Overview of the ImageCLEF 2022 Aware Task

Adrian Popescu[1], Jérôme Deshayes-Chossart[1], Hugo Schindler[1] and Bogdan Ionescu[2]

[1]*Université Paris-Saclay, CEA, LIST, F-91120, Palaiseau, France*

[2]*AI Multimedia Lab, Politehnica University of Bucharest, Bucharest, Romania*

Abstract

The paper presents the overview of the ImageCLEF 2022 Aware task whose final objective is to make users more aware about the consequences of posting information on social networks. This is important insofar as users are often unaware about the effects of personal data sharing. Focus is put on modeling the impact of sharing impactful real-life situations such as searching for a bank loan, an accommodation, or a job. Since photos are one of the main types of data shared online, the task is instantiated as a photographic user profile assessment. Participants receive a training and validation dataset which includes a set of photographic profiles which are manually rated for each situation. They are required to train algorithms which rate and then rank test profiles in each tested situation. The correlation between automatic and manual profile rankings is used to measure the performance of algorithms. The overview discusses the task settings, the dataset constitution process, and the approaches proposed this year.

Keywords

social networks, photo sharing, object detection, use profile rating, situation modeling, ImageCLEF

1. Introduction

Online social networks (OSNs) thrived on the promise to offer their users rich interactions. The personal data shared online is processed and structured into profiles. These profiles are used to personalize the content delivered to each user, and to fuel the OSNs' business activity, which basically consists of selling access to relevant user segments to interested third parties [1]. The success of this business model depends on the richness of user profiles which are available. Consequently, users are incentivized to share large amounts of data [2].

One problem with this functioning model is the lack of control over the effects of data sharing. Users share data primarily to interact with their contacts, but these data are then potentially usable in other contexts which are unknown to them and where the interpretation of data might change compare to the original context [3]. In an early work, the authors of [4] showed that geolocation sharing can become a threat for users if available to malevolent third parties. A framework which claims to automatically detect potential insurance fraud was introduced in [5]. The sharing of the place of origin on OSNs was shown to lead to discrimination in the labor market [6]. These works indicate that sharing data which are seemingly innocuous can be detrimental for users. It is thus important to make them more aware of the impact of their sharing practices.
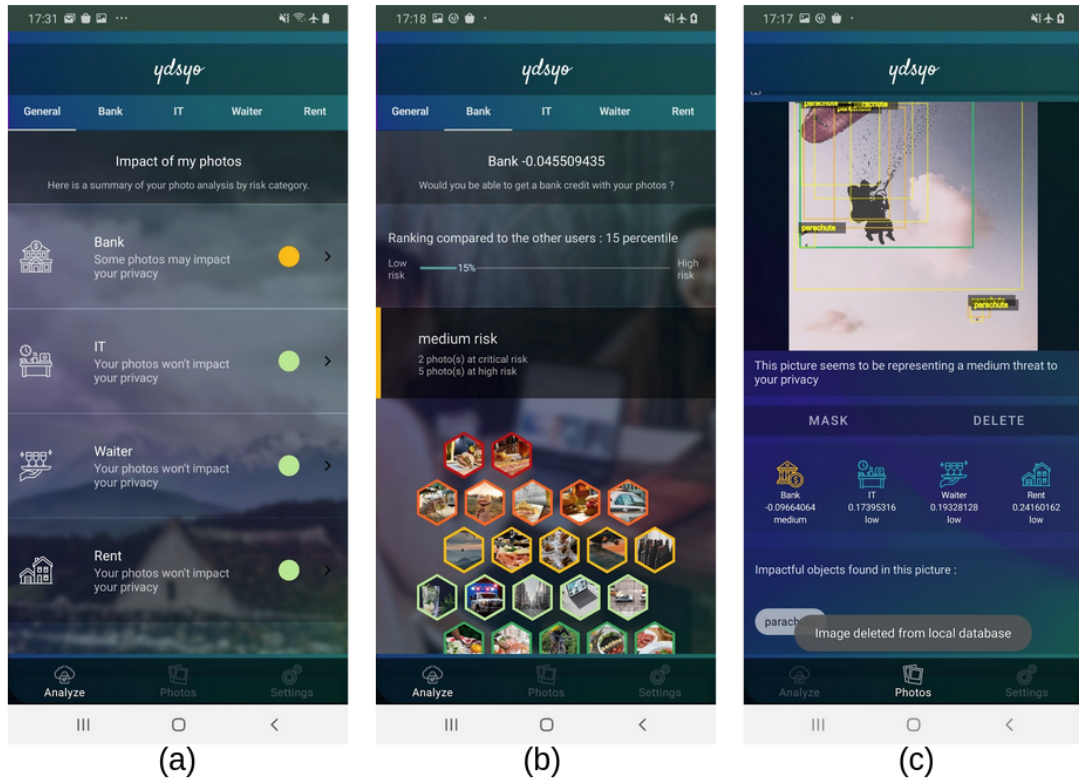
Figure 1: Illustration of the type of feedback which is enabled by the algorithms tested as part of the ImageCLEFaware task. Assuming that a range of reference user profiles were already rated in different situations, subfigure (a) presents the level of exposure of a photographic profile in different situations depending on its relative rating. This relative rating is then detailed in subfigure (b), where the user's profile is situated against the community of reference. A visual explanation of the rating is provided by showing the effect of individual photos in the current situation. Photo-level explanations about how ratings are obtained are proposed in subfigure (c). It depicts the objects which were automatically detected in the photo, gives details about its influence in each modeled situation. Importantly, a control mechanism is introduced since the user is able to remove the photo if it is considered problematic.

