ImageCLEF 2023 Highlight: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications

* Bogdan Ionescu[1], Henning Müller[2], Ana Maria Drăgulinescu[1], Adrian Popescu[3], Ahmad Idrissi-Yaghir[4], Alba Garcia Seco de Herrera[5], Alexandra Andrei[1], Alexandru Stan[6], Andrea M. Storås[7], Asma Ben Abacha[8], Christoph M. Friedrich[4], George Ioannidis[6], Griffin Adams[9], Henning Schäfer[10], Hugo Manguinhas[11], Ihar Filipovich[12], Ioan Coman[1], Jérôme Deshayes[3], Johanna Schöler[13], Johannes Rückert[4], Liviu-Daniel Ştefan[1], Louise Bloch[4], Meliha Yetisgen[14], Michael A. Riegler[7], Mihai Dogariu[1], Mihai Gabriel Constantin[1], Neal Snider[15], Nikolaos Papachrysos[13], Pål Halvorsen[7], Raphael Brüngel[4], Serge Kozlovski[16], Steven Hicks[7], Thomas de Lange[13], Vajira Thambawita[7], Vassili Kovalev[16], and Wen-Wai Yim[8]

[1] Politehnica University of Bucharest, Romania bogdan.ionescu@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] CEA LIST, France
[4] University of Applied Sciences and Arts Dortmund, Germany
[5] University of Essex, UK
[6] IN2 Digital Innovations, Germany
[7] SimulaMet, Norway
[8] Microsoft, USA
[9] Columbia University, USA
[10] University Hospital Essen, Germany
[11] Europeana Foundation, Netherlands
[12] Belarus State University, Belarus
[13] Sahlgrenska University Hospital, Sweden
[14] University of Washington, USA
[15] Microsoft/Nuance, USA
[16] Belarusian Academy of Sciences, Belarus

Abstract. In this paper, we provide an overview of the upcoming ImageCLEF campaign. ImageCLEF is part of the CLEF Conference and Labs of the Evaluation Forum since 2003. ImageCLEF, the Multimedia Retrieval task in CLEF, is an ongoing evaluation initiative that promotes the evaluation of technologies for annotation, indexing, and retrieval of multimodal data with the aim of providing information access to large collections of data in various usage scenarios and domains. In its 21st edition, ImageCLEF 2023 will have four main tasks: (i) a *Medical* task addressing automatic image captioning, synthetic medical images created with GANs, Visual Question Answering for colonoscopy images, and medical dialogue summarization; (ii) an *Aware* task addressing the

* apart from the general organizers, authors are listed in alphabetical order.

prediction of real-life consequences of online photo sharing; (iii) a *Fusion* task addressing late fusion techniques based on the expertise of a pool of classifiers; and (iv) a *Recommending* task addressing cultural heritage content-recommendation. In 2022, ImageCLEF received the participation of over 25 groups submitting more than 258 runs. These numbers show the impact of the campaign. With the COVID-19 pandemic now over, we expect that the interest in participating, especially at the physical CLEF sessions, will increase significantly in 2023.

Keywords: Information retrieval, medical AI, image captioning, GANs, Visual Question Answering, dialogue summarization, social media, user awareness, late fusion, cultural heritage, content recommending, ImageCLEF benchmarking, annotated data.

