

# Aramis at Touché 2022: Argument Detection in Pictures using Machine Learning

Notebook for the Touché Lab on Argument Retrieval at CLEF 2022

Jan Braker<sup>1</sup>, Lorenz Heinemann<sup>2</sup> and Tobias Schreieder<sup>3</sup>

<sup>1</sup>Student at Leipzig University, Computer Science (M.Sc.)

<sup>2</sup>Student at Leipzig University, Computer Science (M.Sc.)

<sup>3</sup>Student at Leipzig University, Data Science (M.Sc.)

## Abstract

This work deals with the classifying and retrieving images in a data set. Images that argue for or against a topic should be recognized and ordered according to their argumentativeness. Therefore, different approaches are tested and compared with each other. The best results are provided by a neural network, which has been trained to recognize argumentative images with a total of 10,000 labeled images. The model received various features as input, including color, image text and other features. In addition, initial attempts are made to classify the images and their websites in relation to a given question according to their stance into "pro" (the thesis from the question is supported), "con" (the thesis from the question is attacked) and "neutral" (the thesis from the question is supported to the same extent as it is attacked).

## Keywords

machine learning, search engine, argument retrieval, neural network, image search

## 1. Introduction

The availability of information on the Internet is constantly growing. All topics are represented on the Internet with public statements of various points of view. Due to the unequal distribution of opinions and one-sided reports, it is often difficult to guarantee a neutral and balanced search result. Even the evaluation of such a search result causes problems. The retrieval of thematically critical search queries is a particular noticeable effort. For this purpose, there are special argument search engines that filter arguments related to a topic [1] [2] [3]. However, this is often only possible in text form at the moment [4]. Although there are opinions that images cannot be argumentative on their own [5], there are also hypotheses to the contrary. Kjeldsen et al. describe in their work the argumentative character of images and graphics and their various functions [6]. Through the visual component, they can clarify a problem to the viewer and highlight arguments in text form. Certain facts can be presented more convincingly by means of a picture than would be possible in written form [7].

---


CLEF'22: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ [jb64vyso@studserv.uni-leipzig.de](mailto:jb64vyso@studserv.uni-leipzig.de) (J. Braker); [lh31gyzy@studserv.uni-leipzig.de](mailto:lh31gyzy@studserv.uni-leipzig.de) (L. Heinemann);

[fp83rusi@studserv.uni-leipzig.de](mailto:fp83rusi@studserv.uni-leipzig.de) (T. Schreieder)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Accident at the finish sprint at a professional bike race. Six drivers were seriously injured.  
Source: Tomasz Markowski/Associated Press [8]

For example, in professional cycling, there is a debate about whether the barriers in the finish-area should be improved and replaced, as too many accidents are caused by the old fences [8]. A presentation of last year's accident statistics can be a useful and convincing argument for increasing safety in cycling. It is also possible to describe the consequences to highlight the scale of the accident.

*"In a professional race, six riders crashed heavily, with three of them suffering brain and bone damage."* [8]

People can generally imagine text worse than a photo [6]. A picture of the accident increases the visualization and may reinforce the importance of renewing safety measures. A photo of the accident can be seen in Figure 1.

The effective interplay of text and visual graphics is also exploited by the law of mandatory pictorial warnings on cigarette packs in Germany. The dissuasive images of long-term consequences of smoking are intended to discourage people from smoking. According to research by the German Bundestag, these warnings are more effective than a simple warning in text form. Especially the combination of picture and text achieves high efficiency [7].

For this reason, it is of great interest to highlight such arguments and argumentative images in a search query. This would be a useful extension, especially for special argument search engines. A method that can classify pixel-based representations as argumentative has not yet been extensively researched in the literature. The aim of this work is to develop a system that can assess and evaluate images according to their argumentative power in order to drive development forward. One possible application would be to obtain arguments from a search query not only in text but also in image form, to gain an even more detailed overview of the searched problem. Finally, an attempt should be made to assign these arguments to a supportive, neutral or negative stance, whereby the focus of this work is clearly on the recognition of argumentativeness.

## 2. Related Work

Many previous works have dealt with the problem of finding arguments in text collections. The three largest search engines specifically designed for this task are *args.me* [1], *IBM-debater* [2] and *ArgumenText* [3]. They all make it possible to search for arguments in texts for a controversial issue and provide arguments in an ordered and clear manner. Wachsmuth et al. have already dealt with this problem in many works [9, 10, 4] and have successfully shown that it is basically possible to extract text excerpts reliably and determine their stance to the topic. However, they are limited to arguments in text form. Currently there is no published argument search engine that includes images in a search, but conventional image search engines can also achieve adequate results if clearly structured search queries are used [11]. This work will investigate whether it is possible to integrate images as a result of argument search queries.

### 2.1. Image search and image features

Compared to working with texts, the steps of indexing and feature extraction have to be adjusted in the retrieval process for images. Latif et al. summarize in their review article the current technologies and procedures very well [12]. The different features of images are presented. Particularly important for this work are the color-based properties. The concept of dominant colors is presented and referred to the work of Shao et al. [13]. They reduce the countless colors of an image to a few representative colors to search for color-level images faster and more effectively. From this it can be concluded that only a few colors are enough to represent the color conditions of a picture. According to Solli et al. it is also possible to establish a connection between emotions and colors. [14]. They show that people experience similar emotions when looking at certain shades in pictures. Emotions are an important part of [14] arguments. From this it can be concluded that colors might also play a central role in the assessment of the argumentativeness of an image.

In addition, represented objects are important for the message of the image. These can be reliably detected by object recognition. Mokshin et al. use distinctive structures and shapes in the images [15]. However, detection often requires specially trained neural networks, which are trained to reliably detect only a few objects using large training data sets [16]. For the assessment and viewing of many different images this method is not suitable for this work, since for each object to be recognized a training data set with several images of it would be necessary.

### 2.2. Can pictures contain an argument?

Finding arguments in pictures is indirectly also critically considered by many works. Fleming et al. have the opinion that images cannot contain an argument on their own, and only can be supportive [17]. According to Fleming et al., an argument must consist of a claim and a support for that claim. An image cannot perform both functions at the same time. This view is also supported by Champagne et. al [5]. Both complain that a visual representation lacks a textual component to be a clear argument.

However, many pictures also contain texts and diagrams. This would allow them to meet the requirements by Flemming et al. for an argument. The text on images gives the visual elements a context in which to understand and classify them, which according to Kjeldsen et al. is an important part when looking at pictures [18]. This fact must be taken into account in the evaluation and creation of the retrieval system and will be dealt with later in this work. However, it can be clearly stated that it is of great importance to examine the additional elements, such as texts and diagrams, on a picture in order to be able to assess the argumentativeness.

### 2.3. Argument search in pictures

Image search has been widespread for many years and has been enhanced by several algorithms and technologies. The most common methods are described and summarized by Meharban et al. [19].

The search for argumentative images is much less researched than the search for arguments in text form. One approach for image search is to use a search query extension [11], whereby the index for the elements to be searched was created only on the texts of the pages of the images. In a search query, only the words *good* and *anti* were added and searched for matches in the documents. The result were already on a good level. However, no information from the image was used and the process is based on the assumption that the text used on a website is representative for each image embedded in it. This assumption has to be questioned critically: several images of different content can be integrated on one page, which makes it difficult to assign the texts clearly.

This paper aims to take a different approach and to focus exclusively on the picture when assessing argumentativeness. Information from the website should only be used later, when classifying the stance of an argument.

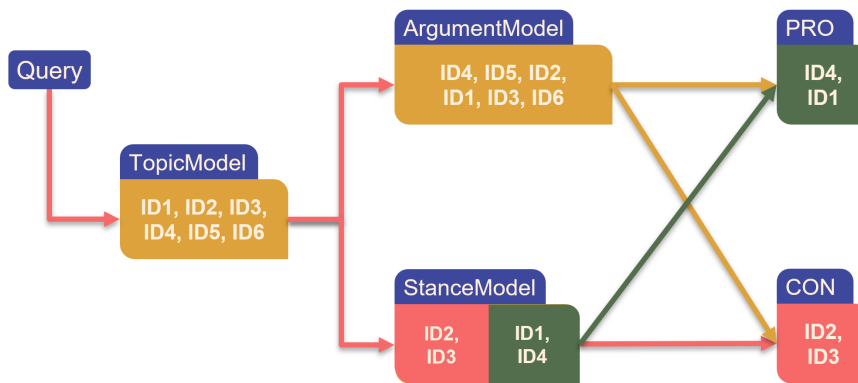
## 3. Retrieval System

To determine the relevance of an image to a given query, Kiesel et al. [11] distinguish three level of relevance. An image is considered relevant to the search engine if it is both topic-relevant, argumentative, and stance-relevant [11]. Stance-relevance refers to the attitude within the discussion.

The tripartite division was adopted for this work and a separate model was developed for each of the three level. The goal of the retrieval system is to find the top k relevant images for both the "pro" and the "con" side. Figure 2 shows the retrieval system in a simplified form. The starting point is always a query as input.

The query and most of the texts processed by the system first go through a preprocessing. It uses the Spacy (en-core-web-sm) language model [20] to get a tokenization on the text first. Subsequently, punctuation marks and stopwords are removed. The remaining tokens are finally lemmatized before they are passed on to the appropriate part of the retrieval model.

A preprocessed query is first entered into the *topic model*, which calculates the affiliation to a topic for each image in the data set by using a DirichletLM model. According to the studies of



**Figure 2:** Overview of the retrieval system with the sketched sequence of the image search via the IDs. Argument and stance models evaluate the images independently. After combining the scores, the pictures are sorted accordingly. The images with the highest score can be considered as pro, and those with the lowest score as con.

