EVENT REPORT

Report on the 15th Conference and Labs of the Evaluation Forum (CLEF 2024): Experimental IR Meets Multilinguality, Multimodality, and Interaction

Giorgio Maria Di Nunzio
University of Padu
Italy
giorgiomaria.dinunzio@unipd.it

Guglielmo Faggioli
University of Padua
Italy
faggioli@dei.unipd.it

Nicola Ferro
University of Padua
Italy
nicola.ferro@unipd.it

Petra Galuščáková
University of Stavanger
Norway
petra.galuscakova@uis.no

Alba García Seco de Herrera
National Distance Education University (UNED)
Spain
alba.garcia@lsi.uned.es

Lorraine Goeuriot
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble
France
lorraine.goeuriot@univ-grenoble-alpes.fr

Philippe Mulhem
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble
France
Philippe.Mulhem@imag.fr

Georges Quénot
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble
France
Georges.Quenot@imag.fr

Didier Schwab
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble
France
didier.schwab@univ-grenoble-alpes.fr

Laure Soulier
Sorbonne Université, CNRS, ISIR, F-75005 Paris
France
laure.soulier@isir.upmc.fr

Abstract

This is a report on the fifteenth edition of the *Conference and Labs of the Evaluation Forum* (CLEF 2024), held on September 9–12, 2024, in Grenoble, France. CLEF was a four-day hybrid event combining a conference and an evaluation forum. This edition also marked a special event, since we celebrated the 25th anniversary of CLEF. The conference featured keynotes by Paula Carvalho and Aurélie Névéol, and presentation of peer-reviewed research papers covering a wide range of topics, in addition to many posters. The evaluation forum consisted of fourteen labs: BioASQ, CheckThat!, ELOQUENT, eRisk, EXIST, iDPP, ImageCLEF, JOKER, LifeCLEF, LongEval, PAN, qCLEF, SimpleText, and Touché, addressing a wide range of tasks, media, languages, and ways to go beyond standard test collections.

Date: 9–12 September, 2024.

Website: https://clef2024.clef-initiative.eu/.

1 Introduction

The 2024 edition of the *Conference and Labs of the Evaluation Forum* (CLEF) was organized by the University of Grenoble Alpes, Grenoble, France, from 9 to 12 September 2024. CLEF 2024 was the 15th year of the CLEF Conference and the 25th year of the CLEF initiative as a forum for IR Evaluation, so it marked an important anniversary for CLEF [Ferro, 2024].

The conference format remained the same as in previous years, and consisted of keynotes, contributed papers, lab sessions, and poster sessions, including reports from other benchmarking initiatives from around the world. All sessions were organized and run in hybrid mode, allowing for both in-presence and remote attendance.

CLEF was established in 2000 as a spin-off of the TREC Cross-Language Track, with a focus on stimulating research and innovation in multimodal and multilingual information access and retrieval [Ferro, 2019; Ferro and Peters, 2019]. Over the years, CLEF has fostered the creation of language resources in many European and non-European languages, promoted the growth of a vibrant and multidisciplinary research community, provided sizable improvements in the performance of monolingual, bilingual, and multilingual information access systems [Ferro and Silvello, 2017], and achieved a substantial scholarly impact [Larsen, 2019; Tsikrika et al., 2011, 2013].

In its first 10 years, CLEF hosted a series of experimental labs that reported their results at an annual workshop held in conjunction with the European Conference on Digital Libraries (ECDL, now TPDL). In 2010, by then a mature and well-respected evaluation forum, CLEF was expanded to include a complementary peer-reviewed conference, focused on discussing the advancement of evaluation methodologies and on reporting evaluations of information access and retrieval systems regardless of data type, format, language, and others. Moreover, the scope of the evaluation labs was broadened, to include not only multilinguality but also multimodality in information access. Multimodality is here intended as the ability to deal with information not only conveyed by multiple media, but also coming in different modalities, e.g. the Web, social media, news streams, specific domains, and so on. Since 2010, the CLEF conference has established a format which includes keynotes, contributed papers, lab sessions, and poster sessions, including reports from other benchmarking initiatives from around the world. Since 2013, CLEF has been supported by an association, a lightweight non-for-profit legal entity that, thanks to the financial support of the CLEF community, takes care of the small central coordination needed to operate CLEF on an ongoing basis and makes it a self-sustaining activity [Ferro, 2019, 2024].

CLEF 2024 continued the initiative introduced in the 2019 edition, during which the *European Conference for Information Retrieval (ECIR)* and CLEF joined forces: ECIR 2024 hosted a special session dedicated to CLEF Labs where lab organizers presented the major outcomes of their Labs and their plans for ongoing activities, followed by a poster session to favour discussion during the conference. This was reflected in the ECIR 2024 proceedings [Nazli et al., 2024a,b], where CLEF Lab activities and results were reported as short papers. The goal was not only to engage the ECIR community in CLEF activities but also to disseminate the research results achieved during CLEF evaluation cycles as submission of papers to ECIR.

CLEF 2024 was attended by 218 participants, out of which 168 in-presence and 50 remotely, denoting a vibrant community, from different academic institutions and industrial organizations. Although the majority (72%) of the participants came from different European countries, there was also considerable worldwide interest in CLEF 2024, with 8% participants from Asia, 8% from the Americas, 2% from Oceania, and 3% from Africa.

