each heuristic for all topics. As the table shows, the topic precision is similar across all topics and all heuristics.

Table 5.3: The topic precision across all topics.

Heuristic	Precision
Kullback-Leibler	0.9200
Sentiment Dictionary	0.9550
Good-Anti	0.9425

5.3.2 Argumentative Precision

The following analysis refers to relevance criterion (2). So, we now determine the proportion of argumentative images. Relevant documents in this case are those with the annotation class *Pro*, *Con* and *Both*.

Table 5.4 shows the heuristics' argumentative precision. The participants assessed more than 83% of the images retrieved with the Sentiment Dictionary and Good-Anti heuristic as argumentative. The Kullback-Leibler heuristic provided relevant images in 75% of the cases. Although the participants rather disagreed on this question, they perceived a significant majority of the images as argumentative. This level of precision shows that both the approach of argumentative query expansion and the heuristics are suitable for this retrieval task.

The tables A.7, A.8, and A.9 provide a breakdown of the argumentative precision by topic. It reveals noticeable differences of precision across the topics. While the finding argumentative images works very well across all heuristics for some topics, the precision is rather low for others. Table 5.5. chooses three topics with low precision as example. For the topic *Bottled Water*, we see that the Kullback-Leibler and Sentiment Dictionary heuristics retrieve less than 50% of relevant documents. The Good-Anti heuristic, on the other hand, retrieves still 90%. For the other topics in the table, the precision across all heuristics is lower than the average in Table 5.3. These differences in the precision once again support the hypothesis that it is rather uncommon to

use argumentative images for certain topics. This impression is also underlined by the low precision of the Good-Anti heuristic for some topics. Because of Google Images' high precision, we expected this method to achieve a consistent precision across all topics.

Table 5.4: The argumentative precision across all topics.

Heuristic	Precision
Kullback-Leibler	0.7625
Sentiment Dictionary	0.8350
Good-Anti	0.8675

Table 5.5: The argumentative precision for three topics with low precision.

Topic	Kullback-Leibler	Sentiment Dictionary	Good-Anti
Bottled Water	0.45	0.35	0.90
E-Cigarettes	0.35	0.60	0.45
Standard Tests	0.50	0.50	0.65

5.3.3 Stance Precision

Now, the precision of the heuristics for the relevance criterion (3) remains to be determined. Relevant documents are now those where the perceived stance matches the inferred stance. This also includes images that have the annotation class Both.

Table 5.6 shows the stance precision of the three heuristics across all topics. The Kullback-Leibler and the Sentiment Dictionary method have an average precision of > 0.65, where the Good-Anti heuristic's is still at ≈ 0.7. The proportion of relevant documents confirms that our system can already be a useful tool for finding stance-specific images on controversial topics.

As the tables A.10, A.11, and A.12 illustrate, there are large differences concerning relevance criterion (3) among the topics as well as in the previous section. Therefore, we can conclude that argumentative image search is better applicable to some topics than it is for others. The data we collected provide

Table 5.6: The stance precision across all topics.

Heuristic	Precision
Kullback-Leibler	0.6175
Sentiment Dictionary	0.6475
Good-Anti	0.7075

no clear explanation for this. Investigating this question in greater depth is an inspiration for future research.

Table 5.7 shows the proportion of *Pro*, *Contra*, and *Both* votes in documents relevant concerning criterion (3). We see that the relevant images of all three heuristics were rated most frequently with *Both*. This supports the assumption from section 3.1 that an image's stance towards a topic also depends on the image's context.

Table 5.7: The percentage of the annotation class assessments of documents relevant to relevance criterion (3) across all topics.

Heuristic	*Pro*	*Con*	*Both*
Kullback-Leibler	13.50%	9.00%	39.25%
Sentiment Dictionary	16.25%	10.75%	37.75%
Good-Anti	15.75%	20.50%	34.50%

In summary, we can state that the presented heuristics are suitable not only to find argumentative images but also to distinguish their stance. Figure 5.5 presents a final overview of the heuristic's precision with respect to the relevance criteria. Considering the precision, we can conclude that the primary assumption of our system is a solid foundation to do further research in this area. The evaluation of the user agreement highlighted the subjectivity of argumentative images. It revealed questions concerning the impact of personal bias when assessing images and indicated that argumentative image search might not apply to every topic.