Potthast et al. the DirichletLM performs significantly better in argument retrieval compared to TFIDF and BM25 [21]. The text of the HTML page of the image is used as input. Since the retrieval of the topic relevance is not be focused by this work, further information provided by the underlying data set is taken into account. This includes, that each image has already been assigned to at least one topic. Additionally, each topic has a handful of example queries. With these and the user query, the topic of the query is determined and a DirichletLM retrieval is performed on the images of the determined topic. Because of that, it is assumed that the topic model returns subject-related images. No further evaluation will take place. As a result, the topic model returns a list of image IDs with a calculated score representing the topic relevance in relation to the query.

The next step is to pass the list of image IDs in parallel to the *argument model* and the *stance model*. The argument model now calculates a score for the argumentativeness of the respective image. At the same time, the stance model classifies the same images into the classes "pro", "con" and "neutral". Note, however, that only the image IDs that have been classified as either "pro" or "con" are returned by the model in two separate lists. Neutral images are ignored. In some cases, this can lead to an unbalanced retrieval result of the search engine if fewer images are classified for one of the sides. This approach was deliberately chosen because the number of arguments per page gives the user additional value in the search query. In the final step, both the "pro" and "con" lists of the stance model are sorted according to the scores calculated by the argument model. The results are two lists of image IDs in descending order, which ideally contain the images, which support or attack the thesis sought the most. The argument and stance models are discussed in much more detail in the following two chapters.

## 4. Argument Model

To be able to rank images according to their argumentativeness, each image must be given a score. The higher this score is, the better an image argues in relation to a topic. An argument that is critical and/or supportive results in a high score. The position for which it argues is not considered. This score makes it possible to search for argumentative images without having to specifically understand their content and assign it to the issue. A diagram for example often has an argumentative supporting character. This can be considered in the argument model and evaluated accordingly. In this case, only the diagram must be recognized, since the expression of the stance is unimportant for the argument model.

There are several such features that can be searched for on an image. Whether and how strongly the occurrence and use of these features correlates with the argumentativeness of an image is to be examined later.

### 4.1. Image Features

The first thing to look at are the features that can be obtained from the image alone. All images are in PNG format and have a rather low resolution, since they are downloads of embedded images on web pages.

#### 4.1.1. Color Features

Colors can be represented on a computer by the RGB color model. It uses three numerical values between 0 and 255 for the colors red, green and blue. From them, a color is described exactly by means of an additive color model. Each color value was normalized between 0 and 1 by means of the maximum value equal to 255.

##### **Average Color**

The first feature used for the argument score is the *average color of an image*. For this purpose, all color values of the pixels of an image are averaged. Possibly, a general color mood and also an emotion when looking at the image, can be detected via this. One hypothesis to be tested is that certain colors are used more often in argumentative images than others. For example, the colors red and green, following the colors of a traffic light, could be used to highlight positive and negative elements as indicators [14]. This feature consists of three values, respectively for the red, green and blue values.

##### **Dominant Color**

However, the average color has a decisive disadvantage. If an image has many red and green elements, which could possibly stand for a strong argumentativeness, this cannot be detected by this feature. The average color mixes the colors together to a new color. Thus, in the additive color model, red and green make yellow, making it impossible to distinguish whether the image is yellow, or red and green. One solution is to use *dominant colors*. Here, the most used colors of an image are considered. The color values of the pixels are grouped and the most used color is output as the dominant color. To avoid grouping by exact colors, the image can optionally be

reduced to fewer colors beforehand, creating color intervals, which results in higher accuracy, especially for photographs and color gradients. As a feature, only the first dominant color is used, which in turn consists of the three RGB values. Consequently, the second most used color is no longer considered. The effects of this decision could be analyzed in further studies.

### **Percentage Color**

In order to investigate the hypothesis that different colors are used more frequently in argumentative images than in non-argumentative images, the color proportions are determined. The colors red, green and blue as the three colors of the color model were considered, additionally yellow as a neutral color between red and green was added. The color proportion is determined by examining each pixel color to see whether it lies within the specified color interval of the color to be examined. The interval is necessary because, for example, the color green describes a variety of hues. These intervals were specified in the HSV model because the brightness and saturation of a color can be considered independently. These have no effect on the hue. A binary image mask is then created, where a pixel is colored white if the pixel color value is in the interval, and black if not. The ratio of black and white pixels in the image gives the color portion of the color being searched for. The following color intervals were defined (Hue, Saturation, Light value):

- *Red*: (0, 50, 80) to (20, 255, 255) and (160, 50, 80) to (255, 255, 255)
- *Green*: (36, 50, 80) to (70, 255, 255)
- *Blue*: (100, 50, 80) to (130, 255, 255)
- *Yellow*: (20, 50, 80) to (36, 255, 255)

Since the color red lies in the middle of the zero point of the color scale, two color intervals are necessary to capture the entire color. The same procedure can be applied to the brightness of the image. The HSV color model makes it very easy to filter the light and dark color areas. The thesis in this context is that possibly light colors could be used more in the positive context and dark colors in the negative context.

- *Light*: (0, 0, 200) to (255, 60, 255)
- *Dark*: (0, 0, 0) to (255, 255, 60)

#### **4.1.2. Optical Character Recognition**

*Optical Character Recognition* is a common technique that Google Inc. used to develop their text recognition *Tesseract* [22]. This open-source software is used to recognize text on images. In this work, each image was checked for recognizable texts using *Tesseract*. It is noticeable that handwriting is not recognized. Standard fonts with high contrast to the background are found best on images. This is problematic for demonstrations with handwritten posters or photographs with low-contrast fonts. Text in languages other than English is also recognized more poorly. Normally, *Tesseract* only outputs letters that could be identified with a high probability. However, for this work, the threshold for the probability of a match was lowered, which allowed more text to be detected on the images. The output contains many single letters and symbols, making

it advisable to filter for words with a letter length greater than two. In addition, the words found were checked against a lexicon and words not in the English lexicon were also filtered out.

### **Text Length**

Text length is a feature, which corresponds to the number of words found after post-processing. The assumption is that a long text has a higher potential to be argumentative. The text length is then normalized with the following function, where  $x$  is the number of words found:

$$\text{textLengthFeature} = 1 - e^{-0.01 \cdot x} \quad (1)$$

This ensures that a text twice as long not results in a feature value twice as large. Furthermore, from a text length of 200 onwards, there is no longer a large change and the value is asymptotically one, because the subtrahend becomes zero by a large divisor.

### **Sentiment Score**

Furthermore, the text is given into a *sentiment analysis*, which returns a *sentiment score* between -1 and 1. A sentiment analysis is a topic of text mining and examines a text snippet for a positive or negative stance. Basically, different words are associated with a certain stance. For example, the word "anti" might indicate a negative stance. The average stance of a text is returned as a continuous numerical value, where -1 indicates a negative text, and 1 indicates a positive text. In this work, a lexicographic approach was taken using the VADER lexicon, with image texts evaluated using NLTK sentiment analysis [23]. However, many images do not include text, or it was not detected. In addition, sometimes individual words can strongly influence the sentiment score, even though the context would suggest the opposite influence. For the argument model, the absolute value of the score was used, since it is irrelevant whether the argument is for the positive or negative side.

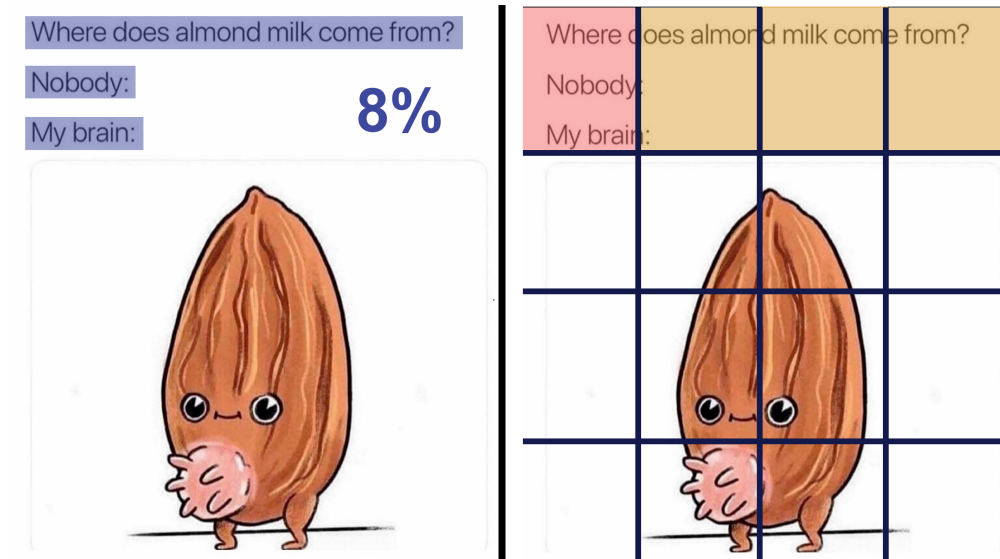
### **Text Area**

In addition to the individual letters, Tesseract also returns the position and size of the text. Two points on the image plane describe the upper left and lower right point of the smallest possible rectangle that can be placed around the font elements. These areas can be added and put into a ratio with the total area of the image. The resulting percentage value describes how much area of the image is taken up by the text (seen in Figure 3). Possibly, the area taken up by the text should be understood as a kind of weighting of the sentiment score. For example, logos or source citations are often included, which are displayed small at the bottom. These may be less important than large headlines.