2 The CLEF Conference

CLEF 2024 continued the focus of the CLEF conference on "experimental IR", as carried out at evaluation forums (CLEF Labs, TREC, NTCIR, FIRE, MediaEval, etc.), with special attention to the challenges of multimodality, multilinguality, and interactive search. We invited submissions on significant new insights demonstrated on IR test collections, on analyses of IR test collections and evaluation measures, and on concrete proposals to push the boundaries of the Cranfield/TREC/CLEF paradigm [Goeuriot et al., 2024a,b]

Keynotes The following scholars were invited to give a keynote talk at the CLEF 2024 conference:

Paula Carvalho (University of Aveiro and INESC-ID Lisboa, Portugal) delivered a talk entitled "An Interdisciplinary and culturally sensitive approach to unpack online hate speech". Here is the abstract of her talk: "In the digital age, online hate speech (OHS) poses a significant threat to democratic societies by fostering discrimination, violence, and social division while normalizing harmful behavior. Automation is essential for providing real-time, effective responses to the vast amount of OHS disseminated daily. However, the lack of consensus on the definition of hate speech and its broad scope presents significant challenges in developing accurate identification strategies. Moreover, automated solutions often overlook the nuanced intricacies underlying hateful discourse, neglecting subtle or indirect manifestations. Advanced language models have the potential to capture complex hate speech manifestations, yet their efficacy depends on access to extensive and diverse data, which is often limited, especially in resource-scarce languages. In addition, understanding hate speech, particularly its subtle forms, requires a deep examination of its content within social, historical, and cultural contexts, which are frequently overlooked in current detection approaches. In this talk, Paula discussed how we address these challenges within the kNOwHATE project (https://knowhate.eu/), employing a comprehensive, culturally sensitive approach grounded in social psychology and language sciences. The talk also covered key findings from the project and discuss future directions."

Aurélie Névéol (LISN, Université Paris-Saclay, France) gave a speech "Evaluation in the era of Large Language Models". Here is the abstract of her talk: "Large Language models have brought about a paradigm shift in Natural Language Processing by creating tools that generate natural language texts with unprecedented fluency in well resourced languages. Furthermore, language models are becoming ubiquitous and easily accessible to the general public. In this talk Aurélie discussed how this impacts the way evaluations are conducted in theory and in practice. She explained how the nature of large language models can make it difficult to efficiently deploy classic evaluation practices. She also outlined critical evaluation dimensions beyond task performance and pointed out the urgency to include social and environmental impact aspects to conduct comprehensive evaluations in the field of Natural Language Processing and Information Retrieval."

Technical Program CLEF 2024 received a total of 25 scientific submissions, of which a total of 11 papers (7 long, 3 short & 1 position) were accepted. Each submission was reviewed by at least two program committee members, and the program chairs oversaw the reviewing and follow-up discussions. Several papers were a product of international collaboration. This year, researchers addressed the following important challenges in the community: factual reporting and political bias; sexism, discrimination, and misinformation; information retrieval and recommendation; information retrieval for decision making; document sanitization for information release and retrieval; evaluation dataset for knowledge acquisition; evaluation with gen-IR; medical entity linking; and classification with large language models.

CLEF 2024 continued the new *result-less review process* introduced at CLEF and inspired by the "Dagstuhl Seminar 23031 on Frontiers of Information Access Experimentation for Research and Education" [Bauer et al., 2023a,b], where the reviewing process is split in two parts. Firstly, the papers are reviewed by their methodological contribution, their research questions, and their experimental design – the submitted papers do not contain an experimentation part. Secondly, those papers which pass the first step are then reviewed by their experimentation, analyses, and insights. The purpose of this new review process is: (i) to avoid accepting papers just because performance improvements with respect to some baseline; (ii) to ensure they have a grounded methodology; and, (iii) to verify that the research questions are driven by the methodology and not, post-hoc, by the experimental results.

Like in previous editions, since 2015, CLEF 2024 continued inviting CLEF lab organizers to nominate a "best of the labs" paper, among those submitted in the CLEF 2023 labs, that was reviewed as a full paper submission to the CLEF 2024 conference, according to the same review criteria and PC. 6 full papers were accepted for this "best of the labs" section.

3 The CLEF Lab Sessions

A total of 23 lab proposals were received and evaluated in peer review based on their innovation potential and the quality of the resources created. To identify the best proposals, well-established criteria from previous editions of CLEF were applied, like, for example, topical relevance, novelty, potential impact on future world affairs, likely number of participants, and the quality of the organizing consortium. This year we further stressed the connection to real-life usage scenarios, and we tried to avoid, as much as possible, overlaps among labs, in order to promote synergies and integration.

The 14 selected labs represented scientific challenges based on new datasets and real-world problems in multimodal and multilingual information access. These datasets provide unique opportunities for scientists to explore collections, to develop solutions for these problems, to receive feedback on the performance of their solutions, and to discuss related challenges with peers at the workshops. In addition to these workshops, the labs reported results of their year-long activities in overview talks and lab sessions.

The 14 labs running as part of CLEF 2024 comprised mainly labs that continued from previous editions at CLEF (BioASQ, CheckThat!, eRisk, EXIST, iDPP, ImageCLEF, JOKER, LifeCLEF, LongEval, PAN, SimpleText, and Touché) and new pilot/workshop activities (ELOQUENT and qCLEF). Details of the individual labs are described by the lab organizers in the CLEF Working Notes [Faggioli et al., 2024]. We only provide a brief overview of them here (in alphabetical order).

BioASQ: Large-scale biomedical semantic indexing and question answering[1] [Nentidis et al., 2024] aims to push the research frontier towards systems that use the diverse and voluminous information available online to respond directly to the information needs of biomedical scientists. It offered the following tasks. *Task 1 - b: Biomedical Semantic Question Answering*: benchmark datasets of biomedical questions, in English, along with gold standard (reference) answers constructed by a team of biomedical experts. The participants have to respond with relevant articles, and snippets from designated resources, as well as exact and "ideal" answers. *Task 2 - Synergy: Question Answering for developing problems*: biomedical experts pose unanswered questions for developing problems, such as COVID-19, receive the responses provided by the participating systems, and provide feedback, together with updated questions in an iterative procedure that aims to facilitate the incremental understanding of developing problems in biomedicine and public health. *Task 3 - MultiCardioNER: Multiple clinical entity detection in multilingual medical content*: focuses on the automatic detection and normalization of mentions of four clinical entity types, namely diseases, symptoms, procedures and medications, in cardiology clinical case documents in Spanish, English, Italian and Dutch. *BioNNE: Nested NER in Russian and English*: deals with nested named entity recognition (NER) in PubMed abstracts in Russian and English. The train/dev datasets include annotated mentions of disorders, anatomical structures, chemicals, diagnostic procedures, and biological functions. Participants are encouraged to apply cross-language (Russian to English) and cross-domain techniques.

