Computational Empirical Research: Automated Visual Content Analysis

Tobias Schreieder[1] and Jan Philipp Zimmer[1]

Leipzig University, Leipzig 04109, Germany
{fp83rusi,uw08qebe}@studserv.uni-leipzig.de

Abstract. The task of image classification is becoming increasingly important in data analysis and scientific research due to the large amount of image data. Araujo et al. [2] have developed a comprehensive framework called AVCA for communication research purposes. The idea is to use commercial computer vision models to automatically label images. The labels should describe the image textually. Subsequently, the labels can be used as input data for machine classification models. In this paper, the AVCA framework will first be introduced and quality implications identified. In the following, the reproducibility and generalizability of AVCA will be tested on an own case study called "Image Classification for Argumentation". For this purpose, AVCA is implemented for the Touché'23 data set and solutions for the quality implications are proposed. The background of the case study is to assign argumentative images to their corresponding topic, with different binary classification models for each topic. The results of the case study show that the best-performing models achieve an average f1-score of 0.78 across the three topics studied.

Keywords: argumentation · computer vision · image classification

1 Introduction to Automated Visual Content Analysis

Over the last few years, images have become increasingly prominent in internet communication. Social media platforms such as Instagram, for example, focus primarily on images. Companies often rely on images to convey advertising messages as well. For these reasons, there is a great need for machine learning image classification methods to reliably process large amounts of image data. Research on such methods is also growing in the field of communication research. A first all-encompassing classification framework called "Automated Visual Content Analysis" (AVCA) was presented by Araujo et al. [2]. in 2020. Their idea is to initially annotate images using commercial pre-trained computer vision models. Computer vision models try to describe an input image as accurately as possible using short words or word sequences (labels). In most cases, these labels are additionally assigned a score that describes the likelihood of the label. Since training your own computer vision models requires a lot of training data and programming knowledge, pre-trained models with API access can make it easier for researchers to use them.

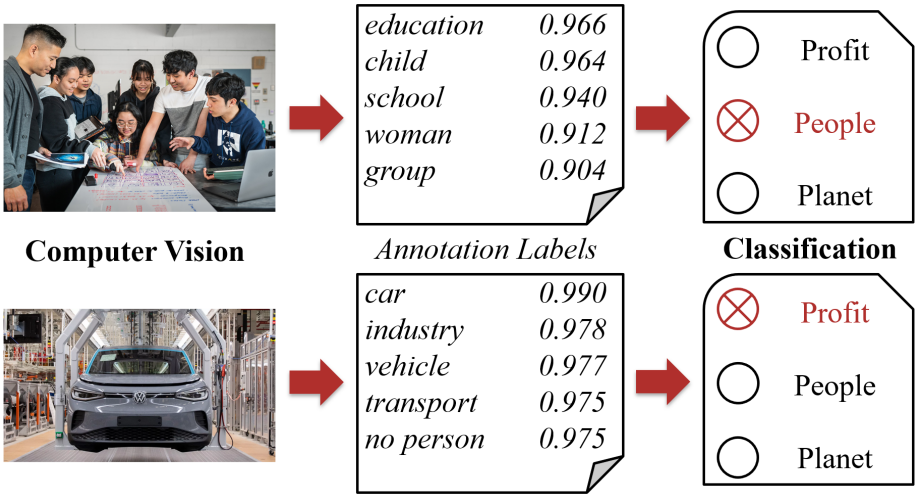


Fig. 1. Example illustration of the image classification process with queried computer vision labels for specific website images. The three class classifications into "profit", "planet" and "people" according to Araujo et al. [2] are shown. The two sample images were crawled from corporate websites as examples. The upper image is from Apple Inc. [1] (07/10/2023) and the lower image is from Volkswagen Germany [21] (07/10/2023).

Figure 1 illustrates this as an example based on the case study presented by Araujo et al. [2]. The authors' intention was to determine whether companies adhere to a tripartite relationship between people, planet and profit in the sustainability communication on their corporate websites. The foundation for this is formed by the "Sustainable Development Goals" established by the United Nations, in which companies are attributed a significant responsibility for people, planet and profits [2, 10, 20]. Whether companies comply with an equal treatment of these three goals was to be analyzed by applying the AVCA framework for corporate communication with images on the corresponding websites. For this purpose, separate binary classification models were trained for each class of people, planet and profit, using the annotation labels of the computer vision models as input parameters [2].

In this work, a brief description and analysis of the paper by Araujo et al. [2] will be given. In section 2, precursors as well as further developments and alternative approaches to image classification will be presented. Subsequently, in section 3 the methodology of the AVCA framework is described in more detail. Through the analysis of the paper, some quality implications have become visible. These will be discussed in section 4. In the following, in section 5 an attempt is made to transfer the framework to another use case. For this purpose, we apply AVCA within the case study "Image Classification for Argumentation" for the classification of argumentative images. Furthermore, an attempt is made to solve the quality implications mentioned above. Following the case study, in

section 6 further applications and suggestions for improvement are proposed. Finally, a short conclusion will be drawn in section 7.

2 Related Work

The last two decades were marked by the ever-growing amount of content, and therefore data, that emerged on various media channels. In order to comprehend and learn from this data, researchers are in need of scalable analysis methods [19]. Computer vision models based on artificial neural networks (ANN) are designed to perform complex classification tasks [14]. Other applications include the detection of objects [8, 18] or the creation of captions for images [15]. Thus, ANNs are one major scientific tool for large-scale analytical tasks. Savage [17] pointed out, that in early 2016 their capability of describing image contents could already be compared to that of a three-year-old.

As of today, commercial models from tech companies like Google, Microsoft and Clarifai are publicly available to perform various high quality computer vision tasks. With this development in mind, Araujo et al. [2] developed a framework to utilize these on demand tools for answering scientific questions in communication research. The main idea is to reuse high quality models, although they are not tailored to a specific task or question, thus reducing the workload for both, scholars and computing systems. Araujo et al. [2] argue, that scholars with little experience in computer science and machine learning shall be able to perform complex analyses of large image data sets.