### **Text Position**

Furthermore, the text position on the image is examined. The theory behind this is that many images contain text captions or headings, with many texts found on the edges of the images. These texts could possibly be attributed a different meaning than the text found in the middle of the image. This was realized using a heat map, with the image divided into an even 8x8 grid. For each field, the amount of text is determined and the 64 values are stored in a two-dimensional array. This can be seen as an example in Figure 3.





**Figure 3:** (left) Text Area calculated based on the area occupied by the text. (right) Text Position calculated as an example of a heat map over a 4x4 grid. In the work, a 8x8 grid was used. A red color represents a high text content.

#### 4.1.3. Additional features

Two additional features were created after an analysis of the data set revealed that many images were graphics and not photographs: *Image Type* and *Diagram Detection*. The assumption was made that graphics have a higher potential to be argumentative than photographs. This was said to be due to the more frequent presence of text and diagrams, which form *assertion* and/or *support* of the argument [17]. To do this, it must be recognized what type of image is involved. Also, it would be reasonable to assume that multiple and larger diagrams have a higher argumentative character.

##### Image Type

A distinction is made between graphics (cartoons, clipart) and photographs. Abd Zaid et al. showed in their work that cartoons generally consist of significantly fewer colors than photographs [24]. This can be used to build a classifier which distinguish between cartoons and photographs. The classifier looks at the ten dominant colors and their image proportion. If these take up more than 30% of the image, it is assumed to be a graphic.

##### Diagram Detection

To recognize diagrams, the image is preprocessed in several steps. The procedure was described this way by the user *nathancy* on the page *Stackoverflow.com* [25]. First, the image is converted to a binary image by a threshold-value. Now the image consists only of black and white pixels. If the contrast between text and background is high enough, the text is clearly visible. The text can be removed using an extended horizontal kernel. The kernel removes all elements and image-areas, which looks like small horizontal lines, just like a line of text. By assuming that the

text was written horizontally, all lines of text will be removed and colored black. All remaining image elements, which are the left over white areas, are no text and combined and extracted. The ratio of the size of these elements to the total image size forms the actual feature. It becomes problematic if several diagrams are contained and recognized as one by the algorithm. Since the smallest possible bounding box is determined, the area portion can have a large error. Also problematic are logos in the corners of the images, which makes the bounding box unnecessarily large if additional diagrams are included on the image. Colored diagrams are recognized worse due to the binary filter, if they are not sufficiently different from the background. This filter is necessary, however, because otherwise the kernel cannot recognize the horizontal structures. Overall, the diagram recognition works well enough to add a benefit to the project. However, the percentage value as a feature implies that a diagram that takes up the entire image is the best. Because of the described error, the feature value should be urgently 0 for a diagram image percentage of 100%. The optimal image proportion is assumed to be 80%. From these defaults, a function can be derived, which converts the image portion accordingly. For this purpose, a log-normal distribution was used. The variable  $x$  is the determined value of the diagram recognition between 0 and 1. The value range of the function is also between 0 and 1.

## 4.2. Argument Standard Model

The first attempt was to build a formula with the normalized features, whose output is a numerical value that describes the argumentativeness of an image. For this purpose, an attempt was made to implement the assumptions made within the formula. A higher result should be associated with a higher argumentativeness.

$$argumentScore = \alpha(colorScore) + \beta(textScore) + \gamma(diagramFeature) \quad (2)$$

The factors  $\alpha, \beta, \gamma$  describe the influence of the individual scores on the *argumentScore*. They should lie between 0 and 1, whereby the optimal weighting is to be determined later by an evaluation. The *colorScore* implements the assumption that light and green colors are more likely to be found on positive images, and dark and red colors on negative images. However, this only applies to photographs, as cliparts contradict the assumption due to their often white background. For this reason, the *colorScore* is included when calculating the *image-type*. Since the stance of the image does not matter in a discussion, the positive and negative assumptions do not need to be distinguished. Therefore, they are added in the formula.

if Image Type == 'photo':

$$colorScore = \sigma(\%green + \%red) + \frac{1}{\sigma}(\%bright + \%dark) \quad (3)$$

$\sigma$  is the weighting ratio between the color component (red, green) and the brightness (light, dark). In this work, a  $\sigma$  of 0.8 is used, which weights the colors significantly higher.

if Image Type == 'clipArt':

$$colorScore = (\%green + \%red) \quad (4)$$

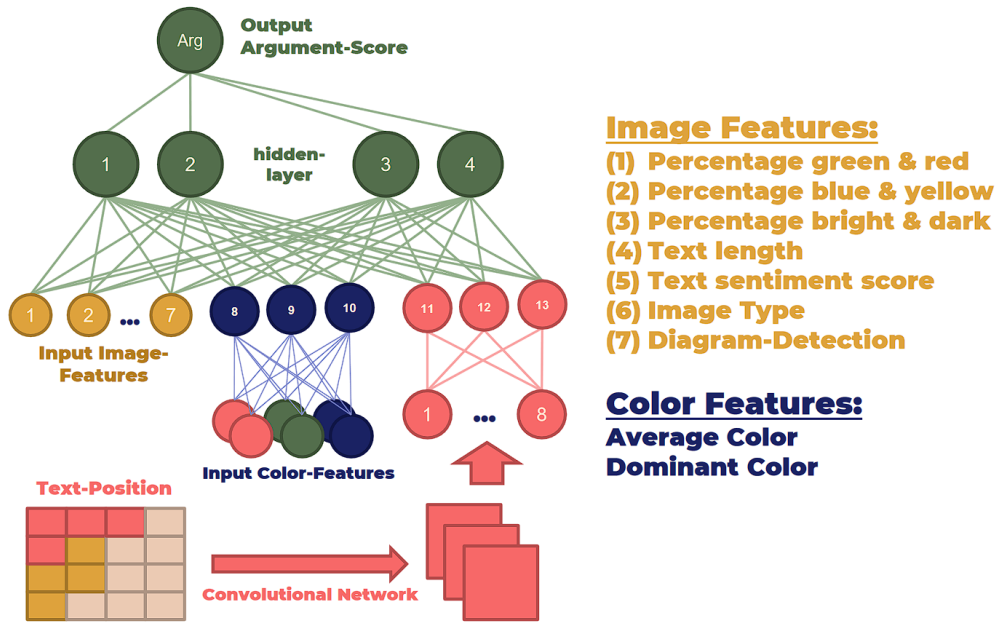


Figure 4: Topology of the used neural network for argument detection.

The *textScore* describes the assumption that longer texts have a higher potential to be argumentative. Furthermore, the sentiment score is included.

$$textScore = textLength \cdot |textSentiment| \quad (5)$$

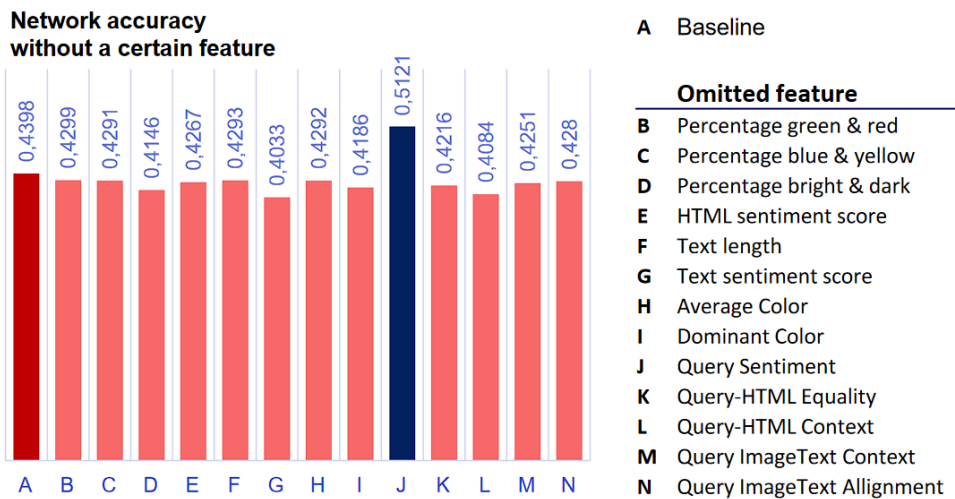
The formula thus formed for calculating the *argumentScore* evaluates images in the value range 0 to 3, where each term can become a maximum of 1.

This formula is based on many assumptions and theories that could not be substantiated in this work due to time constraints. This would require an extensive analysis on a data set labeled in this regard. For this reason, another approach was tested which is based on a neural network and is described below.

### 4.3. Argument NeuralNet

The goal in using a neural network is to not have to set the parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  yourself. In addition, the network should ignore possibly incorrect assumptions, such as the color assumptions on argumentativeness. It first needs a topology that allows all features to be reasonably available to the network.

The simplest variant is to use a fully-connected network, where all neurons of one layer are connected to the neurons of the next layer. The relu function serves as an activation function of the neurons. However, it is not recommended to put all features into the input layer. Although they are already all normalized, the color features receive a presumably higher importance



**Figure 5:** Evaluation of the significance of the input features. If the accuracy of the network is low without a certain feature, this feature has high influence on the predictive performance of the network. For the baseline, all features are used.

due to their number. For this reason, it is advisable to use a combined network in which color features are reduced in number in an upstream network. In addition, it offers itself to analyze the text position by means of a Convolutional network.

The 64 values of the heatmap are reduced to three values by three filters and a small fully-connected layer. Color features are also reduced to three values. Together with the remaining features, the six values are given to a larger fully-connected network (illustrated in Figure 4).