[1] http://www.bioasq.org/workshop2024

CheckThat! Lab on Checkworthiness, Subjectivity, Persuasion, Roles, Authorities and Adversarial Robustness[2] [Barrón-Cedeño et al., 2024] provides a diverse collection of challenges to the research community interested in developing technology to support and understand the journalistic verification process. The tasks go from core verification tasks such as assessing the check-worthiness of a text to understanding the strategies used to influence the audience and identifying the stance of relevant characters on questionable affair. It offered the following tasks. *Task 1 - Check-worthiness estimation*: asks to assess whether a statement, sourced from either a tweet or a political debate, warrants fact-checking. *Task 2 - Subjectivity*: given a sentence from a news article, it asks to determine whether it is subjective or objective. *Task 3 - Persuasion Techniques*: given a news article and a list of 23 persuasion techniques organized into a 2-tier taxonomy, including logical fallacies and emotional manipulation techniques that might be used to support flawed argumentation, it asks to identify the spans of texts in which each technique occurs. *Task 4 - Detecting hero, villain, and victim from memes*: ask to determine the roles of entities within memes, categorizing them as "hero", "villain", "victim", or "other" through a multi-class classification approach that considers the systematic modeling of multimodal semiotic. *ask 5 - Authority Evidence for Rumor Verification*: given a rumor expressed in a tweet and a set of authorities for that rumor, it asks to retrieve up to 5 evidence tweets from the authorities' timelines, and determine if the rumor is supported, refuted, or unverifiable according to the evidence. *Task 6 - Robustness of Credibility Assessment with Adversarial Examples*: the task is realised in five domains: style-based news bias assessment (HN), propaganda detection (PR), fact checking (FC), rumour detection (RD) and COVID-19 misinformation detection (C19). For each domain, the participants are provided with three victim models, trained for the corresponding binary classification task, as well as a collection of 400 text fragments. Their aim is to prepare adversarial examples, which preserve the meaning of the original examples, but are labelled differently be the classifiers.

ELOQUENT shared tasks for evaluation of generative language model quality[3] [Karlgren et al., 2024] provides a set a of tasks for evaluating the quality of generative language models. It offered the following tasks. *Task 1 - Topical competence*: tests and verifies a model's understanding of an application domain and specific topic of interest. *Task 2 - Veracity and hallucination*: tests how the truthfulness or veracity of automatically generated text can be assessed. *Task 3 - Robustness*: tests the capability of a model to handle input variation – e.g. dialectal, sociolectal, and cross-cultural – as represented by a set of equivalent but non-identical varieties of input prompts. *Task 4 - Voight Kampff*: explores whether automatically-generated text can be distinguished from human-authored text. This task is organized in collaboration with the PAN lab at CLEF.

eRisk: Early Risk Prediction on the Internet[4] [Parapar et al., 2024] explores the evaluation methodology, effectiveness metrics and practical applications (particularly those related to health and safety) of early risk detection on the Internet. It offered the following tasks. *Task 1 - Search for symptoms of depression*: consists of ranking sentences from a collection of user

[2] http://checkthat.gitlab.io/

[3] https://eloquent-lab.github.io/

[4] https://erisk.irlab.org/

writings according to their relevance to a depression symptom. The participants will have to provide rankings for the 21 symptoms of depression from the BDI Questionnaire. *Task 2 - Early Detection of Signs of Anorexia*: consists in performing a task on early risk detection of anorexia. The challenge consists of sequentially processing pieces of evidence and detect early traces of anorexia as soon as possible. *Task 3 - Measuring the severity of the signs of Eating Disorders*: consists of estimating the level of features associated with a diagnosis of eating disorders from a thread of user submissions. For each user, the participants will be given a history of postings and the participants will have to fill a standard eating disorder questionnaire.

EXIST: sEXism Identification in Social neTworks[5] [Plaza et al., 2024] aims to capture and categorize sexism, from explicit misogyny to other subtle behaviors, in social networks. Participants will be asked to classify tweets in English and Spanish according to the type of sexism they enclose and the intention of the persons that writes the tweets. It offered the following tasks. *Task 1 - Sexism Identification in Tweets*: is a binary classification. The systems have to decide whether or not a given tweet contains sexist expressions or behaviours (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour). *Task 2 - Source Intention in Tweets*: aims to categorize the message according to the intention of the author, which provides insights in the role played by social networks on the emission and dissemination of sexist messages. *Task 3 - Sexism Categorization in Tweets*: many facets of a woman's life may be the focus of sexist attitudes including domestic and parenting roles, career opportunities, sexual image, and life expectations, to name a few. Automatically detecting which of these facets of women are being more frequently attacked in social networks will facilitate the development of policies to fight against sexism. *Task 4 - Sexism Identification in Memes*: is a binary classification task consisting on deciding whether or not a given meme is sexist. *Task 5 - Source Intention in Memes*: aims to categorize the meme according to the intention of the author, which provides insights in the role played by social networks on the emission and dissemination of sexist messages.