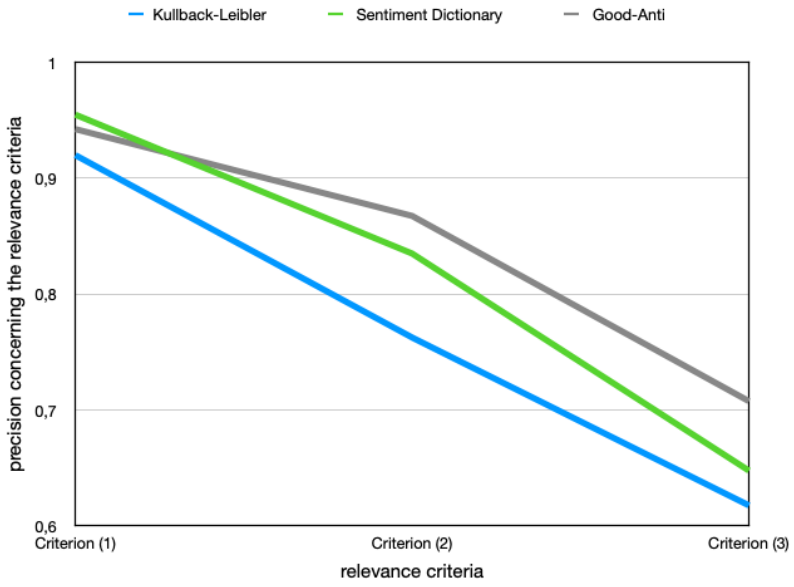


Figure 5.5: Comaprison of the heuristic's precision concerning the relevance criteria.

Chapter 6

Conclusion and Future Work

This paper introduced a new information retrieval system that uses query expansion approaches for finding argumentative images on controversial topics online. The system intends to extend the argument search engine *args* with an image search. Its retrieval pipeline is inspired by the primary assumption that stance-colored images occur on websites expressing a point of view on a topic. Hence, the system uses query expansions to find relevant results. In this work, three different approaches for argumentative query expansion were developed. Besides one approach that provides generic terms, two other methods were presented that generate query expansions depending on the topic. The terms obtained with these three heuristics allow a targeted expansion of the search phrase. This enables the system not only to find positive and negative images on controversial topics but also to distinguish between them. To have a precise objective concerning the search results, we elaborated relevance criteria for argumentative images. A popular online image search with high precision serves as image index. To harvest the search results from it, a search engine scraper was engineered. Due to the system's microservice architecture, the implementation is modular. This structure allows all components to be modified. Because of its extensible structure, the system can be used as a framework for future work in order to experiment with new approaches in this area.

To measure the retrieval performance, we conducted a user study with 12 subjects. For 20 topics, we gathered search results with our image retrieval system using the three heuristics presented in this thesis. The participants annotated the images for the topics with respect to the images' stance that they perceive. The evaluation of this study revealed that, depending on the heuristic, 75% to 85% of the images were perceived as argumentative. In 60% to 70% of all cases, the stance perceived by the participants matched the expected stance. Considering this level of precision, we can make three conclusions:

(1) The primary assumption is applicable for finding argumentative images. (2) The query expansion heuristics are suitable to retrieve relevant argumentative images. (3) The targeted query expansion enables the system to assign an appropriate stance to the images.

By evaluating the user agreement with Fleiss' kappa, the results of the user study emphasize: images are subjective statements. The disagreement of the participants also indicates that the assessment of images by stance is not a trivial task. This claim is also supported by the participants' feedback after the user study. Furthermore, the evaluation of the user agreement raised the question of whether the participants assessed images for certain topics in a biased way. Future user studies in this information retrieval domain could query the subject's attitude towards the topics in advance in order to examine this relationship in more detail.

The participants perceived the majority of argumentative images as bipolar (annotated with *Both*). This suggests that an image's stance towards a topic also depends on the image's context.

Although the heuristics assign an appropriate stance in most cases, the stance-precision between 0.6 and 0.7 encourages to improve these methods:

Kullback-Leibler Heuristic

The heuristic that selects words from the *args* index based on the Kullback-Leibler divergence excludes the debate portal *debate.org* because of the often occurring quotation of opposing arguments. If quotes were detected and filtered, the debate portal would enrich the data basis of this heuristic. Remembering section 5.1.2, the method did not find extensions for several topics. So, including *debate.org* could improve both the query extensions and deliver expansions for more topics. Moreover, other data sources could be considered. ArgumenText, for example, also provides an API[1] to query for arguments. As well as the *args* index, it provides stance information.