The work of Araujo et al. [2] has already been employed for further studies. These include the development of another framework called "Automatic Image Content Extraction" (AICE) by Männistö et al. [16], which takes a closer look at specific machine learning techniques and applicability in a broader range of domains. Moreover, Himelboim et al. [11] use AVCA for studying brand-related communication in social media. Iadanza et al. [12] show that the method can also be transferred to an application, that classifies images of hospital settings, and underline that especially the API training performed very promising.

3 Methodology of the AVCA Framework

Araujo et al. [2] propose a framework with the aim of enabling communication researchers to use commercial image recognition tools for automated visual content analyses. The framework is mainly based on the four steps suggested by Duriau et al. [9]: identification of data sources, data collection, coding and analysis of content and interpretation of results. Additionally, a sample study is conducted to show the effectiveness of the framework. This section discusses the methods that were used by Araujo et al. [2] in their sample study, as well as the measures to verify these methods.

The step identification of data sources mainly deals with the development of actual research questions. For the data collection, Araujo et al. [2] make use of

web-scraping mechanisms. Furthermore, the annotation of the images is carried out by two of the authors manually.

The analytical step of the framework is composed of two main parts: the generation of labels for images with the help of computer vision APIs and the classification of the images with respect to these labels. For the first part, the images are fed to three commercial image recognition APIs of large companies. The results contain labels of concepts present in each image. Additionally, each label is annotated with a probability, representing the likelihood of the concept to in fact be present in the image.

For the classification of the labels, three different approaches are proposed. The first one, an inductive and unsupervised approach, uses topic modelling to assign the provided labels to topics. The other approaches, both of them being deductive, use expert knowledge and supervised machine learning (ML). The ML approach divides itself again into a machine driven approach, for which different ML models are computed, as well as API training, which utilizes an API to an existing commercial classification system. A flow diagram of the analytical step is shown in figure 2.

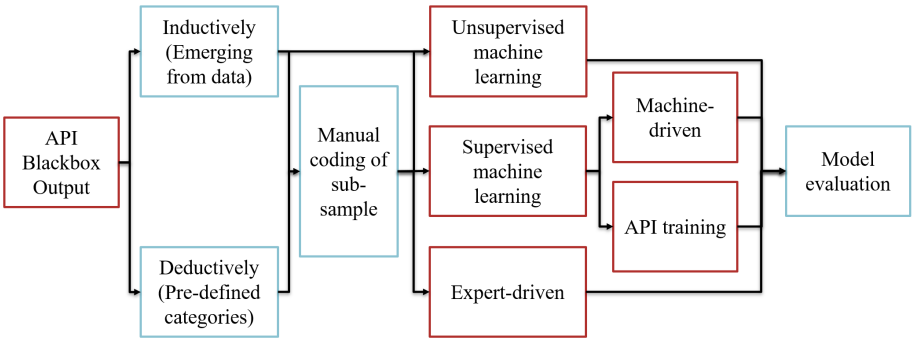


Fig. 2. Simplified representation of the AVCA framework for image classification with computer vision labels according to Araujo et al. [2]

The method performance is measured in different ways. For the coding of the image data, the average agreement of the raters and Krippendorf's alpha are considered as quality measures. The unsupervised approach is only be evaluated qualitatively, as is the expert-driven approach, while the ML approaches are evaluated regarding their precision, recall and f1-scores.

4 Quality Implications

A more detailed analysis of the work of Araujo et al. [2] has revealed some quality implications. In the following, the most important quality implications (QI) will be named and explained. In the area of classification, we restrict ourselves to the

supervised machine learning approaches, which we apply in the case study that follows in section 5. Concrete approaches for solving the quality implications are discussed in section 5.7 on the basis of the case study mentioned.

QI 1 - Calculation of Method Performance: The first quality implication relates to method performance. For evaluation, the "micro", "macro" and "weighted" f1-scores were calculated [2] with the Python package sckit-learn.[1] The micro f1-score is calculated globally, the macro f1-score forms the unweighted mean for each label and the weighted f1-score first calculates the score separately for each label and then averages them weighted into a final score. Araujo et al. [2] show their best-performing approaches separately for each class, by reducing them to a binary classification problem (e.g. people and non-people). However, for the most part, the evaluation with weighted f1-scores is shown for the supervised machine learning models. Reduced to a binary classification, this can lead to the performance of models being overestimated. This is because compared to the example above, for the class "non-people" the data set contains significantly more ground truth images since "non-people" uses the images of the classes "profit" and "planet". By weighting, a model can also achieve a good f1-score with a binary classification if the model classifies all images as "non-people". As default, scikit-learn also offers a binary f1-score for binary classifications. In order to enable a fair comparison of the models, a suitable evaluation measure must be selected in advance, which is applied equally to all models.

QI 2 - Internal Validity of AVCA Framework: The used data set contains a total of 868 annotated images. With a sample of 10%, only 87 images were used for the final testing of the procedures [2]. We find it difficult to make a statement on the internal validity with this small amount of testing data.

QI 3 - External Validity of AVCA Framework: This quality implication relates to the generalizability of the AVCA framework to other use cases and data sets from other domains. In the work of Araujo et al. [2] no explicit statements could be found on this. Besides that, the framework was demonstrated using a highly specific use case with very general classes ("people", "planet", "profit"). Accordingly, it is difficult to make a statement whether the framework can be used for image classification in general, or only for very specific, delimited use cases. Especially data sets with a large variety of different images such as photos or images with text such as memes, diagrams or screenshots of letters could lead to the classification models delivering poor classification results. One reason for this could be that the annotation labels obtained from the computer vision APIs do not describe the images well enough, or there is a lot of label overlap between different classes. We would therefore first like to question the generalizability of the framework, since the results described in the paper do not allow any conclusions to be drawn about external validity.

QI 4 - Validity of Computer Vision Labels: Another quality implication are the Computer Vision labels. Araujo et al. [2] queried and saved the labels from the three APIs from Google, Microsoft and Clarifai. However, the quality of the labels was hardly described, and the labels were not compared. The labels

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

were used directly as input for the classification models as a separate data set per API or in a combined form. Whether the combination of labels from different APIs lead to a syntactically and semantically better description of the images was not considered in the work.