iDPP: Intelligent Disease Progression Prediction[6] [Birolo et al., 2024] Amyotrophic Lateral Sclerosis (ALS) and Multiple Sclerosis (MS) are chronic diseases characterized by progressive or alternate impairment of neurological functions (motor, sensory, visual, cognitive). Patients have to manage alternated periods in hospital with care at home, experiencing a constant uncertainty regarding the timing of the disease acute phases and facing a considerable psychological and economic burden that also involves their caregivers. Clinicians, on the other hand, need tools able to support them in all the phases of the patient treatment, suggest personalized therapeutic decisions, indicate urgently needed interventions. It offered the following tasks. *Task 1 – Predicting ALSFRS-R score from sensor data (ALS)*: focuses on predicting the ALSFRS-R score (ALS Functional Rating Scale - Revised), assigned by medical doctors roughly every three months, from the sensor data collected via the app. The ALSFRS-R score is a somehow "subjective" evaluation performed by a medical doctor and this task will help in answering a currently open question in the research community, i.e. whether it could be derived from objective factors. *Task 2 – Predicting*

[5] http://nlp.uned.es/exist2024/

[6] https://brainteaser.health/open-evaluation-challenges/idpp-2024/

patient self-assessment score from sensor (ALS): focuses on predicting the self-assessment score assigned by patients from the sensor data collected via the app. If the self-assessment performed by patients, more frequently than the assessment performed by medical doctors every three months or so, can be reliably predicted by sensor and app data, we can imagine a proactive application which, monitoring the sensor data, alerts the patient if an assessment is needed. *Task 3 – Predicting relapses from EDDS sub-scores and environmental data (MS))*: focuses on predicting a relapse using environmental data and EDSS (Expanded Disability Status Scale) sub-scores. This task will allow us to assess if exposure to different pollutants is a useful variable in predicting a relapse.

ImageCLEF: Multimedia Retrieval[7] [Ionescu et al., 2024] is aimed at evaluating the technologies for annotation, indexing, classification and retrieval of multimodal data. Its main objective resides in providing access to large collections of multimodal data for multiple usage scenarios and domains. Considering the experience of the last four successful editions, ImageCLEF 2024 will continue approaching a diversity of applications, namely medical, social media and Internet, and recommending, giving to the participants the opportunity to deal with interdisciplinary approaches and domains. It offered the following tasks. *Task 1 - ImageCLEFmedical*: continues the tradition of bringing together several initiatives for medical applications fostering cross-exchanges, namely: (i) caption task with medical concept detection and caption prediction, (ii) GAN task on synthetic medical images generated with GANs, (iii) MEDVQA-GI task for medical images generation based on text input, and (iv) Mediqa task with a new use-case on multimodal dermatology response generation. *Task 2 - Image Retrieval/Generation for Arguments*: given a set of arguments, asks to return for each argument several images that help to convey the argument's premise, that is, suitable images could depict what is described in the argument. *Task 3 - ImageCLEFrecommending*: 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 benchmarking different recommendation systems and methods. *Task 4 - ImageCLEFtoPicto*: aims to provide a translation in pictograms from a natural language, either from (i) text or (ii) speech understandable by the users, in this case, people with language impairments as pictogram generation is an emerging and significant domain in natural language processing, with multiple potential applications, enabling communication with individuals who have disabilities, aiding in medical settings for individuals who do not speak the language of a country, and also enhancing user understanding in the service industry..

JOKER: Automatic Humour Analysis[8] [Ermakova et al., 2024a] aims to foster research on automated processing of verbal humour, including tasks such as retrieval, classification, interpretation, generation, and translation. It offered the following tasks. *Task 1 - Humour-aware information retrieval*: aims at retrieving short humorous texts from a document collection. *Task 2 - Humour classification according to genre and technique*: aims at classifying short texts of humor among the different classes such as Irony, Sarcasm, Exaggeration,

[7] https://www.imageclef.org/2024

[8] http://joker-project.com/

Incongruity, Absurdity, etc. *Task 3 - Pun translation*: aims to translate English punning jokes into French preserving wordplay form and wordplay meaning.

LifeCLEF: species identification and prediction[9] [Joly et al., 2024] is dedicated to the large-scale evaluation of biodiversity identification and prediction methods based on artificial intelligence. It offered the following tasks. *Task 1 - BirdCLEF*: bird species recognition in audio soundscapes. *Task 2 - FungiCLEF*: fungi recognition from images and metadata. *Task 3 - GeoLifeCLEF*: remote sensing based prediction of species. *Task 4 - PlantCLEF*: global-scale plant identification from images. *Task 5 - SnakeCLEF*: snake species identification in medically important scenarios.

LongEval: Longitudinal Evaluation of Model Performance[10] [Alkhalifa et al., 2024] is focused on evaluating the temporal persistence of information retrieval systems and text classifiers. The goal is to develop temporal information retrieval systems and longitudinal text classifiers that survive through dynamic temporal text changes, introducing time as a new dimension for ranking models performance. It offered the following tasks. *Task 1 - LongEval-Retrieval*: aims to propose a temporal information retrieval system which can handle changes over the time. The proposed retrieval system should follow the temporal persistence on Web documents. This task will have 2 sub-tasks focusing on short-term and long-term persistence. *Task 2 - LongEval-Classification* aims to propose a temporal persistence classifier which can mitigate performance drop over short and long periods of time compared to a test set from the same time frame as training. This task will have 2 sub-tasks focusing on short-term and long-term persistence.

PAN: Digital Text Forensics and Stylometry[11] [Ayele et al., 2024] aims to advance the state of the art and provide for an objective evaluation on newly developed benchmark datasets in those areas. It offered the following tasks. *Task 1 - Multi-Author Writing Style Analysis*: given an English document, asks to determine at which paragraphs the author changes. Examples vary in difficulty from easy to hard depending on topical homogeneity of the paragraphs. *Task 2 - Multilingual Text Detoxification*: given a toxic piece of text, asks to re-write it in a non-toxic way while saving the main content as much as possible. Texts are provided in 7 languages. *Task 3 - Oppositional Thinking Analysis*: given an English or Spanish online message, asks to determine if it is a conspiracy theory or critical thinking. In former case, find the core elements of the conspiracy narrative. *Task 4 - Generative AI Authorship Verification*: given a document, asks to determine if the author is a human or a language model. In collaboration with the ELOQUENT lab.