Considering the query expansions that this method provided, it is noticeable that the Kullback-Leibler divergence occasionally selects infrequent terms. As the ranking of this heuristic is only determined by the D_{KL}, a modified ranking, in which the term frequency has an increased influence, could achieve better results. Also, a sentiment analytical estimation of the words in the arguments might help to improve the query expansions.

Sentiment Dictionary Heuristic

This heuristic selects words from a Sentiment Dictionary based on their sentence co-occurrences with the query topic. It obtains the sentence co-occurrence information from an English corpus of the Leipzig Corpora Collection (LCC). In these corpora, the sentence co-occurrences always refer to concrete types,

[1] https://api.argumentsearch.com/en/doc

but not to their lemmas. As query expansion terms are repeating across the topics, we could take sentence co-occurrences of the lemmas into account instead. Although the corpus we used is with more than 120 M sentences very broad, the selection of a more specific corpus like the IAC presented by Walker et al. (2012) could improve the results.

The image retrieval system developed in this paper is a helpful tool to satisfy the information need for argumentative images. Considering the system's precision, we can conclude that the query expansion methods presented in this paper are suitable to retrieve relevant results. Moreover, the elaborated relevance criteria can assist future research in this information retrieval domain. By applying argument search to images, this paper contributes to current research in this research area.

Appendix A

Evaluation Appendix

Table A.1: The 20 randomly sampled Touché Topics for the user study.

Topic ID	Tagged Title
2	E-Cigarettes
4	Corporal Punishment in Schools
8	Abortion
9	School Uniforms
11	Performance-Enhancing Drugs Sports
15	Animal Testing
17	Recreational Marijuana
18	Churches Tax-Xxempt
22	Two-State Solution
23	Euthanasia
26	Standard Tests Education
27	Gun Control
28	Prostitution
35	Violent Video Games
38	Medical Marijuana
39	Minimum Wage
40	Death Penalty
43	Bottled Water
49	Body Cameras Police
50	Basic Income

Table A.2: The Touché Topics along together with the tags we assigned in order to have a reasonable search phrase.

ID	Title	Tagged Title
1	Should teachers get tenure?	teacher tenure
2	Is vaping with e-cigarettes safe?	e-cigarettes
3	Should insider trading be allowed?	insider trading
4	Should corporal punishment be used in schools?	corporal punishment in schools
5	Should social security be privatized?	privatized social security
6	Is a college education worth it?	college education
7	Should felons who have completed their sentence be allowed to vote?	felons voting
8	Should abortion be legal?	abortion
9	Should students have to wear school uniforms?	school uniforms
10	Should any vaccines be required for children?	vaccination children
11	Should performance-enhancing drugs be accepted in sports?	performance-enhancing drugs sports
12	Should birth control pills be available over the counter?	birth control pill non-prescription
13	Can alternative energy effectively replace fossil fuels?	alternative energy sufficient
14	Is sexual orientation determined at birth?	sexual orientation birth determined
15	Should animals be used for scientific or commercial testing?	animal testing
16	Should prescription drugs be advertised directly to consumers?	advertising prescription drugs
17	Should recreational marijuana be legal?	recreational marijuana
18	Should churches remain tax-exempt?	churches tax-exempt
19	Should gay marriage be legal?	gay marriage
20	Is drinking milk healthy for humans?	milk
21	Is human activity primarily responsible for global climate change?	man-made climate change
22	Is a two-state solution an acceptable solution to the Israeli-Palestinian conflict?	two-state solution
23	Should euthanasia or physician-assisted suicide be legal?	euthanasia
24	Does lowering the federal corporate income tax rate create jobs?	low income tax jobs
26	Do standardized tests improve education?	standard tests education
27	Should more gun control laws be enacted?	gun control
28	Should prostitution be legal?	prostitution
29	Should the government allow illegal immigrants to become citizens?	illegal immigrants citizenship
30	Should adults have the right to carry a concealed handgun?	concealed handgun
31	Is obesity a disease?	obesity disease
32	Do electronic voting machines improve the voting process?	voting machines
33	Should people become vegetarian?	vegetarian
34	Are social networking sites good for our society?	social networks
35	Do violent video games contribute to youth violence?	violent video games
36	Is golf a sport?	golf sport
37	Is cell phone radiation safe?	cell phone radiation
38	Should marijuana be a medical option?	medical marijuana
39	Should the federal minimum wage be increased?	minimum wage
40	Should the death penalty be allowed?	death penalty
41	Should student loan debt be easier to discharge in bankruptcy?	interests student loan
42	Should fighting be allowed in hockey?	fighting in hockey
43	Should bottled water be banned?	bottled water
44	Should election day be a national holiday?	election holiday
45	Should the penny stay in circulation?	penny circulation
46	Should net neutrality be restored?	net neutrality
47	Is homework beneficial?	homework
48	Should the voting age be lowered?	voting age
49	Should body cameras be mandatory for police?	body cameras police
50	Should everyone get a universal basic income?	basic income

Table A.3: κ of the agreement concerning the annotation classes by topic.