QI 5 - Reproducibility of Experiments: In order to establish reliability, the experiments shown must be reproducible. In the case related to the experiments on the AVCA framework, it can be seen that only parts of the code can be found in a public repository. Also, for the data set, it was only shown how images from company websites can be crawled. However, the data set itself was not published. Another point of criticism is the missing description of the exact model designs regarding the best-performing methods shown in the result tables. A reproduction of the concrete experiments is therefore not possible. It can only be attempted to generate similar classification results with a similar model structure and another data set.

5 Case Study: Image Classification for Argumentation

In this section, selected approaches from the framework of Araujo et al. [2] will be applied to a case study. Specifically, the Touché'23 data set with topic-relevant argumentative images will be used for image classification. First, a brief overview of the case study will be given in section 5.1. Subsequently, the hypothesis and research questions will be formulated in section 5.2. Section 5.3 then presents the chosen Touché'23 data set. Section 5.4 depicts the annotation process with the customized Aramis ImArg platform. This is followed by a description of the Clarifai API request for comparing computer vision approaches in section 5.5. Section 5.6 describes the implementation of the selected image classification approaches from Araujo et al. [2] on the case study. Finally, section 5.7 and section 5.8 will first show the performance of the selected approaches and then interpret the results and point out solutions for the quality implications described.

5.1 Overview of Image Classification for Argumentation

A few years ago, discussions on social media moved from being primarily text-based towards including more images or videos. Specific platforms that focus on images, like Instagram, then became increasingly popular and still are today. For this purpose, Carnot et al. [6] pointed out that in social media discussions, people often include images to illustrate their stance and arguments on the topic in question or to support written arguments. Whether images can be "argumentative", i.e. whether they can represent arguments in their own right, remains controversial [7]. However, their usefulness for argumentative debates is obvious: Kjeldsen [13] notes that images can underpin and support arguments, clarify facts, and communicate truth more effectively than words.

In 2022 the first shared task CLEF Touché lab "Image Retrieval for Arguments" took place, which had the goal to develop special search engines for

argumentative images [4]. A search engine dedicated to retrieving relevant images on controversial topics can be useful to find images that support a person's stance on social networks or elsewhere, and to get a visual overview of the landscape of different opinions for personal reflections [6]. One of the difficulties of this task is processing the ambiguity of the different images, which can include, for example, simple photos, screenshots of studies, charts, tables, or even memes.

The results of the Touché'22 task [4] and the reproducibility study presented by Carnot et al. [6] show that current state-of-the-art approaches, which use the texts on the websites belonging to the images for indexing, can achieve good results. Carnot et al. [6] also suggest extracting texts on the images with optical character recognition methods and using them as a boost in the indexing step with BM25. However, all of these approaches have the problem that only images that have an associated website on which they are integrated can be processed. Images for which there are no matching textual descriptions on the websites are also disadvantaged by such search engines. We therefore propose to expand existing approaches with an image classification module in order to be able to also consider images that are independent of websites. Such an image classification for argumentation pipeline can also be advantageous for research purposes, for example to monitor the frequency of argumentation with regard to selected topics. For the subsequent implementation of the image classification, we decided to use the AVCA framework introduced by Araujo et al. [2] and to adapt it to the local application.

5.2 Hypothesis and Research Questions

Following, the general hypothesis of this work will be presented and discussed. In addition, research questions (RQ) related to the case study will be identified.
Hypothesis: "The AVCA framework presented by Araujo et al. [2] is reproducible and generalizable to other application domains." In particular, it will be considered whether the described methodology can be translated into a concrete implementation. On the other hand, it will also be considered whether this implementation for the case study "Image Classification for Argumentation" leads to similarly good classification results, in order to be able to conclude whether the AVCA framework can also be successfully applied to domains and data sets with images of higher ambiguity.
RQ 1: With what classification performance can argumentative images be assigned to their associated communication topic using the AVCA framework?
RQ 2: Are there quality differences between the computer vision annotation labels of Google Cloud Vision and Clarifai?
RQ 3: Can a performance boost be achieved when the annotation labels of Google Cloud Vision and Clarifai are combined?

5.3 Touché'23 Data Set

For our research into image classification for argumentation, we employ the data set of the Touché'23 shared task "Image Retrieval for Arguments" [3], which is

located at the CLEF 2023 conference. The data set is freely accessible online.[2] It contains 55,691 images of 50 controversial topics. The topics include, for example, "Do we need sex education in schools?" or "Are video games art?". The images were crawled using regular search engine queries related to the 50 topics. In addition to the image itself, the data set includes, for example, a screenshot of the web page it appeared on, the web page text, or the image's rank in the regular search engine's result list. In addition, the data set contains label and text information recognized by Google Cloud Vision[3], like label, landmark, localized object, face and safe search annotations as well as optical character recognition texts. A detailed description of the used computer vision outputs follows in section 5.5. For our analysis, we use only the computer vision label annotations as input data for the image classification models to predict the topic to which an image is assigned.

5.4 Annotation Process and Results

The Touché'23 data set contains images that were crawled by classical search engines and assigned to a topic according to the query. To check the actual quality of the images in the data set with respect to topic relevance, a sample of the images was manually annotated. For this purpose, the annotation platform Aramis ImArg introduced by Braker et al. [5] and further elaborated by Carnot et al. [6] was used. This platform was developed for the Touché'22 shared task "Image Retrieval for Arguments" and allows labeling images in terms of their topic relevance, their argumentativeness and their stance relevance. We have adapted Aramis ImArg to the Touché'23 data set on the one hand, and on the other hand consider in a simplified form only the topic relevance of the images. A short description of the topic with instructions on which images should be annotated as topic-relevant is already given in the data set and is displayed by Aramis ImArg as well. Figure 3 shows a screenshot of the revised annotation platform. Aramis ImArg can easily be started locally and used via the HTML website. The original code of Aramis ImArg is publicly available.[4]

A total of 500 images, which originated from 5 different topics, were manually labeled through the annotation process. Due to limited human resources, all images were annotated independently by only the two authors of this work. The descriptions of the topics and the instructions for annotating topic-relevant images in the data set were taken into account. Averaged across all topics, an intercoder reliability of 0.87 (average agreement) and 0.64 (Krippendorf's alpha) could be achieved. The intercoder reliability results are shown in table 1. It can be seen that with a range of 0.82 to 0.91, a high average agreement was achieved for all five topics. The lowest intercoder reliability values were found for topic 76 "Are video games art? This could be due to the greater ambiguity of the images in this topic, as well as a broader scope of interpretation for the topic query.