The evaluated images in the data set serve as training data, where an evaluation with *strong* is mapped to a value of 1 and an evaluation with *weak* was mapped to 0.5. The rating scale is defined in more detail in chapter 6.1. To prevent overfitting, an automatic stopping was used that terminates the training as soon as the validation accuracy stops increasing within 10 epochs.

Since in summary 10 features are given to the network, it is advisable to evaluate which features actually add value and which do not. Each feature is based on an assumption for the detection of argumentativeness. However, since these are not proven, the evaluation is at the same time a hypothesis testing.

The features were evaluated by training a baseline model. This contains all the features presented. Subsequently, each feature was omitted once in the training. The average of accuracy was calculated from 10 trained and evaluated models. The results are shown in Figure 5. It can be seen that the network relies heavily on the *textLengths* feature, as the accuracy decreases when the feature is left out. Furthermore, it can be seen that the accuracies increase when *Percentage blue&yellow*, *Image Type* and *Text Position* are excluded during training. In terms of assumptions, this means that argumentativeness does not depend on the yellow and blue content. The same is true for the text position on the image. The argumentativeness of an image

seems to be relatively strongly dependent on the text length found on the image. The result from the evaluation is that *Percentage blue&yellow*, *Image Type* and *Text Position* are no longer considered. It was also tried not to include *Average Color*, but the accuracy decreased in relation to the baseline when excluding all four features.

## 5. Stance Model

Even though the focus of this work is on the recognition of argumentativeness, an attempt was made to recognize the expression of stance relevance within the discussion.

### 5.1. Stance detection features

An image can be *supportive*, neutral or *rejective* towards a question. To recognize this, information about the question itself is required. Since it is difficult to make the content of the question understandable to an algorithm, the following features try to establish a connection between image text and query. However, on many images no text is recognized, or they do not possess any, whereby a comparison is not possible and an attitude concerning the question cannot be recognized. For this reason, the HTML page of the image was included. There, the surrounding text was extracted and special emphasis was placed on captions. Thus, HTML text and image text can be compared with the query regarding features.

#### Query Equality

The *Query Equality* describes the equality between the query and the image or HTML text. This is based on the assumption that if the query itself, or parts of it, are found in the text, there is a high correlation between them. This is implemented by a simple term index, for which the occurrences of the preprocessed query terms in the text are counted.

#### Query Alignment

Similar to the *Query Equality*, the *Query Alignment* also searches for matches between query and text. The Needleman-Wunsch algorithm is used to find optimal alignments [26]. Due to the computation time and the sometimes very long HTML texts, the feature was only computed between query and image text.

#### Query Context

However, finding query terms in the text is not enough to make statements about the attitude of the text. Negations, which can reverse statements using a word, pose a major problem. The *Query Context* feature looks at query term occurrences in the text and evaluates the surrounding words in a  $\sigma$  environment with respect to the sentiment score.  $\sigma$  is to be chosen depending on the text at hand. Since in this work the text was preprocessed,  $\sigma$  is chosen quite low, since filler and stop words are no longer included. Recommended is  $\sigma = 6$ . The assumption underlying the feature is that if a negative sentiment score is found around query term occurrences, there is also a negative correlation between the text excerpt and the query term. For each occurrence in the text, the sentiment score is determined and the average is calculated at the end.

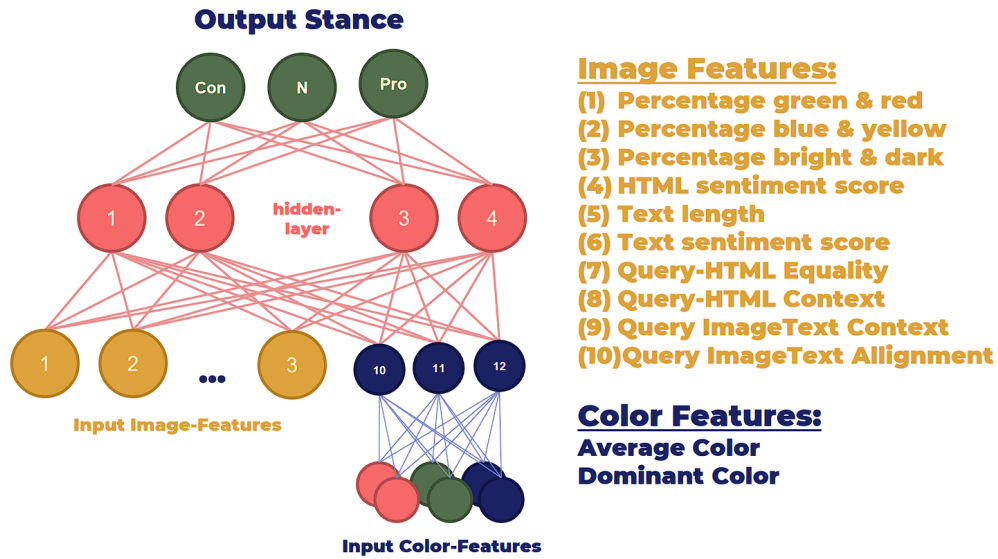


Figure 6: Topology of the neural network for stance determination.

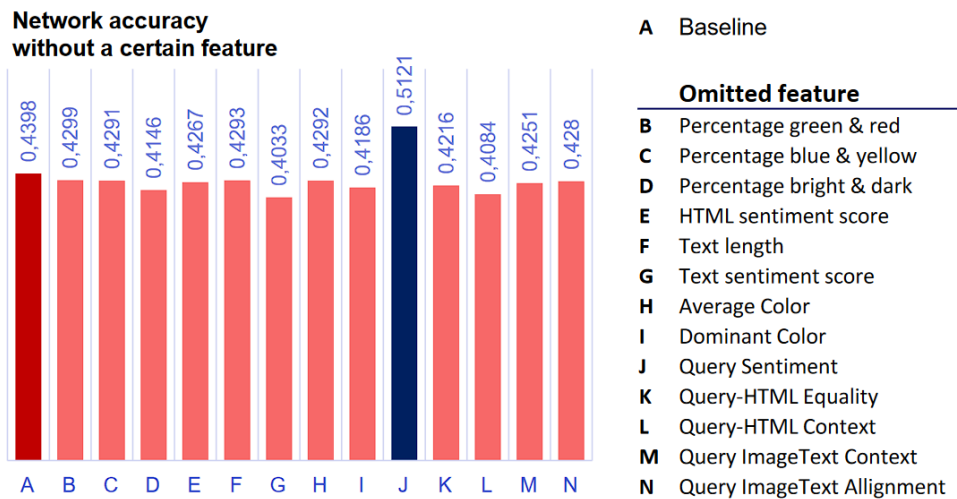
This results in a total of 5 features, since *Query Equality* and *Query Context* are each calculated between query and either image text or HTML text.

## 5.2. Stance NeuralNet

Similar to the argument model, an attempt was made to determine the attitude of an image with respect to a search query by means of a formula. This formula included the described features from section 5.1, but also some features from the standard argument model. However, the accuracy of the used formula and a random assignment couldn't be distinguished. For this reason, a neural network was also trained, which has similarities to the argument model. It uses the stance detection features from section 5.1 and some argument model features. The stance NeuralNet model can be seen in Figure 6.

The three classes have different occurrences in the data set. Because of this, they were weighted differently in the training. The weight of the pro and con classes is the number of neutral images divided by the number of pro/con images. The neutral class has a weight of 1. These weights ensure that the network don't prefer the neutral class due to its frequency.

It should be emphasized that unlike the argument model, no continuous value is to be predicted. The stance model functions as a classifier. Each output neuron represents one of the three stance expressions of the image "con", "neutral" and "pro". During training, the information is converted into binary vectors  $(0, 0, 1)$  for "pro",  $(0, 1, 0)$  for "neutral", and  $(1, 0, 0)$  for "con". In prediction, the largest value of the three neurons is interpreted as the predicted class.



**Figure 7:** Evaluation of the significance of the input features. If the accuracy of the network is low without a certain feature, this feature has high influence on the predictive power of the network. For the baseline, all features are used.

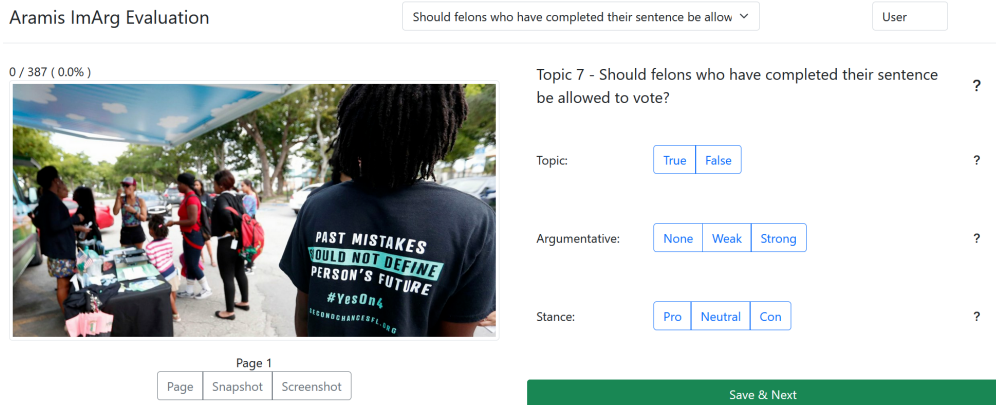
An evaluation of the features can also be made for this network. Only the sentiment score of the query does seem to be disadvantageous for the network. The accuracy in relation to the baseline when excluding this feature increases significantly (to be seen in Figure 7). Consequently, the feature is not further used in the network.