qCLEF: QuantumCLEF[12] [Pasin et al., 2024] Quantum Computing (QC) is a rapidly growing field, involving an increasing number of researchers and practitioners from different backgrounds to develop new methods that leverage quantum computers to perform faster computations. QuantumCLEF provides an evaluation infrastructure to design and develop QC algorithms and, in particular, for Quantum Annealing (QA) algorithms, for Information

[9] http://www.lifeclef.org/

[10] https://clef-longeval.github.io/

[11] http://pan.webis.de/

[12] https://qclef.dei.unipd.it/

Retrieval and Recommender Systems. It offered the following tasks. *Task 1 - Feature Selection*: focuses on applying quantum annealers to find the most relevant subset of features to train a learning model, e.g., for ranking. This problem is very impactful, since many IR and RS systems involve the optimization of learning models, and reducing the dimensionality of the input data can improve their performance. *Task 2 - Clustering*: focuses on using quantum annealing to cluster different documents in the form of embeddings to ease the browsing process of large collections. Clustering can be helpful for organizing large collections, helping users to explore a collection and providing similar search results to a given query. Furthermore, it can be helpful to divide users according to their interests or build user models with the cluster centroids speeding up the runtime of the system or its effectiveness for users with limited data. Clustering is however a very complex task in the case of QA since it is possible to perform clustering only considering a limited number of items and clusters due to the architecture of quantum annealers. A baseline using K-medoids clustering with cosine distance will be used as an overall alternative.

SimpleText: Improving Access to Scientific Texts for Everyone[13] [Ermakova et al., 2024b] addresses technical and evaluation challenges associated with making scientific information accessible to a wide audience, students, and experts. We provide appropriate reusable data and benchmarks for scientific text summarization and simplification. *Task 1 - Retrieving passages to include in a simplified summary*: given a popular science article targeted to a general audience, aims at retrieving passages, which can help to understand this article, from a large corpus of academic abstracts and bibliographic metadata. Relevant passages should relate to any of the topics in the source article. . *Task 2 - Identifying and explaining difficult concepts*: aims to decide which concepts in scientific abstracts require explanation and contextualization in order to help a reader understand the scientific text. *Task 3 - Simplify Scientific Text*: aims to provide a simplified version of sentences extracted from scientific abstracts. Participants will be provided with the popular science articles and queries and matching abstracts of scientific papers, split into individual sentences. *Task 4 - Tracking the State-of-the-Art in Scholarly Publications*: aims to develop systems which given the full text of an AI paper, are capable of recognizing whether an incoming AI paper indeed reports model scores on benchmark datasets, and if so, to extract all pertinent (Task, Dataset, Metric, Score) tuples presented within the paper.

Touché: Argumentation Systems[14] [Kiesel et al., 2024] aims to to foster the development of technologies that support people in decision-making and opinion-forming and to improve our understanding of these processes. It offered the following tasks. *Task 1 - Human Value Detection*: given a text, for each sentence, asks to detect which human values the sentence refers to and whether this reference (partially) attains or (partially) constrains the value. *Task 2 - Ideology and Power Identification in Parliamentary Debates*: given a parliamentary speech in one of several languages, asks to identify the ideology of the speaker's party and identify whether the speaker's party is currently governing or in opposition. *Task 3 - Image Retrieval for Arguments*: given an argument, asks to retrieve or generate images that help to convey the argument's premise.

[13] http://simpletext-project.com/

[14] https://touche.webis.de/

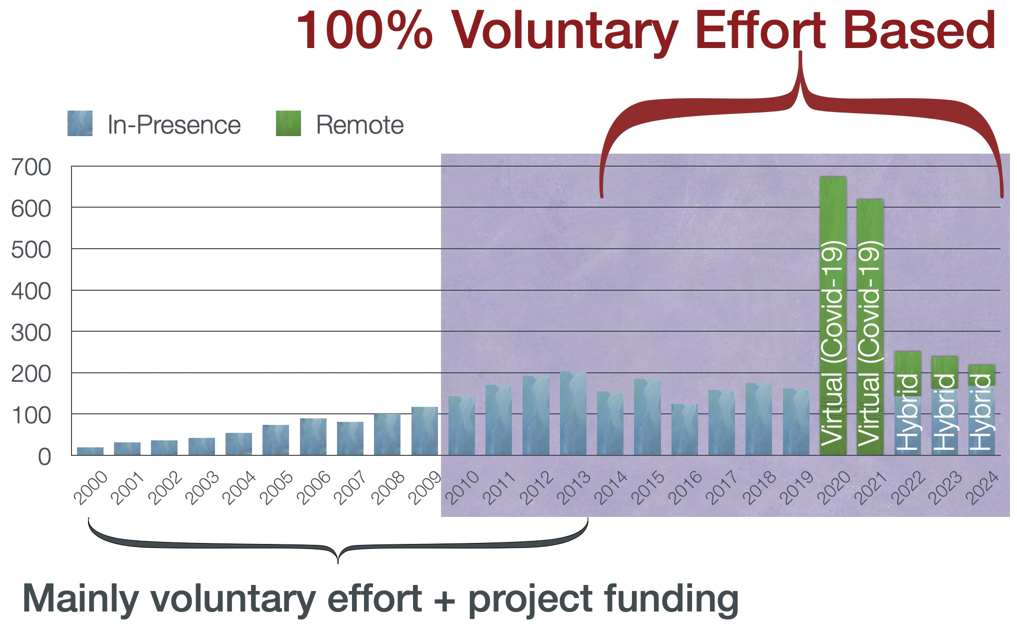

Figure 1. Attendance to CLEF over the years: x-axis reports CLEF editions; y-axis the number of attendees; the shading indicates the change from CLEF as a workshop, co-located with ECDL/TPDL, to CLEF as an independent conference and labs.