These effects are amplified by the use of artificial intelligence tools which extract actionable cues from raw data. For instance, photos, which represent one of the main types of data shared on OSNs, can be analyzed in order to predict their privacy status. An early example of such work was presented [7], where the authors used hand-crafted visual features. Important progress in this task was obtained by the introductions of deep learning representations [8, 9]. While inferring privacy status of a photo is interesting, similar techniques can be used to automatically derive finer grained cues from images.

We built on the works discussed above to propose the Aware shared task. Its technical objective is to automatically score user photographic profiles in a series of impactful situations. Such profiles can then be used to give users feedback about the potential consequences of personal data sharing on their real lives. The long-term objective is to enable and improve

awareness rising applications such as the YDSYO prototype [1], whose functioning is illustrated in Figure 1. The reminder of this paper is organized as follows: Section 2 presents the task, Section 3 the associated dataset, Section 4 discusses the evaluation methodology, Section 5 analyzes this year's results, and Section 6 discusses the current conclusions and future directions.

2. Task

The task aims to raise users' awareness about the effects of personal data sharing. Focus is put on photos since they potentially convey a lot of information which is usable by third parties if an appropriate analysis pipeline is deployed. In contrast with existing works which compute on the impact of single images [8, 9, 10, 7], we hypothesize that sharing effects should be assessed primarily at a user profile level. This choice is made because the data sharing has a cumulative effect, and the final rating of a profile results from an assessment of the entire user profile. Since the same data can be interpreted differently depending on the context [11], we model four real-life situations. The user is assumed to search for: (1) a bank loan, (2) a new accommodation, (3) a job in IT, and (4) a job as a waiter. The task implementation is based on three main components:

- **situation models** – each situation is modeled as an array of visual objects which can be detected in images. Each of these objects has a situation-related rating which was obtained by crowdsourcing.
- **visual objects** – can be automatically detected in images using object detectors, such as Faster-RCNN [12]. These detections are essential for the task since they are aggregated into user profiles.
- **photographic user profiles** – a set of images which were shared by a user. They provide a raw representation of the user. The actual user representation includes the objects which are automatically detected in the images.

These three components are combined into a dataset which is provided to task participants. The constitution of this dataset is described in more details in the next section.