1 Introduction

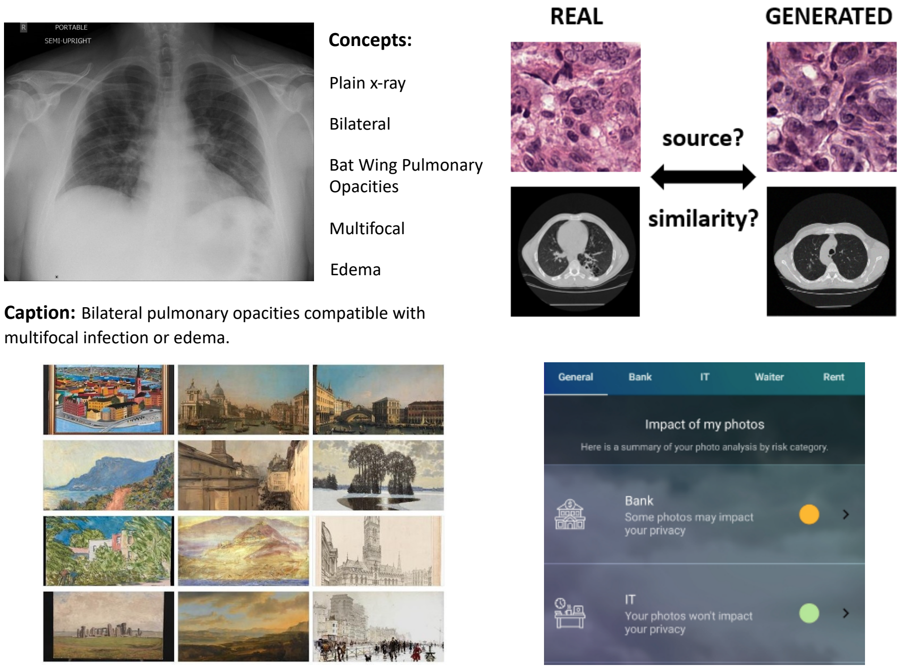
The ImageCLEF evaluation campaign has been organised each year since 2003 and continues to enable the benchmarking activities and research tasks on the cross-language annotation, indexing and retrieval of multimodal data. As part of the Conference and Labs of the Evaluation Forum (CLEF) [17,18], the 21st edition of ImageCLEF will be hosted by CERTH in Thessaloniki, Greece, in September 2023[17]. A set of benchmarking tasks was designed to test different aspects of mono- and cross-language information retrieval systems [14, 17, 18]. Target communities involve (but are not limited to): information retrieval (e.g., text, vision, audio, multimedia, social media, sensor data), machine learning, deep learning, data mining, natural language processing, image and video processing; with special emphasis on the challenges of multi-modality, multi-linguality, and interactive search. Both, ImageCLEF lab [29] and CLEF campaign, have important scholarly impact with 407 publications on Web of Science (WoS) mentioning ImageCLEF (with 2801 WoS citations) and 6730 results on Google Scholar.

The following sections introduce the four tasks that are planned for 2023, namely: ImageCLEFmedical, ImageCLEFaware, ImageCLEFfusion, and the new ImageCLEFrecommeding. Fig. 1 captures representative images for the aforementioned proposed tasks.

2 ImageCLEFmedical

The ImageCLEFmedical task has been carried out every year since 2004 [18]. The 2023 edition will include the following tasks: (i) a sequel of the caption task with medical concept detection and caption prediction, (ii) a new task on synthetic medical images generated with GANs, (iii) a new Visual Question Answering and generation task for colonoscopy images, (iv) a new pilot task on medical doctor-patient conversation summarization for generation of clinical notes.

[17] https://clef2023.clef-initiative.eu/



Impact of my photos

Here is a summary of your photo analysis by risk category.

Bank
Some photos may impact your privacy

IT
Your photos won't impact your privacy

Fig. 1. Sample images from (left to right, top to bottom): ImageCLEFmedical-caption with an image with the corresponding CUIs and caption, ImageCLEFmedical-GAN with an example of real and generated images, ImageCLEFrecommending task with an example of editorial "European landscapes and landmarks" Gallery, and ImageCLEFaware with an example of user photos and predicted influence when searching for a bank loan.

ImageCLEFmedical-caption[18]. The topic of the *caption* task resides in the interpretation of the insights gained from radiology images. In the 7th edition of the task [8, 10, 21–23, 25], there will be two subtasks: concept detection and caption prediction. The *concept detection* subtask aims to develop competent systems that are able to predict the Unified Medical Language System (UMLS®) Concept Unique Identifiers (CUIs) based on the visual image content. The F1-Score [11] will be used to evaluate the participating systems in this subtask. The *caption prediction* subtask focuses on implementing models to predict captions for given radiology images. After using the BLEU [20] score in previous editions, it was decided to change the primary scoring metric for the 2023 challenge because recent studies [3, 15, 31] that investigated the relationship between the BLEU score and human judgment found that there was only a moderate correlation. In reviewing BLEU as an appropriate metric, it was also found that semantically correct sentences tend to be disadvantaged when they differ from

[18] https://www.imageclef.org/2023/medical/caption

the reference in terms of morphology [32]. This is also consistent with recent feedback from the previous edition. To this end, several different metrics alike BERTScore [32] and BLEURT [26] are currently being evaluated, which aim to capture the underlying semantics by leveraging state-of-the-art transformer-based language models like BERT [6]. In 2023, an updated version of the Radiology Objects in Context (ROCO) [24] dataset will be used, further extended in comparison to 2022's edition. As in the previous editions, the updated dataset will be manually curated (e.g., image modalities, anatomy in x-ray images) after using multiple concept extraction methods to retrieve accurate CUIs.