More information on the CLEF 2024 conference, the CLEF initiative and the CLEF Association is provided on the Web:

- CLEF 2024: https://clef2024.clef-initiative.eu/
- CLEF initiative: https://www.clef-initiative.eu/
- CLEF Association: https://www.clef-initiative.eu/#association

4 Overall Trends for CLEF

Figure 1 shows the attendance trends to CLEF since its inception. We can observe that there has been a substantial growth over the years, especially since when it was backed by the CLEF Association. We can also note that CLEF 2020 and CLEF 2021, which were online only and with almost free registration due to COVID-19, represent a spike in the attendance. The in-presence attendance for CLEF 2024 has been substantially comparable to the pre-COVID editions, mostly stable with respect to CLEF 2023 (168 vs 161 in-presence participants) and growing with respect to CLEF 2022 (168 vs 143 in-presence participants), while the overall participation has increased compared to the pre-COVID editions, thanks to the remote participants.

Figure 2 shows the number of papers published in the Working Notes over the years; we report the Working Notes because they contain both the labs overviews and all the participant papers. We can observe how the increase in participation to CLEF has been accompanied by an increase in the publication output, plus 28% with respect to CLEF 2023. Note that both the Working Notes and the Conference Proceedings are fully peer-reviewed venues.

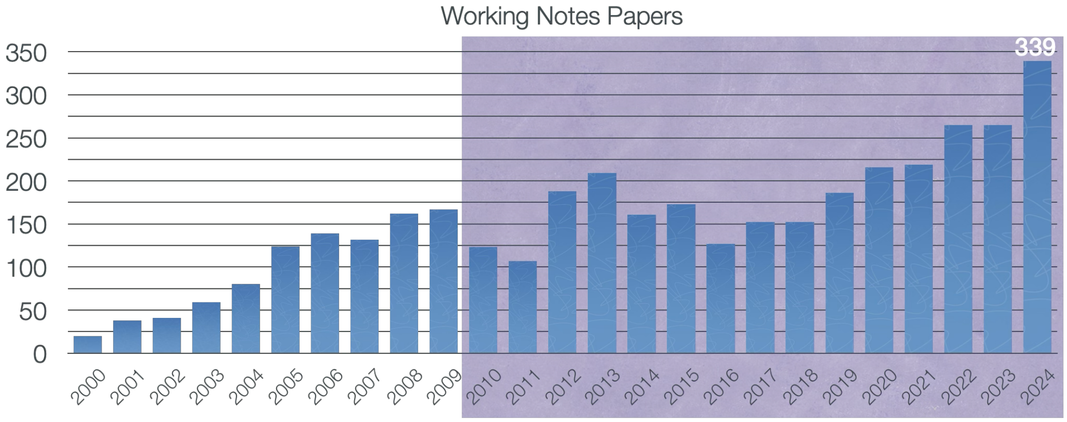

Figure 2. Papers published in the working notes over the years: x-axis reports CLEF editions; y-axis the number of papers in the working notes, highlighting 265, the papers in the CLEF 2023 working notes; the shading indicates the change from CLEF as a workshop co-located with ECDL/TPDL to CLEF as an independent conference and labs.

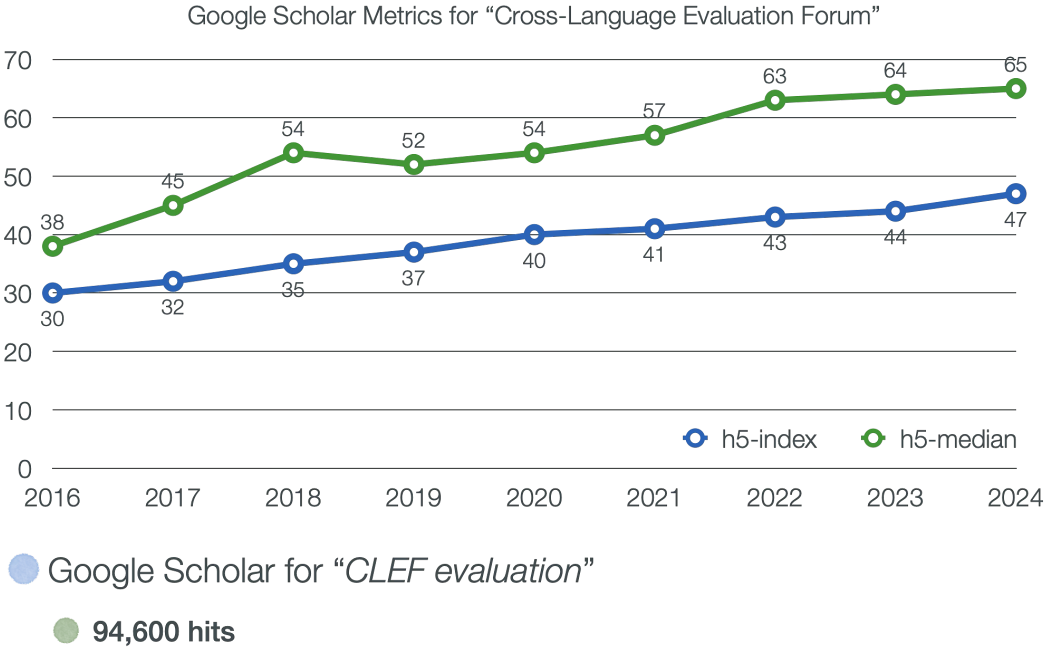

Figure 3. Google Scholar metrics for "Cross-Language Evaluation Forum" since 2016: the x-axis reports years, the y-axis the value for the h5-index (the largest number h such that at least h articles in that publication were cited at least h times each, only those of its articles that were published in the last five complete calendar years) and h5-median (the median number of citations for the articles that make up the h5-index).

Finally, Figure 3 shows the Google Scholar metrics for CLEF[15] since 2016; also in this case we can observe a positive growth trend, giving an idea of the impact of CLEF. In particular, CLEF is listed among the top-20 venues for the sub-category "Databases & Information Systems"[16], together with other important venues for the IR community, like SIGIR, CIKM, RecSys, and WWW.