3. Dataset

3.1. Situation Models

Given the objective of the task, actionable models of the four situations are needed. Focus is on visual objects and we propose a crowdsourcing approach to obtain these models. We start from an initial list of objects which are represented in three object detection datasets: OpenImages [13], ImageNet LSVRC [14], and COCO [15]. The combination of these datasets is important to obtain good coverage of object detectors. The objective is to obtain ratings which encode the influence of these objects in each modeled situation. This is done via a dedicated crowdsourcing interface which is illustrated in Figure 2. The name and a few illustrative images are presented for each object, along with the possible ratings for the the situation in which
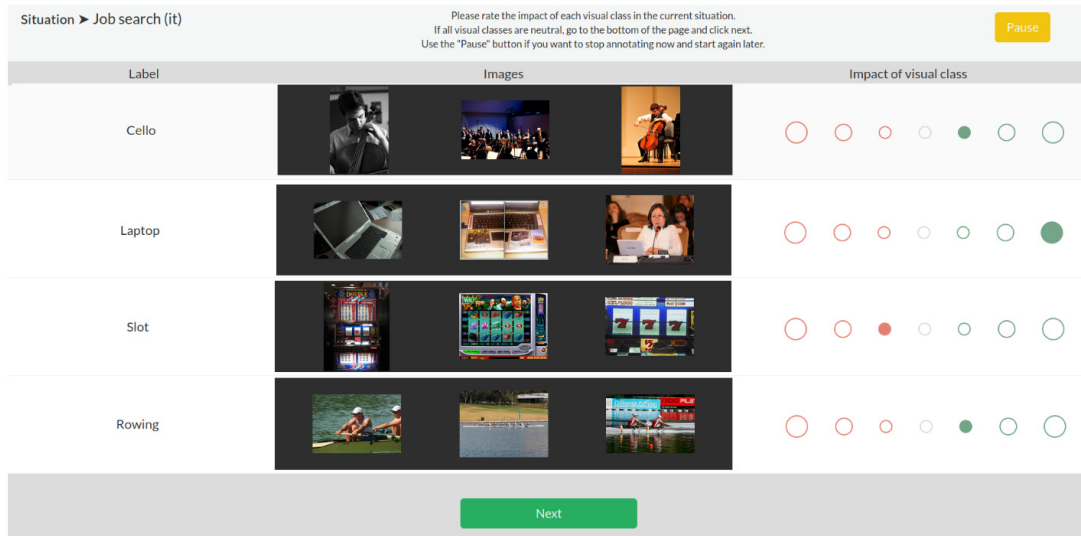
---

[1]https://ydsyo.app

Figure 2: Interface for rating the influence of visual objects in each modeled situation.

they are currently evaluated. Its rating is annotated using a 7-points Likert scale with values between -3 (strongly negative influence) to +3 (strongly positive influence). The final rating is obtained by averaging ratings per situation from 52 annotators (14 per situation) who took part in the experiment. The situation model includes the 269 visual objects for which the final rating is not null for at least one of the four situations. Inter-rater agreement, which is important for task which are prone to bias such as the one proposed here, is computed using the average deviation index (AD) [16]. The obtained AD varies between 0.48 for IT and 0.65 for $WAIT$. These values are well below $AD \leq 1.2$, the maximum acceptable value for a 7-points Likert scale defined in [17].

3.2. Visual Objects Detection

As we mentioned, we merge three existing datasets: OpenImages [13], ImageNet [14] and COCO [15]. Whenever an object is present in more than one dataset, we select images from each dataset in a balanced manner to reduce biases. The number of samples per object is variable across datasets, and we keep a maximum of 1,000 images per object to reduce imbalance. The resulting dataset includes 269 objects and 137,976 images. The average and standard deviation of the distribution are 513 and 305, respectively. We train object detectors by combining an Inception-ResNet-v2 [18] with atrous convolutions backbone with the well-known Faster RCNN module [12] for the detection step. More details about the detector are provided in the supplementary material of [11].