## 6. Evaluation

Subsequent to the presentation of the various models and features, these should now be evaluated. For this purpose, the labeling process and the resulting outcomes will be presented. In connection with this, both the argument model and the stance model will be evaluated and their performance is to be determined using suitable metrics.

### 6.1. Data Set

This work is based on a data set with a total of over 23,000 images, divided into 50 different topics. The data set contains both the images and the HTML text in which each image appears, as well as some additional information. In order to be able to train models on this data, the data must first be labeled. For texts, there are ready-made platforms for semantic annotation, such as the INCEpTION [27] platform. In order to also label images in terms of their topic relevance, argumentativeness, and stance relevance, a separate web frontend was first developed, as can be seen in Figure 8. This allowed a large portion of the data set to be annotated as efficiently as possible. The recognizable tripartition of the labels refers to the considerations of Kiesel et al. [11], whereas the different expressions were adapted to the questions and problems of this work.



**Figure 8:** HTML frontend for labeling the images of a selected topic. Different topics can be labeled by several users at the same time in terms of their topic relevance, their argumentativeness and their stance relevance.

Below is an explanation of the different labels [11]:

- *Topic*
  - **True:** From the image (recognizable content) you can see the topic.
  - **False:** From the image (recognizable content) you can not see the topic.
- *Argumentativeness*
  - **None:** There is no argument recognizable in the image that argues for any position in a topic. If the image does not belong to the topic, the argumentativeness must be set to "none".
  - **Weak:** Few arguments are recognizable in the image or/and the arguments are not clear.
  - **Strong:** Several arguments are recognizable in the image or/and a clear stance is recognizable in each argument.
- *Stance*
  - **Pro:** A clear attitude can be seen throughout the image, which supports the thesis of the topic.
  - **Neutral:** The picture is not argumentative or arguments are made in equal measure for and against a topic. If the image does not belong to the topic, the stance must be "neutral".
  - **Con:** A clear attitude can be seen throughout the picture, which attacks the thesis of the topic.

When assigning labels, it is important to note that images that are not topic relevant (topic = false) cannot be argumentative and are assigned a neutral stance. The underlying assumption is that both argumentativeness and stance relevance are directly dependent on topic relevance.



For example, an image is labeled "neutral" in terms of stance relevance if the topic reference is missing, even if the image could be "pro" or "con" to another topic.

With regard to argumentativeness, a distinction is made between strongly and weakly argumentative images. This is intended to train the model in such a way that strongly argumentative images get a higher score and can therefore ideally be shown first by the search engine. In the further course, a more precise distinction is made in the evaluation with the addition "strong", which means that only strongly argumentative images are considered here.

A total of almost 10,000 images were annotated through the labeling process. There are clear differences between the various topics in the distribution of the label characteristics. Table 1 shows the label results averaged across all annotated images. First of all, it can be deduced from the results that only 72% of the images in the data set that have already been assigned to a topic are actually topic relevant. Since argumentative images must also be topic relevant, the proportion of argumentative images is obviously lower at 46%. However, a clear difference can also be seen with the strongly argumentative images. With 14% of the images, these are represented rather rarely in the data set. A clear stance relevance is still recognisable in 34% of the images, whereby "con" images are represented less frequently in the data set with 14% compared to "pro" images with 20%. In conclusion, 34% of the images are topic relevant, argumentative as well as stance-relevant and are therefore relevant for the search engine. When considering only the strongly argumentative images here, the proportion is reduced to 13%. Even more serious are the differences between the various topics. For example, in the topic "Should the penny stay in circulation?" only 9% of the images are argumentative and only 2% of the images have a stance relevance of the type "pro" (see Table 8 in the appendix). These outliers mean that the retrieval system can only find very few relevant images. In a rank-based evaluation, the Precision@k evaluates the performance of the model for the best k results. If k is chosen to be greater than the number of relevant images contained, the performance of the model is underestimated. This is already the case from a k of 20 for some topics.

Category	Percentage in data set
Topic Relevance	72%
Argumentativeness	46%
Argumentativeness (Strong)	14%
Stance Relevance	34%
Stance Pro	20%
Stance Con	14%
Stance Neutral	66%
Relevant Images	34%
Relevant Images (Strong)	13%

**Table 1**

The analysis shows the ratio of the number of images that fulfil the respective property to the total number of images. All 20 labeled topics are taken into account. The addition of "strong" to the argumentative and relevant images means that only the strong argumentative images are taken into account, whereas without the addition both strong and weak argumentative images are meant. Relevant images are all images that show a topic relevance, are argumentative and have a stance relevance with the characteristic "pro" or "con". In the best case, these should be displayed by the search engine.

## 6.2. Evaluation Argument Model

In the following, the argument model will be evaluated. A differentiation is made between the presented standard and NeuralNet model. All calculated and presented values were cross-validated and averaged from multiple runs to obtain representative data.

For the training of the model, topics are needed that contain enough relevant images. Based on the analysis of the labels, skip topics were defined, which are ignored for the training of the models and partly also for the subsequent testing. The analysis of the data set has shown that some topics (skip topics) lead to a degradation of the model performance when they are used. This is due to the fact that they contain too few argumentative images for the network to train. However, there are two separate skip lists for the argument model and the stance model. In the argument model, all topics that do not have at least 20 strong argumentative images are added to the list and later ignored during training. The stance skip list contains topics that do not have at least 20 "pro" and 20 "con" images. The choice of the value 20 is taken into account in the following subsections for calculating the model performance.

There are two topics in the argument skip list. If the images of these topics are removed from the data, the remaining data is called *valid*. For the evaluation of the standard model, a separate examination and evaluation takes place with all and the valid data. A Precision@20 is calculated in each case. Strong@20 evaluates how many of the best 20 search results were actually labeled as *strong*. Since the valid data for each topic contains at least 20 strongly argumentative images in the data set, in the optimal case these images would also be at the top of the ranking. For Both@20, both strong and weak argumentative images are considered correct and are not distinguished.

$$Precision@k = \frac{|\{r \in I | r \leq k\}|}{|I|} \quad (6)$$

The Precision@k describes according to Berrendorf et al. [28] the proportion of hits or instances for which the true entity appears among the first  $k$  entities in a sorted list. The  $r$  represents the respective rank from the set of individual rank scores  $I$ . The presented model assumes that the inputs are topic relevant. Due to the fact that there is no further check for this, images without topic relevance can also be found in the output. Therefore, only the first  $k$  topic relevant images are considered when calculating the Precision@k.

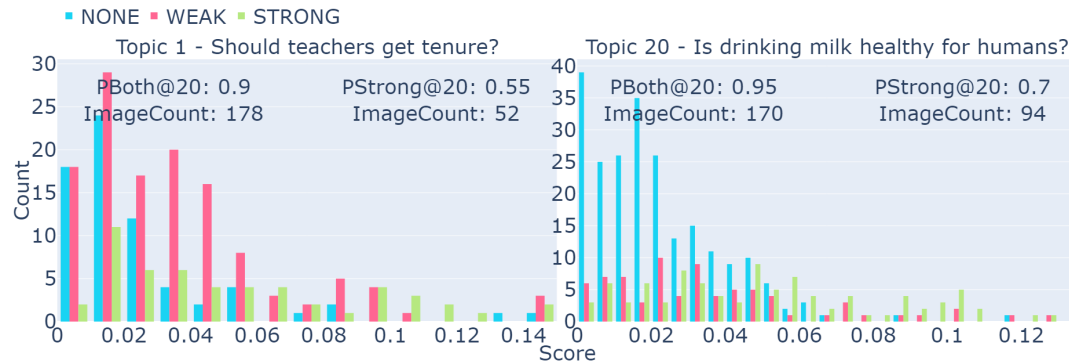
### 6.2.1. Argument Standard Model

The results from table 2 show that the model performance decreases slightly when all topics are considered. This is because the topics in the skip list do not contain 20 strongly argumentative images, and the Precision@k for these topics can never become 1. With a precision Both@20 of 0.875, this means that among the 20 top-rated images, on average 17.5 images are actually strongly or weakly argumentative. The other predicted images were classified as "none" in the labeling process. The predictions can thus be considered acceptable, but not particularly good, since the precision decreases especially with increasing  $k$ .

Topics	Strong@20	Both@20
all	0.5225	0.8475
valid	0.5472	0.8750

**Table 2**

Results of the argument standard model: To calculate the results, the features described in chapter 4.2 are included in the model with equal weighting. Strong@20 means Precision@k with k=20, whereby only the strongly argumentative images are considered. With Both@20, both strong and weak argumentative images are considered. With topics "all" all labeled topics are used, while with topics "valid" all topics are used excluding the previously defined skip topics.



**Figure 9:** Histograms for the argument standard model over a selected sample of valid topics with the distributions of the scores assigned by the model and the calculated Precision@20 values. PBoth@20 shows the determined precision for weakly and strongly argumentative images. With PStrong@20, only strongly argumentative images are considered. ImageCount shows the absolute number of images that are available in the data set for the category ("Strong" or "Both") under consideration.

Figure 9 shows the scores assigned by the standard argument model for two representative topics in the form of histograms. Ideally, strongly argumentative images should be found in the right part of each plot (green), weakly argumentative images in the middle part (red) and non-argumentative images in the left part (blue). Three distributions would thus be visible, which are spatially separated from each other on the x-axis. The argument standard model does not produce such clearly independent distributions. The right-skewed distributions produce good Precision@20 values, but this performance drops sharply as k is increased.