5 CLEF 2025

CLEF 2025 will be hosted by UNED University, Madrid, Spain, on 9–12 September 2025.

More information on CLEF 2025, the call for papers and the ongoing labs is available at:

- https://clef2024.clef-initiative.eu/

As far as labs are concerned, CLEF 2024 will run 14 evaluation activities out of 20 proposals received: thirteen will be a continuation of the labs running during CLEF 2024:

- BioASQ – A challenge in large-scale biomedical semantic indexing and question answering[17];
- CheckThat! – Lab on Subjectivity, Fact-Checking, Claim Extraction & Normalization, and Retrieval[18];
- ELOQUENT – Lab for evaluation of generative language model quality[19];
- eRisk – Early risk prediction on the Internet[20];
- EXIST – sEXism Identification in Social neTworks[21];
- ImageCLEF – Multimodal Challenge in CLEF[22];
- JOKER – Humour in the Machine[23];
- LifeCLEF – Challenges on Species Presence Prediction and Identification, and Individual Animal Identification[24];
- LongEval – Longitudinal Evaluation of Model Performance[25].
- PAN – Lab on Stylometry and Digital Text Forensics[26];
- PAN – Lab on Stylometry and Digital Text Forensics[27];
- QuantumCLEF – Quantum Computing at CLEF[28];
- SimpleText – Simplify Scientific Text (and Nothing More)[29];

[15] Note that Google Scholar still indexes CLEF as "Cross-Language Evaluation Forum", even if the name has changed to "Conference and Labs of the Evaluation Forum" since 2010.

[16] https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_databasesinformationsystems

[17] https://www.bioasq.org/workshop2025

[18] http://checkthat.gitlab.io

[19] https://eloquent-lab.github.io/

[20] https://erisk.irlab.org/

[21] https://nlp.uned.es/exist2025/

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

[23] https://www.joker-project.com/

[24] http://www.lifeclef.org

[25] https://clef-longeval.github.io/

[26] https://pan.webis.de/

[27] https://pan.webis.de/

[28] https://qclef.dei.unipd.it/

[29] https://simpletext-project.com/

- Touché – Argumentation Systems[30];

and one will be a new pilot lab:

- TalentCLEF – Skill and Job Title Intelligence for Human Capital TalentCLEF aims to drive technological advancement in Human Capital Management by establishing a public benchmark for NLP models that facilitates their application in real-world Human Resources (HR) scenarios, incorporating evaluation criteria including multilingualism, fairness, and cross-industry adaptability. The lab also seeks to build a community for researchers and practitioners to generate, evaluate, and discuss ideas on the use of AI in Human Resources, pushing the state-of-the-art of NLP applications for Human Resources.

6 Bids for CLEF 2027

Bids for hosting CLEF 2027 are now open and will close on December 2024. Proposals can be sent to the CLEF Steering Committee Chair at chair@clef-initiative.eu and a template for bids is available here https://www.clef-initiative.eu/assets/CLEF-Template_for_bids.docx.

Acknowledgments

The success of CLEF 2024 would not have been possible without the huge effort of several people and organizations, including the CLEF Association, the program committee, the lab organizing committee, the local organization committee in Thessaloniki, the reviewers, and the many students and volunteers who contributed along the way.

We gratefully acknowledge the support we received from our supporters and sponsors: Université Grenoble Alpes[31], Grenoble INP - UGA[32], MIAI[33], ARIA[34], ATALA[35]. The best papers prices were sponsored by Springer. We also thank the Friends of SIGIR program[36] and the Information Retrieval Specialist Group (IRSG) of the BCS[37] for covering the registration fees for a number of student delegates.

Last but not least, without the important and tireless effort of the enthusiastic and creative authors, the organizers of the selected labs, the colleagues and friends involved in running them, and the participants who contribute their time to making the labs and the conference a success, as well as financially supporting them through the CLEF Association, CLEF would not be possible.

Thank you all very much!

[30] https://touche.webis.de/

[31] https://www.univ-grenoble-alpes.fr

[32] https://www.grenoble-inp.fr

[33] Multidisciplinary Institute in Artificial intelligence – https://miai.univ-grenoble-alpes.fr

[34] Association Francophone de Recherche d'Information et Applications – http://www.asso-aria.org

[35] Association pour le Traitement Automatique des Langues – https://www.atala.org

[36] http://sigir.org/general-information/funding-for-sigir-related-events/

[37] http://irsg.bcs.org/

References

R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, T. Fink, P. Galuščáková, G. N. González Sáez, L. Goeuriot, D. Iommi, M. Liakata, H. Tayyar Madabushi, P. Medina-Alias, P. Mulhem, F. Piroi, M. Popel, and A. Zubiaga. Overview of the CLEF 2024 LongEval Lab on Longitudinal Evaluation of Model Performance. In Goeuriot et al. [2024b].

A. A. Ayele, N. Babakov, J. Bevendorff, X. Bonet Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel Pardo, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, N. Wiegmann, S. Muhie Yimam, and E. Zangerle. Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification. In Goeuriot et al. [2024b].

A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, M. Hasanain, C. Li, J. Piskorski, F. Ruggeri, X. Song, and R. Suwaileh. Overview of the CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness. In Goeuriot et al. [2024b].

C. Bauer, B. Carterette, N. Ferro, N. Fuhr, J. Beel, T. Breuer, C. L. A. Clarke, A. Crescenzi, G. Demartini, G. M. Di Nunzio, L. Dietz, G. Faggioli, B. Ferwerda, M. Fröbe, M. Hagen, A. Hanbury, C. Hauff, D. Jannach, N. Kando, E. Kanoulas, B. P. Knijnenburg, U. Kruschwitz, M. Li, M. Maistro, L. Michiels, A. Papenmeier, M. Potthast, P. Rosso, A. Said, P. Schaer, C. Seifert, D. Spina, B. Stein, N. Tintarev, J. Urbano, H. Wachsmuth, M. C. Willemsen, and J. Zobel. Report on the Dagstuhl Seminar on Frontiers of Information Access Experimentation for Research and Education. *SIGIR Forum*, 57(1):7:1–7:28, June 2023a.