ImageCLEFmedical-GAN[19]. The *GANs* task is a completely new challenge in the ImageCLEFmedical track. The task is focused on examining the existing hypothesis that GANs are generating medical images that contain certain "fingerprints" of the real images used for generative network training. If the hypothesis is correct, artificial biomedical images may be subject to the same sharing and usage limitations as real sensitive medical data. On the other hand, if the hypothesis is wrong, GANs may be potentially used to create rich datasets of biomedical images that are free of ethical and privacy regulations. The participants will test the hypothesis by solving two tasks. The first task is dedicated to the detection of mentioned "fingerprints" in the artificial biomedical image data. Given two sets of real images and a set of images generated by some GAN model, participants will try to detect, which set of real images was used for the generative model training. The second task is focused on the analysis of the similarity of output produced by generative models with different architectures and/or with different training strategies. In this task participants will be given a set of artificial images produced by a set of different generative models and participant will try to group images by their source model. Possible options for a data type in both tasks are histology, X-ray, CT scans.

ImageCLEFmedical-VQA[20]. The *VQA* task, also a new task in this format, combines the task of visual question answering and question generation with detecting diseases within the gastrointestinal (GI) tract. Medical doctors usually examine the GI tract using colonoscopy, gastroscopy or capsule endoscopy. For the VQA task, we combine images taken from the procedures with medically relevant questions and answers. In total, the task has three subtasks: (i) The visual question answering (VQA) subtask asks participants to generate text answers given a text question and image pair. For example, we provide an image containing a colon polyp with the following question: "Where in the image is the polyp located?". Here, the answer should be a textual description of where in the image the polyp is located, like the upper-left or in the center of the image. Example questions could be "How many findings are in the image?", "What are the colors of the findings?", etc. (ii) The visual question generation (VQG) subtask requires participants to generate text questions based on a given text answer and image pair. This task can be seen as the inverse of VQA, where instead of generating the answer, we are asking for the question. An example

[19] https://www.imageclef.org/2023/medical/gans

[20] https://www.imageclef.org/2023/medical/vqa

could be that given the answer "The image contains a polyp" and an image containing a polyp, the question should be "Does the image contain an abnormality?". (iii) The visual location question answering (VLQA) subtask where the participants get an image and a question and are required to answer it by providing a segmentation mask for the image. Example questions are: "Where in the image is the polyp?", "Where in the image is the normal and the diseased part?", "What part of the image shows normal mucosa?" The data is based on the HyperKvasir dataset [2] with additional question-and-answer ground truth verified by medical doctors. It includes images spanning the entire GI tract and will include abnormalities, surgical instruments, and normal findings from gastroscopy, colonoscopy and capsule endoscopy procedures. For VQA and VQG, at least 5,000 image samples, each with five question-and-answer pairs will be provided. For VLQA at least 1,000 images with question and segmentation mask pairs are given. Evaluation will be performed using well know metrics suitable for medical applications such as precision, recall, F1 score, and Matthew correlation coefficient [12].

ImageCLEFmedical-mediqa[21]. The *MEDIQA-Sum* task is a new pilot task that focuses on automatic note generation from patient-clinician conversations, a challenging task that encompasses spoken language understanding and clinical note generation. MEDIQA-Sum 2023 will include three subtasks: (i) the Dialogue2Topic Classification subtask focuses on identifying the topic associated with a conversation snippet between a doctor and patient, (ii) the Dialogue2Note Summarization subtask focuses on producing a clinical note section text summarizing a conversation snippet between a doctor and a patient, and (iii) the Full-Encounter Dialogue2Note Summarization subtask tackles the generation of a full clinical note summarizing a full doctor-patient encounter conversation. New datasets have been created for the MEDIQA subtasks. The Dialogue2Note dataset was created based on clinical notes and corresponding conversations written by domain experts. The Full-Encounter Dialogue2Note dataset consists of full doctor-patient encounters and corresponding notes written by medical scribes. We will measure topic prediction with standard classification metrics such as F1 and accuracy. The two Dialogue2Note subtasks will use SOTA language generation metrics including ROUGE [16], BERTScore [32], and BLEURT [26].