[2] https://webis.de/data.html#touche23-image-retrieval-for-arguments
[3] https://cloud.google.com/vision?hl=de
[4] https://github.com/webis-de/SIGIR-23

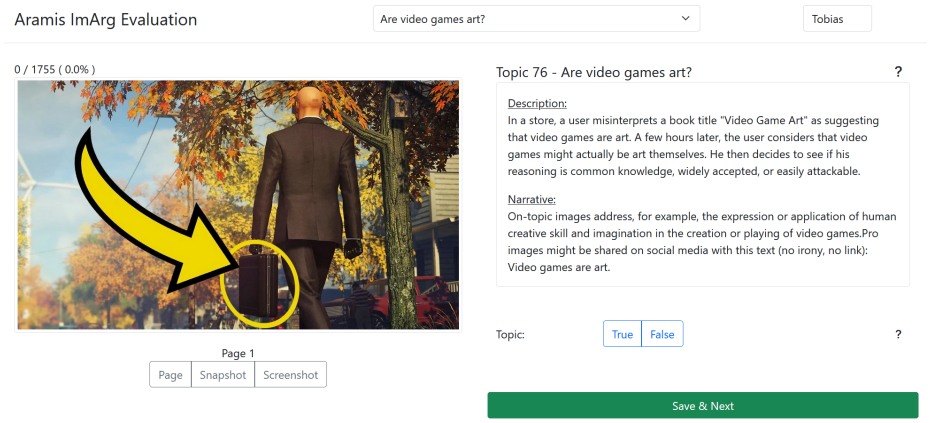


Fig. 3. Screenshot from the revised Aramis ImArg annotation platform for the chosen topic "Are video games art?". For each topic, there is a short description and instructions for labelling topic-relevant (on-topic) images.

The values for Krippendorf's alpha are, as expected, below the values of average agreement in a range of values from 0.48 to 0.69.

Topic-ID	Name	AA	α
51	Do we need sex education in schools?	0.85	0.65
55	Should agricultural subsidies be reduced?	0.90	0.65
76	Are video games art?	0.82	0.48
81	Is genetically modified food unsafe?	0.91	0.69
100	Do we need cash?	0.87	0.69
All	All	0.87	0.64

Table 1. Table with intercoder reliability for topics with manually annotated images. Intercoder reliability was calculated as both average agreement (AA) and Krippendorf's alpha (α). The "Name" column shows the name of the topic, which is also the query used to crawl the images. The row "All" shows average values across all topics.

Table 2 shows the proportion of images labelled as topic-relevant (TR) for each topic. The separate results of the two annotators are first shown with "TR A" and "TR B". It can be seen that the proportion of topic-relevant images for topic 51 is similarly high for both annotators, at 0.69 and 0.68. Topic 76, on the other hand, shows the most difference in topic-relevant images with 0.86 and 0.70, as already described for the intercoder reliability. The column "TR Both" shows the proportion of images that were annotated as topic-relevant by both annotators. According to this observation, the lowest proportion of topic-relevant images can be found in topic 51 with 0.61 and the highest in topic 81 with 0.79. For us, the results make it clear that some topics contain inappropriate

Topic-ID	TR A	TR B	TR Both
51	0.69	0.68	0.61
55	0.86	0.80	0.78
76	0.86	0.70	0.69
81	0.82	0.85	0.79
100	0.67	0.74	0.64

Table 2. Table with the percentage topic relevance (TR) of the images of a topic. "TR A" and "TR B" show the results of the individual results per person, while "TR Both" looks at all images that were labeled the same by both annotators. The row "All" shows average values across all topics.

images due to the crawling process, which may subsequently have a negative influence on the classification performance. At the same time, averaged over all annotated topics, about 70% of the images were annotated as topic-relevant by both annotators (TR Both). We consider this percentage to be high enough to use the Touché'23 data set, and therefore do not restrict ourselves exclusively to the images annotated as topic-relevant.

5.5 Computer Vision

Following the annotation process, we will now show which data is used as input for the classification models. First, the computer vision outputs from Google Cloud Vision that are available in the Touché'23 data set and already described in section 5.3 are used. In order to keep the labels of the various computer vision APIs as consistent as possible, we restrict ourselves to the label annotations and ignore the remaining data.

An exploratory data analysis on the Google Cloud Vision labels showed that a maximum of 10 labels per image were requested. All of these labels contain a name such as "book" or "eyelash" and an associated score that describes the likelihood of the label. We also checked the minimum number of labels received for an image. For example, topic 51 contains at least one image that is only described by a single label. However, no images that do not contain at least one label were found in the data set. Overall, images with few labels seem to be the exception, since the lowest amount of labels averaged over all images of a topic at 9.84 for topic 51 is very close to the maximum number of 10 labels.

In addition to the existing Google Cloud Vision labels, we also used an API request to query Clarifai's label annotations for five topics, each with 900 images. The Clarifai model "general-image-recognition" was used.[5] The same exploratory data analysis has shown significantly more consistent results for the Clarifai data set. Each image contains exactly 20 labels consisting of name and score. To be able to create consistency with the Google Cloud Vision data, we limit the Clarifai data to the 10 labels with the highest scores.