Topic	κ	Interpretation
Abortion	0.595	moderate
Animal Testing	0.707	substantial
Basic Income	0.208	fair
Body Cameras Police	0.298	fair
Bottled Water	−0.082	poor
Churches Tax-Exempt	0.246	fair
Corporal Punishment in Schools	0.231	fair
Death Penalty	0.392	fair
E-Eigarettes	0.501	moderate
Euthanasia	0.330	fair
Gun Control	0.506	moderate
Medical Marijuana	0.268	fair
Minimum Wage	0.246	fair
Performance-Enhancing Drugs in Sports	0.401	moderate
Prostitution	0.413	moderate
Recreational Marijuana	−0.008	poor
School Uniforms	0.082	slight
Standard Tests Education	0.143	slight
Two-State Solution	0.213	fair
Violent Video Games	0.252	fair

Table A.4: The topic precision (relevance criterion (1)) of the Kullback-Leibler heuristic.

Topic	Precision
Abortion	0.95
Animal Testing	1.00
Basic Income	0.85
Body Cameras Police	0.95
Bottled Water	0.90
Churches Tax-Exempt	0.95
Corporal Punishment in Schools	0.80
Death Penalty	0.90
E-Cigarettes	1.00
Euthanasia	0.90
Gun Control	1.00
Medical Marijuana	1.00
Minimum Wage	0.95
Performance-Enhancing Drugs in Sports	0.95
Prostitution	0.85
Recreational Marijuana	1.00
School Uniforms	1.00
Standard Tests Education	0.50
Two-State Solution	1.00
Violent Video Games	0.95

Table A.5: The topic precision (relevance criterion (1)) of the Sentiment Dictionary heuristic.

Topic	Precision
Abortion	1.00
Animal Testing	1.00
Basic Income	0.90
Body Cameras Police	0.95
Bottled Water	1.00
Churches Tax-Exempt	1.00
Corporal Punishment in Schools	0.90
Death Penalty	1.00
E-Cigarettes	1.00
Euthanasia	1.00
Gun Control	1.00
Medical Marijuana	1.00
Minimum Wage	0.95
Performance-Enhancing Drugs in Sports	1.00
Prostitution	0.85
Recreational Marijuana	1.00
School Uniforms	0.95
Standard Tests Education	0.60
Two-State Solution	1.00
Violent Video Games	1.00

Table A.6: The topic precision (relevance criterion (1)) of the Good-Anti heuristic.

Topic	Precision
Abortion	1.00
Animal Testing	1.00
Basic Income	0.95
Body Cameras Police	1.00
Bottled Water	1.00
Churches Tax-Exempt	0.75
Corporal Punishment in Schools	0.95
Death Penalty	0.90
E-Cigarettes	0.95
Euthanasia	1.00
Gun Control	1.00
Medical Marijuana	1.00
Minimum Wage	1.00
Performance-Enhancing Drugs in Sports	1.00
Prostitution	0.70
Recreational Marijuana	1.00
School Uniforms	1.00
Standard Tests Education	0.75
Two-State Solution	1.00
Violent Video Games	0.90

Table A.7: The argumentative precision (relevance criterion (2)) of the Kullback-Leibler heuristic.

Topic	Precision
Abortion	0.75
Animal Testing	0.95
Basic Income	0.75
Body Cameras Police	0.90
Bottled Water	0.45
Churches Tax-Exempt	0.65
Corporal Punishment in Schools	0.80
Death Penalty	0.85
E-Cigarettes	0.35
Euthanasia	0.80
Gun Control	1.00
Medical Marijuana	0.95
Minimum Wage	0.9
Performance-Enhancing Drugs in Sports	0.70
Prostitution	0.70
Recreational Marijuana	1.00
School Uniforms	0.70
Standard Tests Education	0.50
Two-State Solution	0.80
Violent Video Games	0.75

Table A.8: The argumentative precision (relevance criterion (2)) of the Sentiment Dictionary heuristic.