### 6.2.2. Argument NeuralNet

For the NeuralNet model, data splits for training and testing must be defined. This is realised in two different ways. On the one hand, there is a split at the image level, where all labeled topics are taken into account and the images of all topics are divided into 2/3 training data and 1/3 test data. This split allows the model to learn the characteristics of each topic. In order to investigate possible overfitting to the given topics, another split is made at topic level. Here, individual topics are left out from training and only used for testing. These test topics contain enough argumentative and strong argumentative images and are taken from the valid topics.

The retrieval system is set up in such a way that individual topics, but not individual images, can be processed in the evaluation. As a result, a correct evaluation on only the test data is not possible in case of a split on image level. Since the training data is also included in the evaluation data, overfitting is difficult to detect. For this reason, the ratio of images serving as train and test data was varied and different models were trained with them. The test data serves the neural network as a validation data set and, with the accuracy calculated on it, determines when the training is stopped in order to avoid overfitting. It was discovered, that already from a training share of the data of 10% no remarkable changes in the model performance can be recognized (For further results see figure 12 in the appendix). Accordingly, this proportion is already sufficient for the model to learn the essential features of argumentative images. Since, in the worst case, overfitting occurs on the training data, and the test data is significantly worse predicted, there should be a change in overall precision when varying the split between the two. Since this is not the case, overfitting on the data can be excluded with sufficient certainty and an evaluation can also take place with training data contained in the test data.

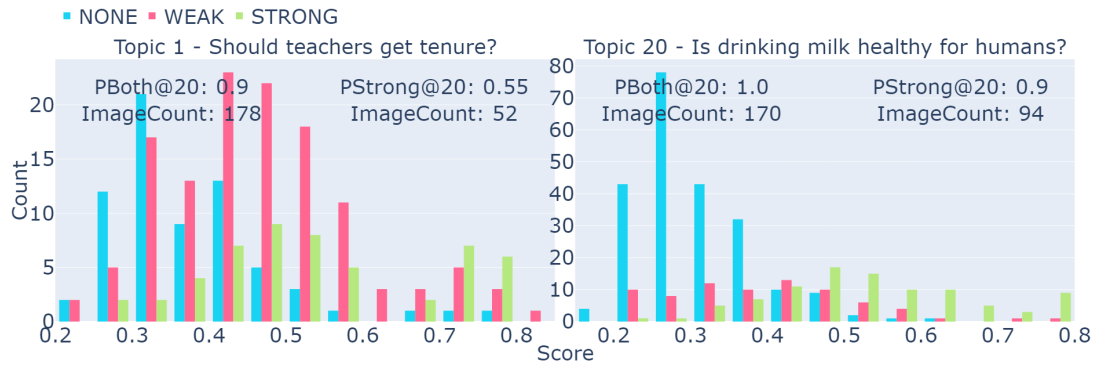
Data Split	Topics	Strong@20	Both@20
Image Level	all	0.5572	0.8472
Image Level	valid	0.5850	0.8764
Topic Level	all	0.5420	0.8568
Topic Level	test	0.5550	0.8729

**Table 3**

Results of the argument NeuralNet model. The table shows the Precision@20 values determined on the one hand for exclusively strong argumentative images (Strong@20) and on the other hand for strong and weak argumentative images (Both@20). A distinction is made as to whether a data split was applied at image level or at topic level, and whether all topics, the valid topics or only the test topics were taken into account.

Table 3 shows the results for the NeuralNet model for both an image level and topic level data split. The column *topics* describes which data was used for the evaluation. The model for the data split at image and topic level was the same in each case and was trained with the valid data (skip topics excluded). As can already be seen with the standard model, the results are also slightly better here when only the valid topics are considered. Furthermore, it can be seen that the results of the data split at the image level are roughly comparable with those at the topic level. This indicates that no overfitting takes place and that the features of certain topics are not learned by heart. With the split at topic level, it can be seen that when predicting unknown data (*topics* = test), approximately the same precision is achieved as when predicting data that has already been partially seen in training (*topics* = all).

Compared to the standard model, the results of the NeuralNet model in Figure 10 show that separate distributions can be seen for all three classes. The majority of the non-argumentative images are found on the left of the plots and the strongly argumentative images on the right. Thus, good Precision@k results can be expected even for larger k values.



**Figure 10:** Histograms for the argument NeuralNet model over a selected section of valid topics with the distributions of the scores assigned by the model and the calculated Precision@20 values. The topics shown were not trained by the model, but are used exclusively as test data. PBoth@20 shows the determined precision for weakly and strongly argumentative images. With PStrong@20, only strongly argumentative images are considered. ImageCount shows the absolute number of images that are available in the data set for the category ("Strong" or "Both") under consideration.

### 6.3. Evaluation Stance Model

Analogous to the argument NeuralNet model, the stance NeuralNet model should now also be evaluated. For this purpose, all valid topics are first searched for, whereby seven topics are no longer considered. The metric accuracy is used for the stance NeuralNet model to measure the performance, since the images are to be classified into the classes "pro", "con" and "neutral".

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

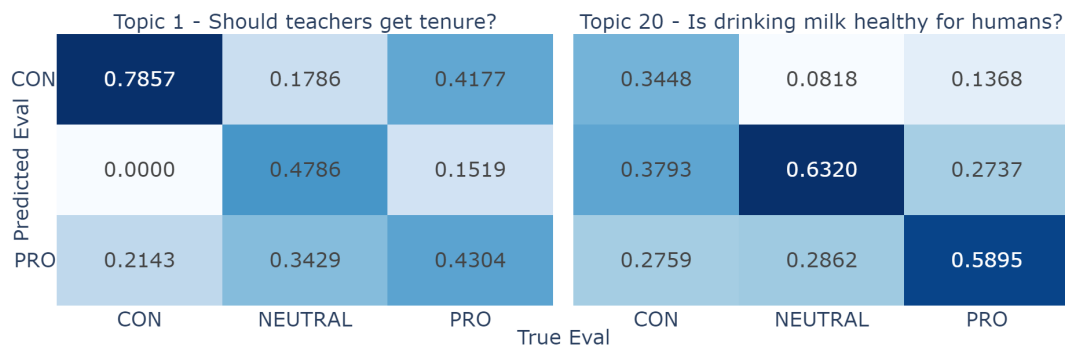
In this application, TP (True Positive) and TN (True Negative) represent the number of images that were correctly classified. The number of incorrectly classified images, FP (False Positive) and FN (False Negative), are added to the denominator of the sum, which means that the total number of all classified images is used here. The quotient of both sums calculates the accuracy.

Data Split	Topics	Accuracy
Image Level	all	0.4723
Image Level	valid	0.4961
Topic Level	all	0.4812
Topic Level	test	0.4215

**Table 4**

Results of the stance NeuralNet model. The table shows the calculated accuracy values. A distinction is made as to whether a data split was applied at image level or at topic level, and whether all topics, the valid topics or only the test topics were taken into account.

Similar conclusions as for the argument models can be drawn from the results in table 4 for the stance NeuralNet model. The model performs better with the split at image level if only valid topics are considered. This is not the case with the split at topic level. Here, the model is 6



**Figure 11:** Confusion matrices for the stance NeuralNet model over a selected section of valid topics with the predicted classes and the calculated accuracy values. The topics shown were not trained by the model, but are used exclusively as test data.

percentage points better when all topics are considered than when predicting only unknown topics. This difference could be due to the variance of the trained models. However, it is also conceivable that the model increasingly classifies "neutral", which results in a higher accuracy for the non-valid topics, which achieves a noticeable difference on average across all topics. A slight overfitting would also be conceivable.

Figure 11 shows the results of the out-of-sample prediction on the test data in the form of confusion matrices. Ideally, a dark blue diagonal line from top-left to bottom-right should be visible. This would be the case if the actual stance of the images corresponds to the majority of the predicted stance. This diagonal is not visible in the plots shown. Mostly one or two of the classes "pro", "con" or "neutral" are predicted well, but it is not possible for the neural network to predict all stances of a topic well. It can also be seen that "neutral" is classified more often, which can explain the differences in the table 4 at topic level. It can be seen that the stance model is not able to distinguish "pro" from "con". The features do not seem to make it possible to establish a connection between query and stance. The results are significantly worse than those of the argument model.

#### 6.4. Evaluation of the overall system

First of all, it should be noted once again that the focus of this work was on the argument model. No solution was found to improve the stance model and bring it to a similar quality. Since the overall system is based on both models, the results of the overall system are not satisfactory due to a poor stance model.

Table 5 shows that the system offers on average only 5 strongly argumentative images among the first 20 images viewed for a search query, which were also assigned to the correct stance ("pro" or "con"). If the weakly argumentative ones are also evaluated, the value increases to 8 out of 20. A more precise evaluation of the overall system is considered to make little sense, since the main focus was on the argument model. The stance model is clearly the limiting factor for a better precision here.

	Precision
Strong@20	0.2510
Both@20	0.4030

**Table 5**

Results of the overall system consisting of the argument NeuralNet model and stance NeuralNet model trained with all topics and a data split at the image level. A Precision@20 was calculated for strong argumentative images only, as well as for strong and weak argumentative images.

## 6.5. External evaluation with Tira

In addition to our own evaluation, another external evaluation of the approaches via Tira took place. Tira is a software that tries to solve the problem of reproducibility of scientific work, especially for shared tasks [29]. Table 6 and table 7 show the results for the submitted runs. The four runs each come from combining one of the two argument models (NeuralNet or standard) with one of the two stance models (NeuralNet or standard).