3.3. Photographic User Profiles

The third core component of the dataset is the set of photographic user profiles which should be automatically rated in each situation. Photos were sampled from the YFCC dataset [19], a dataset
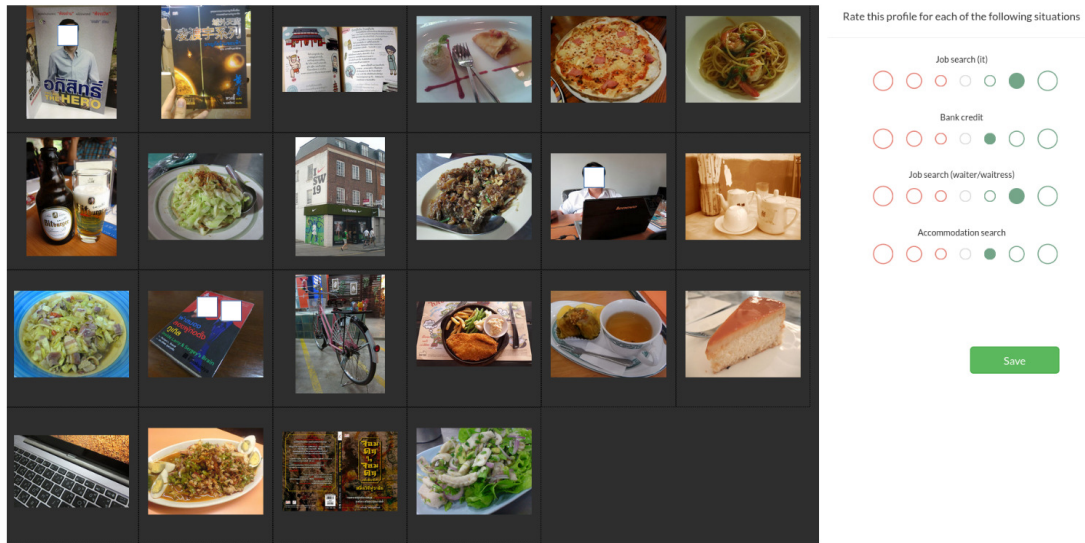
Figure 3: Interface for rating the influence of visual objects in each modeled situation.

which includes only data shared under Creative Commons licenses. The labeling approach used here is similar to the one deployed for situation modeling. Visual profiles are evaluated using a 7-points Likert scale which goes from -3 (strongly unappealing) to +3 (strongly appealing). Each profile includes a total of 100 images which were sampled from the user's contribution to YFCC. The 2021 version of the dataset included rating collected from 9 annotators for a set of 500 users. In 2022, the dataset was enriched and it now includes 1,000 profiles. The newly includes profiles were labeled by up to 10 annotators.

The annotation interface is illustrated in Figure 3. The images of each visual profile are shown on a single page, along with the possible ratings in all four situations. Participants are asked to look at all the photos and provide a global rating of the user profile in each situation. The order in which profiles were presented to annotators was randomized to avoid any ordering bias. Similar to situation modeling, the inter-rater agreement was measured in each situation using the AD index [16]. The obtained values are below 0.9, which is within the acceptability bounds for a 7-points Likert scale ($AD \leq 1.2$) [17]. The disagreement is higher when rating profiles compared to object rating. This is intuitive insofar profile rating is a more complex task which involves simultaneous evaluation of a set of 100 images for each profile.

3.4. Dataset Distribution

Automatic object detection (Section 3.2) is applied to photographic user profiles (Section 3.3) to create usable representations of the profiles. Profile representation are an aggregation of image-related vectors, with each vector being composed of all objects detected in the image, along with their position and detection confidence scores.

The data included in the dataset is personal, and even sensitive according to Article 9 of

GDPR [2]. It is important to distribute the dataset so as to minimize any deanonymization risks. Images themselves are not provided, but only their representative vectors. In addition, the user, image and visual object names are anonymized in the distributed version of the dataset. All data were shared using a JSON format to facilitate their parsing by participants.

The profiles subset was split into three parts which include 600, 200 and 200 profiles intended for training, validating and testing algorithms, respectively. Training and validation data were provided along with their associated manual ratings. The dataset equally includes situation models (i.e., visual object ratings), which provide complementary information to that available in the profiles.

4. Evaluation Methodology

The performance of the task submissions is measured by their ability to produce profile ratings which are similar to those provided by human annotators. The similarity between automatic and manual rankings is measured using the Pearson correlation coefficient, which is a normalized measure of the covariance of two variables. Its values range from -1 (inverse correlation) to 1 (perfect correlation), with 0 standing for no linear dependency between variables. Correlation is measured for each of the four situations and the submissions are evaluated using the average of individual values. The obtained scores can be analyzed using Cohen's interpretation of the Pearson correlation coefficient [20]. Correlation is considered weak for values between 0.1 and 0.3, moderate between 0.3 and 0.5 and strong above 0.5.

5. Results

Three teams submitted a total of 9 runs, with 5 submitted by SSNCSE_KS_NA_AKR_CB [21], 2 from JBTTM [22] and 2 from ssnce-cse-JT (no working notes provided). An overview of results is presented in Table 1.