Topic	Precision
Abortion	0.95
Animal Testing	1.00
Basic Income	0.85
Body Cameras Police	0.95
Bottled Water	0.35
Churches Tax-Exempt	0.90
Corporal Punishment in Schools	0.80
Death Penalty	0.95
E-Cigarettes	0.60
Euthanasia	1.00
Gun Control	0.95
Medical Marijuana	0.95
Minimum Wage	0.90
Performance-Enhancing Drugs in Sports	0.90
Prostitution	0.65
Recreational Marijuana	1.00
School Uniforms	0.55
Standard Tests Education	0.50
Two-State Solution	1.0
Violent Video Games	0.95

Table A.9: The argumentative precision (relevance criterion (2)) of the Good-Anti heuristic.

Topic	Precision
Abortion	0.90
Animal Testing	0.95
Basic Income	0.95
Body Cameras Police	1.00
Bottled Water	0.90
Churches Tax-Exempt	0.70
Corporal Punishment in Schools	0.90
Death Penalty	0.90
E-Cigarettes	0.45
Euthanasia	1.00
Gun Control	0.95
Medical Marijuana	1.00
Minimum Wage	0.9
Performance-Enhancing Drugs in Sports	1.00
Prostitution	0.70
Recreational Marijuana	1.00
School Uniforms	0.90
Standard Tests Education	0.65
Two-State Solution	0.80
Violent Video Games	0.80

Table A.10: The stance precision (relevance criterion (3)) of the Kullback-Leibler heuristic.

Topic	Precision
Abortion	0.75
Animal Testing	0.70
Basic Income	0.40
Body Cameras Police	0.90
Bottled Water	0.30
Churches Tax-Exempt	0.40
Corporal Punishment in Schools	0.70
Death Penalty	0.65
E-Cigarettes	0.35
Euthanasia	0.75
Gun Control	0.65
Medical Marijuana	0.85
Minimum Wage	0.7
Performance-Enhancing Drugs in Sports	0.65
Prostitution	0.65
Recreational Marijuana	0.90
School Uniforms	0.50
Standard Tests Education	0.40
Two-State Solution	0.60
Violent Video Games	0.55

Table A.11: The stance precision (relevance criterion (3)) of the Sentiment Dictionary heuristic.

Topic	Precision
Abortion	0.80
Animal Testing	0.60
Basic Income	0.70
Body Cameras Police	0.90
Bottled Water	0.10
Churches Tax-Exempt	0.65
Corporal Punishment in Schools	0.65
Death Penalty	0.80
E-Cigarettes	0.50
Euthanasia	1.00
Gun Control	0.65
Medical Marijuana	0.75
Minimum Wage	0.75
Performance-Enhancing Drugs in Sports	0.90
Prostitution	0.65
Recreational Marijuana	0.55
School Uniforms	0.30
Standard Tests Education	0.45
Two-State Solution	0.70
Violent Video Games	0.55

Table A.12: The stance precision (relevance criterion (3)) of the Good-Anti heuristic.

Topic	Precision
Abortion	0.80
Animal Testing	0.80
Basic Income	0.80
Body Cameras Police	1.00
Bottled Water	0.55
Churches Tax-Exempt	0.60
Corporal Punishment in Schools	0.65
Death Penalty	0.65
E-Cigarettes	0.40
Euthanasia	0.90
Gun Control	0.70
Medical Marijuana	0.80
Minimum Wage	0.80
Performance-Enhancing Drugs in Sports	0.90
Prostitution	0.60
Recreational Marijuana	0.65
School Uniforms	0.75
Standard Tests Education	0.60
Two-State Solution	0.50
Violent Video Games	0.70

Bibliography

Alex Abdo. Facebook is shaping public discourse. We need to understand how | Alex Abdo. *The Guardian*, September 2018. ISSN 0261-3077. URL https://www.theguardian.com/commentisfree/2018/sep/15/facebook-twitter-social-media-public-discourse. 2.2

Joseph Kwame Adjei. The Role of Social Media in National Discourse and Mobilization of Citizens. In *2016 International Conference on Collaboration Technologies and Systems (CTS)*, pages 559–563, Orlando, FL, USA, October 2016. IEEE. ISBN 978-1-5090-2300-4. doi: 10.1109/CTS.2016.0103. URL http://ieeexplore.ieee.org/document/7871041/. 2.2

Rebecca Adler-Nissen, Katrine Emilie Andersen, and Lene Hansen. Images, emotions, and international politics: the death of Alan Kurdi. *Review of International Studies*, 46(1):75–95, January 2020. ISSN 0260-2105, 1469-9044. doi: 10.1017/S0260210519000317. URL https://www.cambridge.org/core/product/identifier/S0260210519000317/type/journal_article. 2.2