ID	Tira Timestamp	Models used
1	2022-02-25-11-07-15	NeuralNet argument model and NeuralNet stance model
2	2022-02-25-11-49-41	NeuralNet argument model and standard stance model
3	2022-02-25-19-11-54	standard argument model and NeuralNet stance model
4	2022-02-25-09-41-56	standard argument model and standard stance model

**Table 6**

The table shows the Tira timestamps and the argument and stance models (standard or NeuralNet) used in each case.

ID	Topic	Argument	Stance	Argument (adj)	Stance (adj)
Minsc - Baseline	0.736	0.686	0.407	0.932	0.553
1	0.673	0.624	0.354	0.927	0.526
2	0.687	0.632	0.365	0.920	0.531
3	0.664	0.609	0.344	0.917	0.518
4	0.701	0.634	0.381	0.904	0.544

**Table 7**

Results of the Tira runs compared to the Minsc baseline. The quality of topic relevance, argumentativeness and stance relevance is calculated with a Precision@10. To relate the values to the evaluation shown in the previous subsections, the argumentativeness and stance relevance have to be adjusted by dividing them with the topic precision. This is indicated by the addition of "adj".

To measure the performance, a Precision@10 was calculated for topic relevance, argumentativeness, and stance relevance. Comparing only these precision values with the baseline, they are significantly lower than the precision values for the argument and stance model from the previous subsections. This is mainly due to the fact that the topic model was not evaluated for this work and therefore only topic relevant images were used for the evaluation of the argument and stance model. For better comparability, the Tira Precision@10 values for the argument and stance model were additionally adjusted by dividing the respective value by the topic precision. This is possible because there is a continuous dependency between the values. This means that

an image can only be argumentative if it is also topic relevant and an image can only have a positive or negative stance value if it is also argumentative.

If we now look at the adjusted values, we can see, as in the previous evaluation, that the NeuralNet argument model perform better than the standard argument model, although the deviations between the different runs are small. Furthermore, the four runs with the adjusted precision values achieve about the same high performance as the Minsc baseline. The opposite is evident in the stance models. Here, the standard stance model performs better than the NeuralNet, although the precision can generally be rated as low. Thus, the Tira results also confirm the conclusions of the evaluations in the previous subsections.

## 7. Conclusion

The work has shown that it is in principle possible to recognise argumentative images. Two feature-based approaches were tested. Both involve information derived exclusively from images. These include colour, image text and structural features such as the recognition of diagrams. The standard argument model, which is based on a formula, achieves worse results in this work than a trained neural network with the same features. This is because too many assumptions have to be made in a single formula to accommodate the complexity of an image. The interaction of colours and text cannot be considered because each feature is added up individually and as part of the formula. The neural network, on the other hand, achieves remarkable results and can deliver on average more than 17 matching images among the first 20 results for a search query. Among them, more than 10 images are even strongly argumentative. As the number of search results considered increases, the neural network also performs significantly better than the standard model. It has been shown that not all information on the images is equally important for the network. The presence of text seems to be particularly decisive for an argumentative image. However, this should be seen as positive, since according to the considerations in chapter 2.2, an image can only be argumentative in interaction with text or diagrams.

The word count of the recognised text is the most important feature for the argument model. However, this feature is based on the unreliable text recognition of Tesseract, which can neither recognise handwriting nor low-contrast writing. Perhaps an improvement of this optical character recognition technique would also lead to an improvement of the model.

In further work, other features could be integrated and examined for their usefulness. It might be conceivable to recognise simple symbols, which could indicate an argumentative character.

As a result of this work, in addition to the argument model, the analysis and labeling of the data set provided for the Touché Lab task should be mentioned. There it was shown that many topics are not suitable for evaluating an argument search engine. Too few images are actually argumentative (46%). Furthermore, 28% of the images that were indicated as topic relevant are not topic relevant. With actually only 13% strongly argumentative and at the same time topic relevant images, the data set is only of very limited use for a good training of a neural network and for evaluation for this task. A better data set could possibly lead to better results here as well.



Even though it was not the focus of this work, some attempts were made to create a stance model. However, only slightly less than 50% of the images are correctly classified as "pro", "neutral" and "con". This is probably due to the distinction between "pro" and "con". The neural network was not able to establish the connection between the given query and the image. Even after including the HTML pages of the images, the performance improved only slightly. This classification requires a deep understanding of the question, including negations and other rhetorical devices. For this reason, the overall system cannot achieve good results. On average, only 5 of the first 20 images in a search query are strongly argumentative and assigned to the correct side ("pro" or "con") and thus of great interest to the user. Improvements must therefore be made primarily to the stance model for a good overall system. One possibility would be to use a language model such as BERT [30]. This might make it possible to process the complexities of the language and establish a connection between query and text.

## Acknowledgments

We would like to thank all the people who helped with the evaluation. Without them, we would not have been able to use nearly 10,000 labeled images as a data set. We would like to thank Lena, Yasmin, Sören and Roman. Additionally, we would like to thank Theresa Elstner for the detailed and fast review of our paper.

## References

- [1] Y. Ajjour, H. Wachsmuth, J. Kiesel, M. Potthast, M. Hagen, B. Stein, Data acquisition for argument search: The args.me corpus, in: *KI 2019: Advances in Artificial Intelligence*, Springer International Publishing, 2019, pp. 48–59. URL: [https://doi.org/10.1007/978-3-030-30179-8\\_4](https://doi.org/10.1007/978-3-030-30179-8_4). doi:10.1007/978-3-030-30179-8\_4.
- [2] R. Levy, B. Bogin, S. Gretz, R. Aharonov, N. Slonim, Towards an argumentative content search engine using weak supervision, in: *COLING, 2018*, pp. 2066–2081. URL: <https://aclanthology.org/C18-1176/>.
- [3] C. Stab, J. Daxenberger, C. Stahlhut, T. Miller, B. Schiller, C. Tauchmann, S. Eger, I. 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*, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 21–25. URL: <https://aclanthology.org/N18-5005>. doi:10.18653/v1/N18-5005.
- [4] H. Wachsmuth, M. Potthast, K. A. Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, B. Stein, Building an argument search engine for the web (2017). URL: <https://doi.org/10.18653/v1/w17-5106>. doi:10.18653/v1/w17-5106.
- [5] M. Champagne, A.-V. Pietarinen, Why images cannot be arguments, but moving ones might, *Argumentation* 34 (2019) 207–236. URL: <https://doi.org/10.1007/s10503-019-09484-0>. doi:10.1007/s10503-019-09484-0.

- [6] J. E. Kjeldsen, The rhetoric of thick representation: How pictures render the importance and strength of an argument salient, *Argumentation* 29 (2014) 197–215. URL: <https://doi.org/10.1007/s10503-014-9342-2>. doi:10.1007/s10503-014-9342-2.
- [7] D. Bundestag, Wirksamkeit von bildlichen warnhinweisen auf zigarettenpackungen (2017). URL: <https://www.bundestag.de/resource/blob/511122/8ae51b807ef2d0ebd58e4f4747c4bee7/wd-5-024-17-pdf-data.pdf>. doi:10.1007/s10503-014-9342-2.
- [8] N. Busca, How a horrifying cycling crash set up a battle over safety, 2021. URL: <https://www.nytimes.com/2021/01/30/sports/cycling/riders-crashes-uci-safety.html>.
- [9] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, B. Stein, Computational argumentation quality assessment in natural language, in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, 2017. URL: <https://doi.org/10.18653/v1/e17-1017>. doi:10.18653/v1/e17-1017.
- [10] M. Potthast, L. Gienapp, F. Euchner, N. Heilenkötter, N. Weidmann, H. Wachsmuth, B. Stein, M. Hagen, Argument search, in: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2019. URL: <https://doi.org/10.1145/3331184.3331327>. doi:10.1145/3331184.3331327.
- [11] J. Kiesel, N. Reichenbach, B. Stein, M. Potthast, Image retrieval for arguments using stance-aware query expansion, in: *Proceedings of the 8th Workshop on Argument Mining*, Association for Computational Linguistics, 2021. URL: <https://doi.org/10.18653/v1/2021.argmining-1.4>. doi:10.18653/v1/2021.argmining-1.4.
- [12] A. Latif, A. Rasheed, U. Sajid, J. Ahmed, N. Ali, N. I. Ratyal, B. Zafar, S. H. Dar, M. Sajid, T. Khalil, Content-based image retrieval and feature extraction: A comprehensive review, *Mathematical Problems in Engineering* 2019 (2019) 1–21. URL: <https://doi.org/10.1155/2019/9658350>. doi:10.1155/2019/9658350.
- [13] H. Shao, Y. Wu, W. Cui, J. Zhang, Image retrieval based on MPEG-7 dominant color descriptor, in: *2008 The 9th International Conference for Young Computer Scientists*, IEEE, 2008. URL: <https://doi.org/10.1109/icycs.2008.89>. doi:10.1109/icycs.2008.89.
- [14] M. Solli, R. Lenz, Color emotions for multi-colored images, *Color Research & Application* 36 (2011) 210–221. URL: <https://doi.org/10.1002/col.20604>. doi:10.1002/col.20604.
- [15] V. Mokshin, I. Sayfudinov, S. Yudina, L. Sharnin, Object detection in the image using the method of selecting significant structures, *International Journal of Engineering & Technology* 7 (2018) 1187. URL: <https://doi.org/10.14419/ijet.v7i4.38.27759>. doi:10.14419/ijet.v7i4.38.27759.
- [16] J. K. L. and, Image classification and object detection algorithm based on convolutional neural network, *Science Insights* 31 (2019) 85–100. URL: <https://doi.org/10.15354/si.19.re117>. doi:10.15354/si.19.re117.
- [17] D. Fleming, Can pictures be arguments?, *Argumentation and Advocacy* 33 (1996) 11–22.
- [18] J. E. Kjeldsen, Virtues of visual argumentation: How pictures make the importance and strength of an argument salient, 2013.
- [19] M. Meharban, D. Priya, A review on image retrieval techniques, *Bonfring International Journal of Advances in Image Processing* 6 (2016) 07–10. URL: <https://doi.org/10.9756/bijaip.8136>. doi:10.9756/bijaip.8136.