JBTTM [22] tested approaches based on random forest regressors and dense neural networks (NNs). Preprocessing was the same for both approaches. It consisted in creating a stacked matrix per user which included the location and confidence scores per situation were concatenated. They compared a random forest regressor and an extra tree regressor to combine predictions from individual prediction trees. A 7-layers deep NN was tested and the authors noted that performance for dense NNs is suboptimal. They explain this finding by an insufficient amount of data which is available for training the neural net. The reported results are similar for regressor and NN approaches since PCC reaches 0.139.

SSNCSE_KS_NA_AKR_CB [21] tested random forest regressors with different preprocessing of the data, and also explored the effect of fine-grained parameter tuning. Their baseline run (179994) used user profile descriptions made of a combination of average confidence score and situation impact scores for each detected object. Focus was put on optimizing the number of estimator used by the regressor. The range from 10 to 1000 was explored and the best results were obtained with 650 estimator. The correlation score reported for this run is 0.288. The

---

Team	Run ID	Method	PCC
SSNCSE_KS_NA_AKR_CB	179994	random forest regressor	0.288
SSNCSE_KS_NA_AKR_CB	182709	179994 + object detection matrix	0.544
SSNCSE_KS_NA_AKR_CB	182888	182709 + parameter tuning	0.542
SSNCSE_KS_NA_AKR_CB	182890	-	0.540
SSNCSE_KS_NA_AKR_CB	182892	182888 + object bounding boxes	0.519
JBTTM	181730	random forest regressors	0.139
JBTTM	181665	dense neural network	0.139
ssnce-cse-JT	182457	-	0.0
ssnce-cse-JT	182300	-	0.0

Table 1
Aware task result measured using the Pearson Correlation Coefficient (PCC). Higher values indicate better performance. A short description of each the method is proposed whenever available. Details for SSNCSE_KS_NA_AKR_CB and JBTTM teams are available in their working notes papers ([21] and [22], respectively).

second run (182709) improved over the first by adding an object-confidence score matrix to the input. This addition is highly effective since the PCC score of this run is 0.544. The third run (182888) focused on a fine-tuning of model parameters. The team varied parameters such as bootstrapping, max number of features, max depth, max samples per leaf, number of estimator. This exploration did not lead to a performance improvement compared to the second run since PCC reached 0.542. The last run described by the team added the area of bounding boxes which delimit the object detection as a proxy for object importance. The use of the area did not prove useful since PCC dropped to 0.519.

The results submitted this year show that both the machine learning methods and an appropriate preprocessing of the data are important to obtain good quality profile ratings. This is notably underlined by the experiments run by team SSNCSE_KS_NA_AKR_CB [21], with run performance almost doubling when input data are appropriately prepared. Globally, the obtained scores show that the task is doable, since good a good correlation level is reported following Cohen's interpretation of PCC [20]. However, the correlation is still far from perfect and the task cannot be considered as solved.

6. Conclusion

Participation was comparable between 2021 and 2022, and so where the best scores reported. The low participation might be explained by a combination of factors which include: the novelty of the task, its niche orientation, and the effectiveness of the communication effort made to advertise it. Of these three dimensions, the latter one can be improved by more widespread and early diffusion of the call for participation and, potentially, by adding a financial incentive in order to stimulate participation. From a technical perspective, we will try to: (1) expand the dataset with new user profiles to the dataset to make it more robust, (2) add a more recent object detection, such as EfficientDet [23], which improves the intrinsic quality of detections, or Detic [24], which scales-up the number of detectable objects with image-level supervision.

Acknowledgements

References

[1] K. Curran, S. Graham, C. Temple, Advertising on facebook, International Journal of E-business development 1 (2011) 26–33.

[2] P.-W. Fu, C.-C. Wu, Y.-J. Cho, What makes users share content on facebook? compatibility among psychological incentive, social capital focus, and content type, Computers in Human Behavior 67 (2017) 23–32.

[3] M. Burke, J. Cheng, B. de Gant, Social comparison and facebook: Feedback, positivity, and opportunities for comparison, in: R. Bernhaupt, F. F. Mueller, D. Verweij, J. Andres, J. Mc-Grenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, R. Kocielnik (Eds.), CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020, ACM, 2020, pp. 1–13. URL: https://doi.org/10.1145/3313831.3376482. doi:10.1145/3313831.3376482.