M. Afgani, S. Sinanovic, and H. Haas. Anomaly detection using the Kullback-Leibler divergence metric. In *2008 First International Symposium on Applied Sciences on Biomedical and Communication Technologies*, pages 1–5, October 2008. doi: 10.1109/ISABEL.2008.4712573. URL https://ieeexplore.ieee.org/document/4712573. ISSN: 2325-5331. 4.1.3

M. Azar. Argumentative Text as Rhetorical Structure: An Application of Rhetorical Structure Theory. *Argumentation*, 13(1):97–114, February 1999. ISSN 1572-8374. doi: 10.1023/A:1007794409860. URL https://doi.org/10.1023/A:1007794409860. 4

Anne Barnard and Karam Shoumali. Image of Drowned Syrian, Aylan Kurdi, 3, Brings Migrant Crisis Into Focus. *The New York Times*, September 2015. ISSN 0362-4331. URL https://www.nytimes.com/2015/09/04/world/europe/syria-boy-drowning.html. 2.2

Martin A Berger. Fixing Images: Civil Rights Photography and the Struggle Over Representation. *RIHA Journal*, (0010), October 2010. URL `https://journals.ub.uni-heidelberg.de/index.php/rihajournal/article/download/68538/61784`. 2.2

Gemma Boleda. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, 6(1):213–234, January 2020. doi: 10.1146/annurev-linguistics-011619-030303. URL `https://arxiv.org/pdf/1905.01896`. arXiv: 1905.01896. 3.2.1

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of Touché 2020: Argument Retrieval: Extended Abstract. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 12260, pages 384–395. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-58219-7_26. Series Title: Lecture Notes in Computer Science. 5.1.2, 5.1, A

J. Clement. Facebook: number of monthly active users worldwide 2008-2020, August 2020a. URL `https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/`. 2.2

J. Clement. Search engine market share worldwide, October 2020b. URL `https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/`. 3.2.3

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. doi: 10.1177/001316446002000104. 5.2

Francesco D'Orazio. 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, 2015. URL `https://www.pulsarplatform.com/blog/2016/journey-of-an-image-from-a-beach-in-bodrum-to-twenty-million-screens-across-th` 2.2

Ian J. Dove. 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, 2012. doi: 10.1007/978-94-007-4041-9_15. URL https://link.springer.com/chapter/10.1007/978-94-007-4041-9_15. 3

Finis Dunaway. Images, Emotions, Politics. *Modern American History*, 1(3): 369–376, November 2018. ISSN 2515-0456, 2397-1851. doi: 10.1017/mah.2018.17. URL https://www.cambridge.org/core/product/identifier/S2515045618000172/type/journal_article. 3

Ruth Eisenreich. Was den nächsten Aylan Kurdi retten könnte. *Süddeutsche Zeitung*, September 2015. URL https://www.sueddeutsche.de/politik/fluechtlinge-was-den-naechsten-aylan-kurdi-retten-koennte-1.2640109. 2.2

Xénia Farkas and Márton Bene. Images, Politicians, and Social Media: Patterns and Effects of Politicians' Image-Based Political Communication Strategies on Social Media. *The International Journal of Press/Politics*, page 194016122095955, September 2020. ISSN 1940-1612, 1940-1620. doi: 10.1177/1940161220959553. URL http://journals.sagepub.com/doi/10.1177/1940161220959553. 2.2

Susann Fegter. Die Macht der Bilder – Photographien und Diskursanalyse. In Gertrud Oelerich and Hans-Uwe Otto, editors, *Empirische Forschung und Soziale Arbeit: Ein Studienbuch*, pages 207–219. VS Verlag für Sozialwissenschaften, Wiesbaden, 2011. ISBN 978-3-531-92708-4. doi: 10.1007/978-3-531-92708-4_13. URL https://doi.org/10.1007/978-3-531-92708-4_13. 2.2

Ronald Aylmer Fisher and Frank Yates. *Statistical tables for biological, agricultural and medical research, edited by R.A. Fisher and F. Yates. 6th ed.* Edinburgh: Oliver and Boyd, 6 edition, 1963. URL https://digital.library.adelaide.edu.au/dspace/handle/2440/10701. Accepted: 2006-06-27T07:57:52Z. 5.1.3

J.L. Fleiss. The measurement of interrater agreement. In *Statistical methods for rates and proportions*, pages 212–236. John Wiley & Sons, New York, 2 edition, 1981. 5.2