- [20] M. Honnibal, I. Montani, spacy - industrial-strength natural language processing in python, 2022. URL: <https://spacy.io/>.
- [21] M. Potthast, L. Gienapp, F. Euchner, N. Heilenkötter, N. Weidmann, H. Wachsmuth, B. Stein, M. Hagen, Argument search: Assessing argument relevance, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (2019) 1117–1120. doi:10.1145/3331184.3331327.
- [22] R. Smith, An overview of the tesseract OCR engine, in: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2, IEEE, 2007. URL: <https://doi.org/10.1109/icdar.2007.4376991>. doi:10.1109/icdar.2007.4376991.
- [23] V. Bonta, N. Kumares, N. Janardhan, A comprehensive study on lexicon based approaches for sentiment analysis, Asian Journal of Computer Science and Technology 8 (2019) 1–6. URL: <https://doi.org/10.51983/ajcst-2019.8.s2.2037>. doi:10.51983/ajcst-2019.8.s2.2037.
- [24] M. Zaid, L. George, G. Al-Khafaji, Distinguishing cartoons images from real-life images, International Journal of Advanced Research in Computer Science and Software Engineering 5 (2015) 91–95.
- [25] user12526469 nathancy, How to detect diagram region and extract(crop) it from a research paper's image, 2022. URL: <https://stackoverflow.com/a/59315026>.
- [26] R. A. Wagner, M. J. Fischer, The string-to-string correction problem, Journal of the ACM 21 (1974) 168–173. URL: <https://doi.org/10.1145/321796.321811>. doi:10.1145/321796.321811.
- [27] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, Iryna Gurevych, The inception platform: Machine-assisted and knowledge-oriented interactive annotation, Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations (2018) 5–9. URL: <https://aclanthology.org/C18-2002/>.
- [28] M. Berrendorf, E. Faerman, L. Vermue, V. Tresp, On the ambiguity of rank-based evaluation of entity alignment or link prediction methods, 2020. URL: <https://arxiv.org/pdf/2002.06914>.
- [29] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1\_5.
- [30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.

## A. Reproducibility

Our complete code basis and our results of the labeling process can be found in our GitLab repository <sup>1</sup>. We used the dataset from the Touché 2022 Task 3: Image Retrieval for Arguments. It can be found on the official task website <sup>2</sup>.

## B. Complete analysis of the labeled data set

Topic number	Topic Relevance	Argumentativeness	Arg. (Strong)	Stance Pro	Stance Con	Relevant	Relevant (Strong)
1	0.5 (247)	0.36 (178)	0.11 (52)	0.16 (79)	0.06 (28)	0.22 (106)	0.1 (47)
2	0.85 (466)	0.73 (400)	0.11 (60)	0.48 (266)	0.17 (91)	0.64 (353)	0.1 (54)
4	0.64 (294)	0.58 (268)	0.15 (70)	0.05 (22)	0.21 (99)	0.26 (121)	0.15 (68)
8	0.68 (333)	0.61 (300)	0.14 (68)	0.3 (149)	0.14 (69)	0.44 (217)	0.12 (58)
9	0.78 (380)	0.21 (103)	0.1 (48)	0.08 (38)	0.08 (41)	0.16 (76)	0.07 (33)
10	0.79 (348)	0.72 (317)	0.14 (60)	0.53 (231)	0.13 (58)	0.66 (289)	0.12 (53)
15	0.87 (466)	0.54 (290)	0.18 (97)	0.03 (15)	0.3 (160)	0.32 (173)	0.16 (88)
20	0.77 (393)	0.34 (171)	0.19 (95)	0.19 (95)	0.06 (30)	0.24 (124)	0.14 (71)
21	0.34 (157)	0.29 (135)	0.14 (64)	0.18 (83)	0.05 (23)	0.23 (106)	0.12 (58)
22	0.65 (263)	0.42 (169)	0.05 (19)	0.1 (41)	0.06 (24)	0.16 (65)	0.05 (19)
27	0.93 (447)	0.88 (421)	0.37 (177)	0.26 (123)	0.34 (165)	0.6 (288)	0.35 (170)
31	0.76 (402)	0.63 (333)	0.21 (110)	0.45 (241)	0.03 (14)	0.48 (253)	0.18 (94)
33	0.61 (345)	0.57 (317)	0.22 (121)	0.4 (222)	0.1 (57)	0.5 (279)	0.21 (118)
36	0.79 (413)	0.14 (75)	0.07 (35)	0.11 (60)	0.02 (10)	0.13 (70)	0.06 (32)
37	0.79 (367)	0.46 (213)	0.16 (76)	0.01 (6)	0.3 (138)	0.31 (144)	0.15 (72)
40	0.79 (392)	0.49 (245)	0.09 (45)	0.14 (71)	0.32 (157)	0.46 (227)	0.06 (32)
43	0.93 (438)	0.26 (121)	0.1 (45)	0.18 (84)	0.04 (17)	0.21 (101)	0.07 (35)
45	0.48 (174)	0.09 (31)	0.05 (18)	0.02 (7)	0.06 (21)	0.08 (28)	0.04 (16)
47	0.69 (329)	0.25 (119)	0.09 (41)	0.04 (20)	0.2 (98)	0.2 (98)	0.06 (27)
48	0.58 (193)	0.56 (186)	0.22 (74)	0.26 (88)	0.05 (15)	0.31 (103)	0.21 (69)

**Table 8**

Results of the analysis of the data set for all labeled topics. The addition "strong" means that only strong argumentative images are used here, while otherwise strong and weak argumentative images are considered. Relevant images are images which are topic relevant, argumentative as well as stance-relevant.

<sup>1</sup><https://git.informatik.uni-leipzig.de/jb64vyso/aramis-image-argument-search>

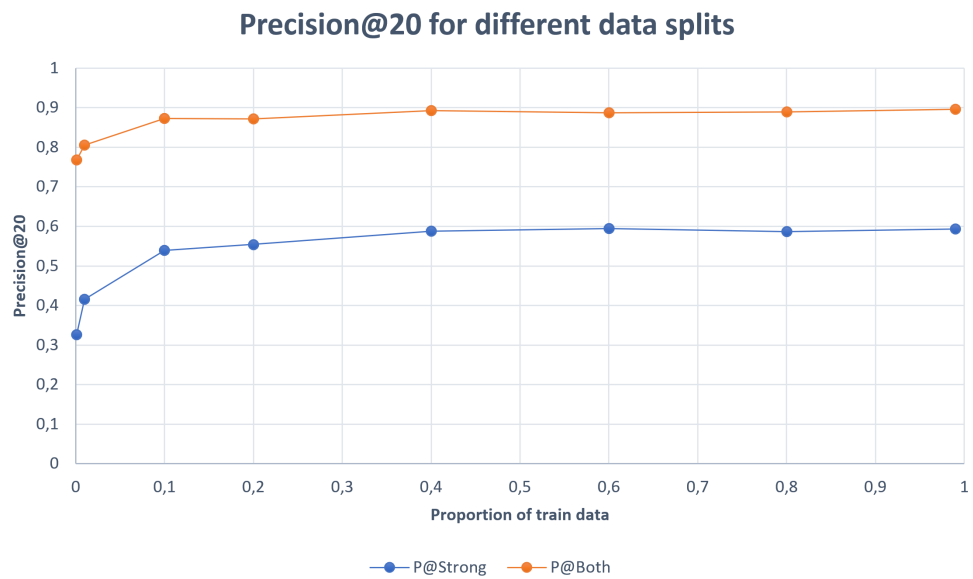
<sup>2</sup><https://webis.de/events/touche-22/shared-task-3.html>

Topic number	Topic name
1	Should teachers get tenure?
2	Is vaping with e-cigarettes safe?
4	Should corporal punishment be used in schools?
8	Should abortion be legal?
9	Should students have to wear school uniforms?
10	Should any vaccines be required for children?
15	Should animals be used for scientific or commercial testing?
20	Is drinking milk healthy for humans?
21	Is human activity primarily responsible for global climate change?
22	Is a two-state solution an acceptable solution to the Israeli-Palestinian conflict?
27	Should more gun control laws be enacted?
31	Is obesity a disease?
33	Should people become vegetarian?
36	Is golf a sport?
37	Is cell phone radiation safe?
40	Should the death penalty be allowed?
43	Should bottled water be banned?
45	Should the penny stay in circulation?
47	Is homework beneficial?
48	Should the voting age be lowered?

**Table 9**

Assignment of topic numbers to topic names for all labeled topics.

## C. Results for different data splits



**Figure 12:** Precision@20 values for the argument NeuralNet model with image level data split, with the x-axis showing the proportion of training data. All data points represent an arithmetic mean over the results of 10 trained models. It can be seen that already from a training share of the data of 10% no remarkable changes in the model performance can be recognized.