In order to make a statement about the similarity of the labels between the Google Cloud Vision and the Clarifai data set, a syntactic matching was carried

[5] https://clarifai.com/clarifai/main/models/general-image-recognition

Topic-ID	Google	Clarifai	Exact Matches	Levenshtein Matches
51	624	687	140	194
55	646	691	153	218
76	717	784	170	249
81	815	868	202	265
100	731	727	151	220

Table 3. Table with the results of the label matching process. The columns "Google" and "Clarifai" show the number of unique (duplicate-reduced) labels per topic. The "Google" labels correspond to the labels in the Touché'23 data set. For both data sets, the number of exact matches and the number of matches with a Levenshtein similarity greater than or equal to a set threshold (t=0.8) were calculated.

out. First, all labels of the images of a topic were merged and duplicates were removed. On the one hand, an exact match approach was chosen for the unique labels, which marks labels as a match if they are identical. On the other hand, a Levenshtein similarity between the labels of both data sources of a topic was calculated. A match occurs when the Levenshtein similarity is greater than or equal to a threshold of 0.8. The Levenshtein similarity should also recognize matches when two labels are syntactically similar but are not identical, for example due to a different spelling or tense.

The label matching results are shown in table 3. The duplicate-reduced labels for each topic are shown in the "Google" and "Clarifai" columns. It can be seen that, except for topic 100, Clarifai was able to generate slightly more unique labels. However, the number of unique labels is very similar between the two data sources, with a maximum difference of 67 for topic 76. Thus, both APIs describe the images of a topic with a similar number of labels. The number of exact matches ranges from 140 to 202, while for the Levenshtein similarity 194 to 265 matches could be found. Thus, only a small fraction of just over 20% labels of both data sources show a syntactic similarity. For our further analyses, a combined data set with the Google Cloud Vision and Clarifai labels is therefore created in addition to the separate data sets. If the syntactically heterogeneous labels also show a semantic heterogeneity, this combination could describe the images better and more comprehensively. Overall, this could be reflected in a higher classification performance for the combined data set.

5.6 Image Classification

Our classification process is strictly oriented on the supervised machine learning approaches of Araujo et al. [2]. Two different subsets of the Touché'23 data set were created for training and evaluation of the classifiers. The first one consists of 900 images out of the topics 55, 76 and 81, which showed both good topic relevance and intercoder reliability. The amount of three topics was adapted from Araujo et al. [2], who also used three different categories for their classifications ("people", "planet" and "profit"). To examine whether the approach can also handle more categories, a second subset includes 900 images of each of the five

topics from table 1. Out of the first subset, three actual data sets were created, one including the up to 10 labels retrieved from the Google API, one including the up to 10 labels of the Clarifai API and one containing the combined up to 20 labels of both APIs. However, during classification, each data set is processed once with a binary coding, indicating whether a specific label is present in an image or not, and once with the likelihood values for each label that were provided by the respective API. For the second subset, only the data sets with Google labels are created and likewise processed both, with binary and likelihood coding. This results in eight different data sets as input for the classification task.

For the classification process itself, a wrapper class is implemented. This class serves for the purposes of training and evaluation of four different types of classifiers, which are selected from the Python package scikit-learn[6]. This selection is again based on the best resulting classifiers in the underlying paper of Araujo et al. [2]. Specifically the Support Vector Classifier (SVC), Gradient Boosting Classifier (GBC), Stochastic Gradient Descent (SGD) and Passive Aggressive Classifier (PAC) can be processed by this class. To assess the hypothesis and answer the research questions from section 5.2, the relevant information about each trained classifier is saved. This information includes: name of the model, model parameters, name of the data file, the predicted topic, whether likelihood or binary coding was used, the best f1-score during training and the evaluation metrics of the test set, namely accuracy, precision, recall, binary f1-score and weighted f1-score.

5.7 Evaluation

In this section, we present the results of the classification task. First, the three-topic data sets are discussed before a brief look is taken into the five-topic case.

Derived from Araujo et al. [2], all data set are split into a training, validation and test set. 90 percent of the data is used for the training and validation, while 10 percent of the data are withheld and utilized for testing the models' performance on previously unseen data. To ensure reproducibility, the split is performed with a seeding-mechanism.

Various configurations are then trained for each classifier, using a grid search with five-fold cross-validation. We compute 36 SVMs, 18 SGDs, 27 PACs and 24 GBCs per configuration, resulting in an amount of 105 models with different hyperparameters. The varied hyperparameters for each classifier are shown in table 4. For an in-depth description of the hyperparameters as well as others available, we refer to the official documentation of scikit-learn. One grid search is carried out per classifier. For each run, the best model with regard to its binary f1-score is saved for further evaluation. All models are computed for likelihood and binary coded data sets. Moreover, they are computed for each included topic as the predicted variable, resulting in 630 models per data set. This adds up to a total amount of 1890 computed models for three-topic data sets of which the best 72 are saved. In addition to the binary f1-score of the validation results,

[6] https://scikit-learn.org/stable/

precision, recall and binary f1-score with respect to the test set are calculated and saved for further evaluation.

Classifier	Hyperparameters
SVC	C, kernel, degree, gamma
SGD	loss, penalty, alpha, learning rate
PAC	C, fit intercept, max iter, loss
GBC	loss, learning rate, n estimators

Table 4. The hyperparameters that were modified during the grid search for each classifier. Note that the SVC degree parameter was only used for the "poly" kernel.

The overall best results are achieved by two SVC models and an SGD model, all employing the combined data set of Google and Clarifai labels. For the SVCs both, the likelihood and the binary coded data set, result in a binary f1-score of 0.85 after validation and 0.89 on the test set. With a binary f1-score after validation of 0.84 and a test binary f1-score of 0.89, the SGD classifier on the likelihood coded data set indicated the third best results. The best resulting model, that was not trained on the combined data set, is also an SGD classifier with a binary f1-score of 0.83 after validation and 0.85 on the test data. This model was trained on the binary coded Google data set.

To assess the quality of the data sets, we compute the mean scores per data set. They are displayed in table 5. The combined data set shows the best average performance across all metrics, followed by the Google data set. The Clarifai data sets evaluation scores remain slightly lower.

Data Set	F1 Val.	Precision	Recall	F1
Google	0.74	0.81	0.69	0.74
Clarifai	0.72	0.76	0.65	0.70
Combined	0.76	0.81	0.72	0.77

Table 5. The mean scores of each metric per data set. "F1 Val." (F1 Validation) was computed during training. "Precision", "Recall" and "F1" were computed with the unseen test set after the training was finished.