Bernd Frohmann. THE POWER OF IMAGES: A DISCOURSE ANALYSIS OF THE COGNITIVE VIEWPOINT. *Journal of Documentation*, 48(4):365–386, April 1992. ISSN 0022-0418. doi: 10.1108/eb026904. URL https://www.emerald.com/insight/content/doi/10.1108/eb026904/full/html. 2.2

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*, 2012. URL `https://wortschatz.uni-leipzig.de/de/download`. 4.2

Ioana Grancea. Types of Visual Arguments. *Argumentum. Journal of the Seminar of Discursive Logic, Argumentation Theory and Rhetoric*, 15(2): 16–34, 2017. 2.2

Rohmat Gunawan, Alam Rahmatulloh, Irfan Darmawan, and Firman Firdaus. Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath. In *Proceedings of the 2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018)*, Yogyakarta, Indonesia, 2019. Atlantis Press. ISBN 978-94-6252-689-1. doi: 10.2991/icoiese-18.2019.50. URL `https://www.atlantis-press.com/article/55914830`. 3.2.3

Ben He. Probability Ranking Principle. In LING LIU and M. TAMER ÖZSU, editors, *Encyclopedia of Database Systems*, pages 2168–2169. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_930. URL `https://doi.org/10.1007/978-0-387-39940-9_930`. 3.2.2

Benita Heiskanen. Meme-ing Electoral Participation. *European journal of American studies*, 12(2), July 2017. ISSN 1991-9336. doi: 10.4000/ejas.12158. URL `http://journals.openedition.org/ejas/12158`. 1

Allen Kent, Madeline M. Berry, Fred U. Luehrs, and J. W. Perry. Machine literature searching VIII. Operational criteria for designing information retrieval systems. *Journal of the Association for Information Science and Technology*, 6(2):93–101, 1955. ISSN 1936-6108. doi: https://doi.org/10.1002/asi.5090060209. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090060209`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.5090060209. 5.3

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729694. URL `http://projecteuclid.org/euclid.aoms/1177729694`. 4.1.3

Deborah Kurniawati and Deny Triawan. Increased information retrieval capabilities on e-commerce websites using scraping techniques. In *2017 International Conference on Sustainable Information Engineering and Technology*

(SIET), pages 226–229, November 2017. doi: 10.1109/SIET.2017.8304139. URL `https://ieeexplore.ieee.org/document/8304139`. 3.2.3

J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159, March 1977. ISSN 0006341X. doi: 10.2307/2529310. URL `https://www.jstor.org/stable/2529310?origin=crossref`. 5.2

Sabine Maasen, Torsten Mayerhauser, and Cornelia Renggli. *Bilder als Diskurse - Bilddiskurse*. Velbrück, Weilerswist, 2006. URL `https://www.velbrueck.de/out/media/978-3-938808-19-1.pdf`. 2.2

Sanjay Kumar Malik and Sam Rizvi. Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation. In *2011 International Conference on Computational Intelligence and Communication Networks*, pages 465–469, Gwalior, India, October 2011. IEEE. ISBN 978-0-7695-4587-5 978-1-4577-2033-8. doi: 10.1109/CICN.2011.97. URL `http://ieeexplore.ieee.org/document/6112910/`. 3.2.3

G. Moy, N. Jones, C. Harkless, and R. Potter. Distortion estimation techniques in solving visual CAPTCHAs. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages 23–28, Washington, DC, USA, 2004. IEEE. ISBN 978-0-7695-2158-9. doi: 10.1109/CVPR.2004.1315140. URL `http://ieeexplore.ieee.org/document/1315140/`. 3.2.3

Kate Ormiston and Mariki M. Eloff. Denial-of-Service & Distributed Denial-of-Service on The Internet. In Jan H. P. Eloff, Les Labuschagne, Mariki M. Eloff, and Hein S. Venter, editors, *Proceedings of the ISSA 2006 from Insight to Foresight Conference, 5-7 July 2006, Balalaika Hotel, Sandton, South Africa*, pages 1–14. ISSA, Pretoria, South Africa, 2006. URL `http://icsa.cs.up.ac.za/issa/2006/Proceedings/Full/18_Paper.pdf`. 3.2.3

Andreas Peldszus and Manfred Stede. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31, January 2013. ISSN 1557-3958. doi: 10.4018/jcini.2013010101. URL `https://doi.org/10.4018/jcini.2013010101`. 2.1

Georges Roque. 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, 2012. ISBN 978-94-007-4040-2 978-94-007-4041-9. doi: 10.