3 ImageCLEFaware

When users contribute to online platforms they share data in a given context, which is controlled and understood by them. These data are then stored and can later be reused in contexts which were not anticipated initially. Such reuse can be directed toward impactful situations, and can have serious consequences for the users' real lives. For instance, future employers often search online information about prospective candidates. This process can involve humans or be automated using Artificial Intelligence tools. Users should be aware that such inferences are possible, with potentially detrimental effects for them.

[21] https://www.imageclef.org/2023/medical/mediqa

Photos represent a large part of the data shared online, and the ImageCLEFAware[22] task focuses on their usage. Given a set of user photographic profiles, and four modeled situations (search for a bank loan, an accommodation, a waiter job, or an IT job), the objective consists in providing feedback to the users about how their profiles compare to those of a community of reference. Users' photos are manually labeled with an appeal score by several annotators, and an average score per situation is computed and serves a ground truth. User profiles are created by automatically detecting visual objects in users' images, and the resulting profiles can be used to automatically rate and rank profiles. Correlation between automatic and manual profile rankings will be measured using a classical measure such as the Pearson correlation coefficient.

While the global objective remains unchanged since the first edition, the dataset will be enriched and updated for the third edition of the task. The main changes refer to: (1) a larger number of user profiles to make the dataset more robust, and (2) a new object detector based on EfficientDet, to provide better profiles. Among the resources associated to the proposed tasks that will be provided to the participants in different communities, we include: (i) visual object ratings per situation obtained through crowdsourcing; (ii) automatically extracted visual object detections for over 350 objects which have non-null rating in at least one situation, using a new object detector compared to 2022.

The dataset includes personal data, and strong anonymization is performed in order to comply with EU's General Data Protection Regulation. Participants receive only the object detections which compose the profile. Furthermore, the names of these objects are anonymized.

4 ImageCLEFfusion

Late fusion approaches represent one of the goto methods of improving machine learning performance in particular domains, where single-system performance may not be acceptable, or even in critical systems, where every improvement is vital. There are numerous examples in the literature where the top performers on specific datasets or even entire domains are represented by late fusion systems that use several models and fuse their predictions. Some of these examples would be the prediction of media memorability [1] and interestingness [30], the detection of violent video scenes [5], and human action recognition [28]. As the previous edition of the ImageCLEFfusion task shows [27], numerous types of approaches to fusion exist, ranging from simple statistical approaches to more complex systems that use traditional machine learning methods like KNN or SVR, or deep neural network-based learning, and even using more than one stage of fusion in order to create the final set of predictions.

In this context we propose the second edition of the ImageCLEFfusion[23] task, a follow-up to last year's edition [27]. In the first edition, two tasks are defined as two different machine learning task types, namely: (i) a regression scenario that

[22] https://www.imageclef.org/2023/aware

[23] https://www.imageclef.org/2023/fusion

uses data associated with the prediction of multimedia interestingness, extracted from the Interestingness10k dataset [4], and (ii) a retrieval scenario, using data that targets the retrieval of diverse social images extracted from the Div150 challenge [13]. Annotation data, metrics, inducers, and all other tools are developed and published during the respective benchmarking campaigns, and are provided to participants to the fusion task. For this edition, we propose integrating a third task, that targets another type of machine learning task, namely multi-label classification. Thus, we will integrate data associated with the Concept Detection task from the ImageCLEFmedical caption task [25].

5 ImageCLEFrecommending

ImageCLEFrecommending[24] is a new task which focuses on content-recommendation for cultural heritage content. Despite current advances in content-based recommendation systems, there is limited understanding how well these perform and how relevant they are for the final end-users. This task aims to fill this gap by developing ground truth data of recommendations and by allowing benchmarking different recommendation systems and methods.