[4] G. Friedland, J. Choi, Semantic computing and privacy: a case study using inferred geo-location, Int. J. Semantic Computing 5 (2011) 79–93. URL: https://doi.org/10.1142/S1793351X11001171. doi:10.1142/S1793351X11001171.

[5] M. Diaz-Granados, J. Diaz-Montes, M. Parashar, Investigating insurance fraud using social media, in: 2015 IEEE International Conference on Big Data (Big Data), IEEE, 2015, pp. 1344–1349.

[6] M. Manant, S. Pajak, N. Soulié, Can social media lead to labor market discrimination? evidence from a field experiment, Journal of Economics & Management Strategy 28 (2019) 225–246.

[7] S. Zerr, S. Siersdorfer, J. S. Hare, E. Demidova, Privacy-aware image classification and search, in: W. R. Hersh, J. Callan, Y. Maarek, M. Sanderson (Eds.), The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, August 12-16, 2012, ACM, Portland, OR, USA, 2012, pp. 35–44. URL: https://doi.org/10.1145/2348283.2348292. doi:10.1145/2348283.2348292.

[8] T. Orekondy, B. Schiele, M. Fritz, Towards a visual privacy advisor: Understanding and predicting privacy risks in images, in: IEEE International Conference on Computer Vision, ICCV 2017, October 22-29, 2017, IEEE Computer Society, Venice, Italy, 2017, pp. 3706–3715. URL: https://doi.org/10.1109/ICCV.2017.398. doi:10.1109/ICCV.2017.398.

[9] E. Spyromitros-Xioufis, S. Papadopoulos, A. Popescu, Y. Kompatsiaris, Personalized privacy-aware image classification, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 71–78. URL: https://doi.org/10.1145/2911996.2912018. doi:10.1145/2911996.2912018.

[10] A. Tonge, C. Caragea, Image privacy prediction using deep neural networks, ACM Transactions on the Web (TWEB) 14 (2020) 1–32.

[11] V.-K. Nguyen, A. Popescu, J. Deshayes-Chossart, Unveiling real-life effects of online photo sharing (2022) 2898–2908.

[12] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2015, pp. 91–99.

[13] A. Kuznetsova, H. Rom, N. Alldrin, J. R. R. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, V. Ferrari, The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale, CoRR abs/1811.00982 (2018). URL: http://arxiv.org/abs/1811.00982. arXiv:1811.00982.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, F. Li, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (2015) 211–252. URL: https://doi.org/10.1007/s11263-015-0816-y. doi:10.1007/s11263-015-0816-y.

[15] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, in: D. J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, volume 8693 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 740–755. URL: https://doi.org/10.1007/978-3-319-10602-1_48. doi:10.1007/978-3-319-10602-1_48.

[16] M. J. Burke, L. M. Finkelstein, M. S. Dusig, On average deviation indices for estimating interrater agreement, Organizational Research Methods 2 (1999) 49–68.

[17] M. J. Burke, W. P. Dunlap, Estimating interrater agreement with the average deviation index: A user's guide, Organizational research methods 5 (2002) 159–172.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: S. P. Singh, S. Markovitch (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, AAAI Press, 2017, pp. 4278–4284. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806.

[19] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L. Li, YFCC100M: the new data in multimedia research, Commun. ACM 59 (2016) 64–73.

[20] J. Cohen, Statistical power analysis for the behavioral sciences, Academic press, New York, 2013.

[21] A. Nunna, A. K. Rathinasapabathi, C. B. P. K, K. Srinivasan, Ssn cse at imageclefaware 2022: Contextual job search feedback score based on photographic profile using a random forest regression technique, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[22] A. S. Kumar, A. A, J. G. M, K. R. A, B. Jayaraman, M. T.T., Multi regressor based user rating predictor for imageclef aware 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[23] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, June 13-19, 2020, Computer Vision Foundation / IEEE, Seattle, WA, USA, 2020,

pp. 10778–10787. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Tan_ EfficientDet_Scalable_and_Efficient_Object_Detection_CVPR_2020_paper.html. doi:`10. 1109/CVPR42600.2020.01079`.

[24] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, I. Misra, Detecting twenty-thousand classes using image-level supervision, arXiv preprint arXiv:2201.02605 (2022).