In addition to the rather general statements of the overall data set performance, a closer look is taken at the influence of the use of a likelihood coding in the data set as well as the predicted topic. While no greater varieties appear dependent on the use of the likelihood coding in the data set, the predicted topic has a large impact on the outcome of the evaluation scores. The models predicting topic 76 show significantly larger scores compared to topic 55 and 81 with a validation binary f1-score of 0.85 in comparison to 0.77 and 0.72 respective 0.73. The scores computed with the test set indicate equal results, with the precision score of topic 76 having the greatest difference to the other precision scores.

The comparison of the evaluation metrics dependent on the binary or likelihood coded data set and predicted topic are depicted in table 6.

Likelihood	Topic	F1 Val.	Precision	Recall	F1
False	55	0.77	0.77	0.75	0.76
	76	0.85	0.97	0.82	0.89
	81	0.72	0.73	0.68	0.71
True	55	0.77	0.76	0.74	0.75
	76	0.85	0.96	0.84	0.89
	81	0.73	0.72	0.68	0.70

Table 6. The best resulting model of each combination of likelihood or binary coding and the predicted topic. "Likelihood" indicates whether the binary or the likelihood coded data set was used. "Topic" defines the topic, that was predicted. "F1 Val." (F1 Validation) was computed during training and is the relevant value for the selection of the best models. "Precision", "Recall" and "F1" were computed with the unseen test set after the training was finished.

At last, combinations of classifier and predicted topics are examined. To achieve this, the best model based on the f1-validation-score of each combination of classifier and predicted topic was determined. Again, topic 76 indicates the best results independent of the classifier type. All classifiers perform likewise good, with the SVCs showing slightly higher and the PACs slightly lower scores than the other classifiers. The results are displayed in table 7.

Classifier	Topic	F1 Val.	Precision	Recall	F1
GBC	55	0.74	0.78	0.70	0.74
	76	0.82	0.96	0.76	0.85
	81	0.70	0.76	0.60	0.67
PAC	55	0.71	0.73	0.69	0.71
	76	0.82	0.90	0.81	0.85
	81	0.67	0.70	0.66	0.68
SGD	55	0.76	0.79	0.70	0.74
	76	0.84	0.96	0.84	0.89
	81	0.72	0.80	0.65	0.71
SVC	55	0.77	0.76	0.74	0.75
	76	0.85	0.96	0.84	0.89
	81	0.73	0.72	0.68	0.70

Table 7. The best f1-validation results for each combination of classifier and predicted topic. "Classifier" indicates the used classifier. "Topic" defines the topic, that was predicted. "F1 Val." (F1 Validation) was computed during training. "Precision", "Recall" and "F1" were computed with the unseen test set after the training was finished.

The five-topic data set is computed solely with Google labels. The same grid search is performed for both likelihood and binary coded data sets, resulting

in an output of 20 best performing models, one for each grid search. The best performing models per topic, as can be seen in table 8, indicate that again topic 76 shows the best results with a binary f1-score of 0.72 after validation and 0.72 on the test set. The best model for all topics is the SVC. Two of the five best models were trained with the binary data set, whereas the other three were trained with likelihood values. The performance values are generally lower than in the three-topic case.

Model	Topic	Likelihood	F1 Val.	Precision	Recall	F1
SVC	51	True	0.62	0.65	0.56	0.60
SVC	55	True	0.69	0.78	0.66	0.71
SVC	76	False	0.72	0.89	0.61	0.72
SVC	81	True	0.60	0.72	0.57	0.64
SVC	100	False	0.63	0.78	0.48	0.59

Table 8. The best resulting model per predicted topic, computed on the five-topic data set. "Model" is the type of model that shows the best results. "Likelihood" indicates whether the binary or the likelihood data set was used. "Topic" defines the topic, that was predicted. "F1 Val." (F1 Validation) was computed during training and is the relevant value for the selection of the best models. "Precision", "Recall" and "F1" were computed with the unseen test set after the training was finished.

All the code for retrieving, processing and analyzing the data is publicly available via GitLab of the University of Leipzig[7]. With respect to the reproducibility of this work, all the used settings and resulting files are also openly available within the mentioned repository.

5.8 Discussion

Following the evaluation, the results will first be interpreted and discussed. In particular, the research questions and hypothesis outlined in section 5.2 will be addressed. Subsequently, validity and reliability control mechanisms will be presented, which attempt to solve the quality implications discussed in section 4 in relation to the work of Araujo et al. [2].

Interpretation of the Results: At first, the research questions will be discussed, before a closer look will be taken on the hypothesis. With regard to RQ 1, this study indicates an average performance based on the binary f1-score of 0.78 for the task of assigning argumentative images to a correct topic. However, the best model reached a binary f1-score of 0.89 on the test set. As it can be seen in the evaluation section of this paper, the models' performance is significantly above chance and resembles the overall performance values of Araujo et al. [2], while being measured with the more appropriate f1-score instead of the weighted f1-score. Although closely followed by the other classifiers, the SVC seems to be

[7] https://git.informatik.uni-leipzig.de/fp83rusi/cer-automated-visual-content-analysis

the best fitting type of classification model for our purposes. Furthermore, it is notable that the data set that consisted of five topics resulted in overall lower scores than the three-topic data set. This might be the case because of class imbalances which occur because of the binary classification.

RQ 2 stated whether performance differences can be detected when using different commercial computer vision APIs. As table 5 indicates, the labels provided by the Google API were slightly better suited for the task of image classification for argumentation than the Clarifai labels. It has to be noted, that the differences are rather small and could be dependent on a specific task.

With respect to RQ 3, we observed that the combined data set of Clarifai and Google labels performs slightly better than the isolated API outputs. This might be the case because Clarifai and Google labels were merged into one data set, resulting in up to 20 labels. Compared to the up to 10 labels, that were included in the standalone data sets, this might have led to a more informative data set, resulting in the observed increase of classification performance.

Regarding the hypothesis of this work, it can be shown that the findings of Araujo et al. [2] were in fact reproducible and generalizable for our application. We found, that the performance of AVCA for the presented case study was equally good, despite us using a better suited metric for performance measurement, that reduced its positive bias. Additionally, our results suggest that AVCA is also a suitable choice in domains with higher image ambiguity.