The task targets a key infrastructure for researchers and heritage professionals, namely Europeana [9]. With over 53 million records, the single search bar that served as the main access point was identified as a bottleneck by many users. Thus, the strategy has gradually shifted towards exploration of the available collections based on themes. Now users can explore over 60 curated digital exhibitions, countless galleries and blog posts (to be refered to as *editorials*). The metadata of the content items available on Europeana is in most cases very rich and must follow a very well defined structure given by the Europeana Data Model [7]. The current Europeana Recommendation Engine [19] focuses only on providing recommended items based on the content of a gallery, which is a collection of items with a title and optional description. The recommendations are based on similarity of the most important metadata fields of the items (*dc:title, dc:creator, dc:subject, dc:date and dc:description*) into a sentence embedding. However, recommendations for editorials are done at the moment only manually. For instance when a new blog is created, the author would manually provide a list of related galleries, blogs or exhibitions that have been already published.

The task requires participants to devise recommendation methods and systems, apply them in the supplied data set gathered from Europeana and provide a series of recommendations for items and editorials. The task is thus divided into two sub-tasks: (i) given a list of items, provide a list of recommended items; (ii) given an editorial (Europeana blog or gallery), provide a list of recommended editorials. For the task a new dataset based on Europeana items and editorials will be provided to the participants. The individual items in the dataset will include a wealth of metadata based on the Europeana Data Model schema. Performance will be evaluated on the basis of the recommendations that are provided computing Mean Average Precision at X (Map@X) compared to the ground truth.

[24] https://www.imageclef.org/2023/recommending

Moreover, because black-box systems make it difficult for users to assess why the recommendation should be trusted, the systems in this task that can provide an explanation for the results provided will be awarded additional points in terms of evaluation metrics.

6 Conclusions

The current paper provides an overview of the tasks proposed by the 2023 ImageCLEF evaluation campaign organised in the framework of the 14th CLEF Conference and Labs of the Evaluation Forum scheduled for September 2023, in Thessaloniki, Greece. Since 2003, the tasks proposed by the ImageCLEF lab gained popularity being held for the evaluation of technologies for *annotation*, *indexing*, *classification* and *retrieval* of multimodal data, with the objective of providing information access to large collections in various usage scenarios and domains. The 21st edition brings several new tasks, e.g., content recommending, generative adversarial networks, visual question answering and generation, medical dialogue summarization, while some old tasks were discontinued. All the tasks are solving challenging current issues and will provide a set of new test collections simulating real-world situations. Such collections are important to enable researchers to assess the performance of their systems and to compare their results with others following a common evaluation framework.

Acknowledgement

The lab is supported under the H2020 AI4Media “A European Excellence Centre for Media, Society and Democracy” project, contract #951911, as well as the ImageCLEFaware, ImageCLEFfusion and ImageCLEFrecommending tasks. The work of Louise Bloch and Raphael Brüngel was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Ahmad Idrissi-Yaghir and Henning Schäfer was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed).

References

1. Azcona, D., Moreu, E., Hu, F., Ward, T.E., Smeaton, A.F.: Predicting media memorability using ensemble models. In: Working Notes Proceedings of the MediaEval 2019 Workshop. CEUR Workshop Proceedings, vol. 2670. CEUR-WS.org (2019)
2. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1), 1–14 (2020)
3. Cao, Y., Shui, R., Pan, L., Kan, M.Y., Liu, Z., Chua, T.S.: Expertise style transfer: A new task towards better communication between experts and laymen. In: Proceedings of the 58th Annual Meeting of the Association

for Computational Linguistics. pp. 1061–1071. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.100, https://aclanthology.org/2020.acl-main.100