1007/978-94-007-4041-9_18. URL http://link.springer.com/10.1007/978-94-007-4041-9_18. Series Title: Argumentation Library. 2.2

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x. Conference Name: The Bell System Technical Journal. 4.1.3

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, New Orleans, Louisiana, 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-5005. URL http://aclweb.org/anthology/N18-5005. 2.1

Christian Stab, Tristan Miller, and Iryna Gurevych. Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks. *arXiv:1802.05758 [cs]*, February 2018b. URL http://arxiv.org/abs/1802.05758. arXiv: 1802.05758. 2.1

Statista. Facebook - Anzahl der geteilten Fotos pro Tag 2014, July 2014. URL https://de.statista.com/statistik/daten/studie/312268/umfrage/taeglich-auf-facebook-hochgeladene-und-geteilte-fotos/. 2.2

Benno Stein, Tim Gollub, and Maik Anderka. Retrieval Models. In Reda Alhajj and Jon Rokne, editors, *Encyclopedia of Social Network Analysis and Mining*, pages 1–7. Springer New York, New York, NY, 2017. ISBN 978-1-4614-7163-9. doi: 10.1007/978-1-4614-7163-9_117-1. URL http://link.springer.com/10.1007/978-1-4614-7163-9_117-1. 3.2.2

Werner Stelzner. Diskussion und Logik. *Deutsche Zeitschrift für Philosophie*, 26(2), January 1978. ISSN 2192-1482, 0012-1045. doi: 10.1524/dzph.1978.26.2.209. URL http://www.degruyter.com/view/j/dzph.1978.26.issue-2/dzph.1978.26.2.209/dzph.1978.26.2.209.xml. 2.2

Sung-min Kim and Young-guk Ha. Automated discovery of small business domain knowledge using web crawling and data mining. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, pages 481–484, Hong Kong, China, January 2016. IEEE. ISBN 978-1-4673-8796-5. doi: 10.1109/BIGCOMP.2016.7425974. URL http://ieeexplore.ieee.org/document/7425974/. 3.2.3

Anne Trafton. In the blink of an eye, January 2014. URL `https://news.mit.edu/2014/in-the-blink-of-an-eye-0116`. 4

UNHCR. UNHCR viewpoint: 'Refugee' or 'migrant' – Which is right?, 2016. URL `https://www.unhcr.org/news/latest/2016/7/55df0e556/unhcr-viewpoint-refugee-migrant-right.html`. 6

Ahmet Uyar and Rabia Karapinar. Investigating the precision of Web image search engines for popular and less popular entities. *Journal of Information Science*, 43(3):378–392, June 2017. ISSN 0165-5515, 1741-6485. doi: 10.1177/0165551516642929. URL `http://journals.sagepub.com/doi/10.1177/0165551516642929`. 3.2.3

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5106. URL `http://aclweb.org/anthology/W17-5106`. 1, 2.1, 2.1.1, 3, 4.1

Marilyn A Walker, Pranav Anand, Jean E Fox Tree, Rob Abbott, and Joseph King. A Corpus for Research on Deliberation and Debate. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 812–817, Istanbul, 2012. URL `https://www.aclweb.org/anthology/L12-1643/`. 2.1, 6

Nicholas Watt. David Cameron says UK will take thousands more Syrian refugees. *The Guardian*, September 2015. ISSN 0261-3077. URL `https://www.theguardian.com/world/2015/sep/04/david-cameron-syrian-refugees-uk-will-take-thousands-more`. 2.2

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. page 8, 2005. URL `http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/`. 4.2

Heather Suzanne Woods and Leslie Ann Hahner. *Make America meme again: the rhetoric of the alt-right.* Number vol. 45 in Frontiers in political communication. Peter Lang, New York, 2019. ISBN 978-1-4331-5974-9. 1

Xin Xu, Lei Liu, and Bo Li. A survey of CAPTCHA technologies to distinguish between human and computer. *Neurocomputing*, May 2020. ISSN 0925-2312. doi: 10.1016/j.neucom.2019.08.109. URL `http://www.sciencedirect.com/science/article/pii/S0925231220304896`. 3.2.3

Bo Zhao. Web Scraping. In Laurie A. Schintler and Connie L. McNeely, editors, *Encyclopedia of Big Data*, pages 1–3. Springer International Publishing, Cham, 2017. ISBN 978-3-319-32001-4. doi: 10.1007/978-3-319-32001-4_483-1. URL `http://link.springer.com/10.1007/978-3-319-32001-4_483-1`. 3.2.3

List of Figures

List of Tables