Validity and Reliability Controlling: For the case study "Image Classification for Argumentation", attempts were made to develop validity and reliability control mechanisms that address the quality implications of the AVCA framework. In the following, these control mechanisms and potential solutions are explained. In general, we adopt the control mechanisms of Araujo et al. [2] and extend them in the five areas shown below.

QI 1 - Calculation of Method Performance: We find it necessary to determine a suitable quality criterion for the evaluation of the classification results in advance. It is important not to choose as many different evaluation criteria as possible, of which only the highest scores are shown in the evaluation, since this can lead to a possible overestimation of the models. Furthermore, this makes it difficult to compare different models. In our evaluation pipeline, we deliberately use only the binary f1-score for decision-making purposes, which is designed for binary classifications. The best grid search model is thus calculated by the highest binary f1-score for the validation data. These best-performing approaches are then additionally evaluated with the unseen test data. Through this simple and straightforward evaluation pipeline, we aim to ensure that models can be fairly evaluated and the results of different models can be compared with each other.

QI 2 - Internal Validity of AVCA Framework: To remedy the internal validity quality implications, we rely on a much larger data set for our own experiments. Thus, for each class, we use 900 images with a given ground truth topic. For the comparison of the classification models for different topics, we deliberately used an equal distribution of images per topic. Thus, we use 2700 images for the 3-class experiments and 4500 images for the 5-class experiments. In comparison,

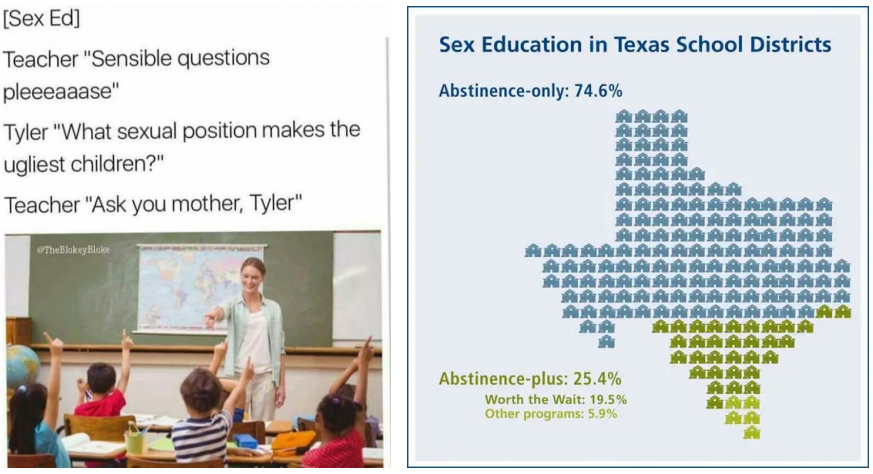
for the experiments of Araujo et al. [2] a total of 868 annotated images were used. This should ensure a greater significance of the results.

(a) (b)

Fig. 4. Sample images for topic 51 "Do we need sex education in schools?". (a) Photo of a demonstration for the protection of LGBTIQ rights. (b) Drawing for illustrating different contraceptives.

QI 3 - External Validity of AVCA Framework: In this work, a large focus was placed on testing the external validity of AVCA. In particular, the generalizability of the framework to other application domains and data sets was to be tested. For our case study, we deliberately chose the Touché'23 data set. The classification of topic-relevant and argumentative images is not a trivial undertaking, as argumentative images show a great deal of ambiguity. Figure 4 and figure 5 illustrate this using images selected from the Touché'23 data set for topic 51 "Do we need sex education in schools?". Images can be argumentative themselves, contain arguments, or be used to support one's own argumentation. This can be done, for example, by showing a photo of a demonstration or by posting a meme. Through this variety of different images, we want to show to what extent AVCA can be generalized and where potential limits of the approach lie.

QI 4 - Validity of Computer Vision Labels: Our process for determining the validity of the computer vision labels starts with an extensive exploratory data analysis. We analyzed various characteristics of the labels of the different data sets in terms of, for example, average, minimum and maximum number of labels, completeness of labels or the amount of unique labels. After the data analysis, matches between the labels of the data sets can be calculated for the data sets that meet the defined minimum requirements. These matches can be used to make a statement about the syntactic similarity of the labels. If the number of matches is low, it can also be useful to create a combined data set from the labels of different APIs and use it for classification. With significantly better classification results with combined labels, it may be concluded that the semantic similarity of the labels is small and that these in combined form thus serve the classification performance.



(a) (b)

Fig. 5. Sample images for topic 51 "Do we need sex education in schools?". (a) Meme on sex education in schools. (b) Infographic on sex education in Texas school districts.

QI 5 - Reproducibility of Experiments: The analysis of the work of Araujo et al. [2] showed that the experiments are not reproducible for outsiders and thus the reliability suffers strongly. It was therefore very important for us to choose with Touché'23 a publicly accessible data set that can be used for research purposes. Furthermore, all our implementations, (intermediate) results, annotations and Clarifai labels are accessible in a public Git repository. We suggest this approach to other researchers using the AVCA framework. We are aware that computer vision approaches are black boxes and thus difficult to analyze and reproduce. Nevertheless, Clarifai shows version numbers for the general-image-recognition model.[8] Since we did not crawl the Google Cloud Vision labels for the Touché'23 data set, we cannot make any statements about versioning.

6 Further Applications

In this section, further applications of the framework are briefly discussed. As it lies in a frameworks' nature, the possibilities of application are rather versatile. Both Araujo et al. [2] and this work implement a case study in a communication science context. A promising approach would be an application in other scientific disciplines. A vast variety of scientific questions and disciplines are imaginable, that could benefit from utilizing the AVCA framework.

To further enhance the existing approach, more APIs or combinations of APIs could be covered. Moreover, modern APIs are often capable of performing more

[8] Clarifai version number used: aa7f35c01e0642fda5cf400f543e7c40

specialized tasks than a label annotation, for instance the detection of business logos, faces or the detection of image text using optical character recognition. This could also lead to new opportunities of larger data sets with more detailed information about images, which in turn might benefit classification results.