4. Constantin, M.G., Ştefan, L.D., Ionescu, B., Duong, N.Q., Demarty, C.H., Sjöberg, M.: Visual interestingness prediction: A benchmark framework and literature review. International Journal of Computer Vision pp. 1–25 (2021)
5. Dai, Q., Zhao, R.W., Wu, Z., Wang, X., Gu, Z., Wu, W., Jiang, Y.G.: Fudan-Huawei at MediaEval 2015: Detecting violent scenes and affective impact in movies with deep learning. In: Working Notes Proceedings of the MediaEval 2015 Workshop. CEUR Workshop Proceedings, vol. 1436. CEUR-WS.org (2015)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
7. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., Sompel, H.: The europeana data model (edm). World Library and Information Congress: 76th IFLA General Conference and Assembly pp. 10–15 (01 2010)
8. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Müller, H.: Overview of ImageCLEFcaption 2017 – the image caption prediction and concept extraction tasks to understand biomedical images. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2017). CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017)
9. Foundation, E.: Europeana (2022), https://www.europeana.eu/
10. García Seco De Herrera, A., Eickhof, C., Andrearczyk, V., Müller, H.: Overview of the ImageCLEF 2018 caption prediction tasks. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2018). CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018)
11. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Advances in Information Retrieval – 27th European Conference on IR Research (ECIR 2005). pp. 345–359. Springer (2005)
12. Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P., Parasa, S.: On evaluation metrics for medical applications of artificial intelligence. Scientific Reports **12**(1), 1–9 (2022)
13. Ionescu, B., Rohm, M., Boteanu, B., Gînscă, A.L., Lupu, M., Müller, H.: Benchmarking image retrieval diversification techniques for social media. IEEE Transactions on Multimedia **23**, 677–691 (2020)
14. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics **39**(0), 55 – 61 (2015)
15. Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: a simple approach to sentiment and style transfer. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1865–1874. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1169, https://aclanthology.org/N18-1169

16. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
17. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): ImageCLEF – Experimental Evaluation in Visual Information Retrieval, The Springer International Series On Information Retrieval, vol. 32. Springer, Berlin Heidelberg (2010)
18. Müller, H., Kalpathy-Cramer, J., García Seco de Herrera, A.: Experiences from the ImageCLEF medical retrieval and annotation tasks. In: Information Retrieval Evaluation in a Changing World, pp. 231–250. Springer (2019)
19. Pangeanic, Anacode, Foundation, E.: The recommendation system (2022), https://pro.europeana.eu/page/the-recommendation-system
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL 2002). pp. 311–318 (2002)
21. Pelka, O., Abacha, A.B., García Seco de Herrera, A., Jacutprakart, J., Friedrich, C.M., Müller, H.: Overview of the ImageCLEFmed 2021 concept & caption prediction task. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2021). CEUR Workshop Proceedings, vol. 2936. CEUR-WS.org (2021)
22. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2019 concept detection task. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2019). CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019)
23. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2020). CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020)
24. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 180–189. Springer (2018)
25. Rückert, J., Ben Abacha, A., García Seco de Herrera, A., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Müller, H., Friedrich, C.M.: Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection. In: CLEF2022 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy (September 5-8 2022)
26. Sellam, T., Das, D., Parikh, A.: BLEURT: Learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7881–7892. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.704, https://aclanthology.org/2020.acl-main.704
27. Ştefan, L.D., Constantin, M.G., Dogariu, M., Ionescu, B.: Overview of imagecleffusion 2022 task-ensembling methods for media interestingness prediction and result diversification. In: Working Notes of Conference and Labs of the Evaluation Forum (CLEF 2022). CEUR Workshop Proceedings, CEUR-WS.org (2022)
28. Sudhakaran, S., Escalera, S., Lanz, O.: Gate-shift networks for video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020). pp. 1102–1111 (2020)
29. Tsikrika, T., García Seco de Herrera, A., Müller, H.: Assessing the scholarly impact of ImageCLEF. In: CLEF 2011. pp. 95–106. Springer Lecture Notes in Computer Science (LNCS) (2011)

30. Wang, S., Chen, S., Zhao, J., Jin, Q.: Video interestingness prediction based on ranking model. In: Proceedings of the joint workshop of the 4th workshop on affective social multimedia computing and first multi-modal affective computing of large-scale multimedia data (ASMMC-MMAC 2018). pp. 55–61. Association for Computing Machinery (ACM) (2018)
31. Xu, W., Saxon, M., Sra, M., Wang, W.Y.: Self-supervised knowledge assimilation for expert-layman text style transfer. Proceedings of the AAAI Conference on Artificial Intelligence **36**(10), 11566–11574 (jun 2022). https://doi.org/10.1609/aaai.v36i10.21410, https://ojs.aaai.org/index.php/AAAI/article/view/21410
32. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020 (2020), https://openreview.net/forum?id=SkeHuCVFDr