C. Bauer, B. A. Carterette, N. Ferro, N. Fuhr, and G. Faggioli, editors. *Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education*, Dagstuhl Reports, Volume 13, Number 1, 2023b. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany.

G. Birolo, P. Bosoni, G. Faggioli, H. Aidos, R. Bergamaschi, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. Garcia Dominguez, M. Gromicho, A. Guazzo, E. Longato, S. Madeira, U. Manera, S. Marchesin, L. Menotti, G. Silvello, E. Tavazzi, E. Tavazzi, I. Trescato, M. Vettoretti, B. Di Camillo, and N. Ferro. Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2024. In Goeuriot et al. [2024b], pages 118–139.

L. Ermakova, A.-G. Bosser, T. Miller, V. M. Palma Preciado, G. Sidorov, and A. Jatowt. Overview of JOKER @ CLEF-2024: Automatic Humour Analysis. In Goeuriot et al. [2024b].

L. Ermakova, E. Sanjuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, and J. Kamps. Overview of the CLEF 2024 SimpleText Track: Improving Access to Scientific Texts for Everyone. In Goeuriot et al. [2024b].

G. Faggioli, N. Ferro, P. Galuščáková, and A. García Seco de Herrera, editors. *CLEF 2024 Working Notes*, 2024. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, https://ceur-ws.org/Vol-3740/.

N. Ferro. What Happened in CLEF... For a While? In F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019)*, pages 3–45. Lecture Notes in Computer Science (LNCS) 11696, Springer, Heidelberg, Germany, 2019.

N. Ferro. What Happened in CLEF... For Another While? In Goeuriot et al. [2024a], pages 3–57.

N. Ferro and C. Peters, editors. *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, 2019. Springer International Publishing, Germany.

N. Ferro and G. Silvello. 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc-*ish* CLEF? *Information Processing & Management*, 53(1):175–202, January 2017.

L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, and N. Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024) – Part I*, 2024a. Lecture Notes in Computer Science (LNCS) 14958, Springer, Heidelberg, Germany.

L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, and N. Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024) – Part II*, 2024b. Lecture Notes in Computer Science (LNCS) 14959, Springer, Heidelberg, Germany.

B. Ionescu, H. Müller, A.-M. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco De Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A.-G. Andrei, Y. Prokopchuk, D. Karapenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W.-W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthas, and B. Stein. Overview of the ImageCLEF 2024: Multimedia retrieval in medical applications. In Goeuriot et al. [2024b].

A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Šulc, M. Hrúz, M. Servajean, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet, and H. Müller. Overview of LifeCLEF 2024: Challenges on Species Distribution Prediction and Identification. In Goeuriot et al. [2024b].

J. Karlgren, L. Dürlich, E. Gogoulou, L. Guillou, J. Nivre, M Sahlgren, A. Talman, and S. Zahra. Overview of ELOQUENT 2024 — shared tasks for evaluating generative language model quality. In Goeuriot et al. [2024b].

J. Kiesel, C. Çöltekin, H. Heinrich, M. Fröbe, M. Alshomary, B. De Longueville, T. Erjavec, N. Handke, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, T. Reitis-Münstermann, M. Scharfbillig, N. Stefanovitch, H. Wachsmuth, M. Potthast, and B. Stein. Overview of Touché 2024: Argumentation Systems. In Goeuriot et al. [2024b].

B. Larsen. The Scholarly Impact of CLEF 2010-2017. In N. Ferro and C. Peters, editors, *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 547–554. Springer International Publishing, Germany, 2019.

G. Nazli, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, and I. Ounis, editors. *Advances in Information Retrieval. Proc. 46th European Conference on IR Research (ECIR 2024) – Part V*, 2024a. Lecture Notes in Computer Science (LNCS) 14612, Springer, Heidelberg, Germany.

G. Nazli, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, and I. Ounis, editors. *Advances in Information Retrieval. Proc. 46th European Conference on IR Research (ECIR 2024) – Part VI*, 2024b. Lecture Notes in Computer Science (LNCS) 14613, Springer, Heidelberg, Germany.

A. Nentidis, G. Katsimpras, A. Krithara, S. Lima López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, and G. Paliouras. Overview of BioASQ 2024: The Twelfth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In Goeuriot et al. [2024b].

J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani. Overview of eRisk 2024: Early Risk Prediction on the Internet. In Goeuriot et al. [2024b].

A. Pasin, M. Ferrari Dacrema, P. Cremonesi, and N. Ferro. Overview of QuantumCLEF 2024: The Quantum Computing Challenge for Information Retrieval and Recommender Systems at CLEF. In Goeuriot et al. [2024b], pages 260–282.

L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, B. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, and D. Spina. Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes. In Goeuriot et al. [2024b].

T. Tsikrika, A. Garcia Seco de Herrera, and H. Müller. Assessing the Scholarly Impact of ImageCLEF. In P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, and M. de Rijke, editors, *Multilingual and Multimodal Information Access Evaluation. Proceedings of the Second International Conference of the Cross-Language Evaluation Forum (CLEF 2011)*, pages 95–106. Lecture Notes in Computer Science (LNCS) 6941, Springer, Heidelberg, Germany, 2011.

T. Tsikrika, B. Larsen, H. Müller, S. Endrullis, and E. Rahm. The Scholarly Impact of CLEF (2000–2009). In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the Fourth International Conference of the CLEF Initiative (CLEF 2013)*, pages 1–12. Lecture Notes in Computer Science (LNCS) 8138, Springer, Heidelberg, Germany, 2013.