A concrete application scenario would be the inclusion of an image classification component in today's state-of-the-art argumentative image search engines. For example, Carnot et al. [6] use BM25 to index images on the basis of the texts on the images' websites. Especially images that are not embedded on web pages or whose web pages do not fit thematically, get only a low score by this procedure, whereby they end up far behind in the search engine's result list. In further work, an adaptation of the AVCA framework for image classification into the indexing process could be investigated. In this case, an individual classification model can already be trained for each predefined topic during the indexing of the images. Subsequently, each image is classified by each classifier. The results can be stored as a bitlist in the index, where 0, for example, stands for not relevant for a topic and 1 for relevant. Each image can thus be assigned to several topics. If a search query is sent, the classification results can be included in the ranking of the images, for example, as a boosting factor for the final score.

7 Conclusion

In our analysis of the AVCA framework, we have identified five quality implications in terms of method performance, validity and reliability. Through the case study "Image Classification for Argumentation", we were able to analyze validity and reliability control mechanisms that address these quality implications and provide possible solutions. In our view, these solutions serve as useful extensions of the AVCA framework, which can be considered in future implementations.

It was particularly important for us to test the reproducibility and generalizability of the framework using our own case study with a high diversity of images. For this purpose, four different classification models were tested for the Touché'23 data set, first on a corpus with three and then with five different topics. In addition, the influence of the annotation label data sets of different computer vision models was analyzed. Our evaluation showed that the best classification results were achieved with the SVC and SGD classification models. The best-performing models averaged over the three selected topics achieved a f1-score of 0.78. The choice of computer vision APIs also plays a role in the classification quality. On average, the best results were achieved with a combined data set with labels from Google Cloud Vision and Clarifai. The choice of a binary classifier or a classifier with likelihood values does not seem to have a significant influence on the model performance. Even though we could not reproduce the concrete experiments and implementations of Araujo et al. [2], we would still conclude from the results of our case study that the AVCA framework is reliable and can be generalized to other application domains.

References

1. Apple Inc.: Education (2023), https://www.apple.com/v/education/home/w/images/overview/spotlight_sisler__evj8rginc38m_large_2x.jpg
2. Araujo, T., Lock, I., van de Velde, B.: Automated visual content analysis (avca) in communication research: A protocol for large scale image classification with pre-trained computer vision models. Communication Methods and Measures 14(4), 239–265 (2020), https://doi.org/10.1080/19312458.2020.1810648
3. Bondarenko, A., Fröbe, M., Kiesel, J., Schlatt, F., Barriere, V., Ravenet, B., Hemamou, L., Luck, S., Reimer, J., Stein, B., Potthast, M., Hagen, M.: Overview of Touché 2023: Argument and Causal Retrieval. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association (CLEF 2023). Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Sep 2023)
4. Bondarenko, A., Fröbe, M., Kiesel, J., Syed, S., Gurcke, T., Beloucif, M., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of touché 2022: Argument retrieval. In: Barrón-Cedeño, A., Martino, G.D.S., Esposti, M.D., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Lect. Notes Comput. Sci., vol. 13390. Springer (2022)
5. Braker, J., Heinemann, L., Schreieder, T.: Aramis at touché 2022: Argument detection in pictures using machine learning. Working Notes Papers of the CLEF (2022)
6. Carnot, M.L., Heinemann, L., Braker, J., Schreieder, T., Kiesel, J., Fröbe, M., Potthast, M., Stein, B.: On stance detection in image retrieval for argumentation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2562–2571. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023), https://doi.org/10.1145/3539618.3591917
7. Champagne, M., Pietarinen, A.V.: Why images cannot be arguments, but moving ones might. Argumentation 34(2), 207–236 (Jun 2019)
8. Dhillon, A., Verma, G.K.: Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence 9(2), 85–112 (2020)
9. Duriau, V.J., Reger, R.K., Pfarrer, M.D.: A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. Organizational Research Methods 10(1), 5–34 (2007), https://doi.org/10.1177/1094428106289252
10. Elkington, J.: Partnerships fromcannibals with forks: The triple bottom line of 21st-century business. Environmental Quality Management 8(1), 37–51 (1998)

11. Himelboim, I., Maslowska, E., Araujo, T.: Integrating network clustering analysis and computational methods to understand communication with and about brands: Opportunities and challenges. Journal of Advertising pp. 1–11 (2023)
12. Iadanza, E., Benincasa, G., Ventisette, I., Gherardelli, M.: Automatic classification of hospital settings through artificial intelligence. Electronics 11(11), 1697 (2022)
13. Kjeldsen, J.E.: The rhetoric of thick representation: How pictures render the importance and strength of an argument salient. Argumentation 29 (2014)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25 (2012)
15. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Babytalk: Understanding and generating simple image descriptions. IEEE transactions on pattern analysis and machine intelligence 35(12), 2891–2903 (2013)
16. Männistö, A., Seker, M., Iosifidis, A., Raitoharju, J.: Automatic image content extraction: Operationalizing machine learning in humanistic photographic studies of large visual archives. arXiv preprint arXiv:2204.02149 (2022)
17. Savage, N.: Seeing more clearly. Commun. ACM 59(1), 20–22 (dec 2015), https://doi.org/10.1145/2843532
18. Szegedy, C., Toshev, A., Erhan, D.: Deep neural networks for object detection. Advances in neural information processing systems 26 (2013)
19. Trilling, D., Jonkman, J.G.F.: Scaling up content analysis. Communication Methods and Measures 12(2-3), 158–174 (2018)
20. United Nations: Take action for the sustainable development goals - united nations sustainable development (2023), https://www.un.org/sustainabledevelopment/sustainable-development-goals/
21. Volkswagen Germany: Weltmarkt (2023), https://assets.volkswagen.com/is/image/volkswagenag/weltmarkt_DB2022FA00082_16-9-f-cc-2?Zml0PWNyb3AsMSZmbXQ9d2VicCZxbHQ9Nzkmd2lkPTE5MjAmaGVp_PTEwODAmYWxpZ249MC4wMCwwLjAwJmJmYz1vZmYmM2E1Nw==