Preface

The CLEF 2024 conference is the twenty-fifth edition of the popular CLEF campaign and workshop series that has run since 2000 contributing to the systematic evaluation of multilingual and multimodal information access systems, primarily through experimentation on shared tasks. In 2010 CLEF was launched in a new format, as a conference with research presentations, panels, poster and demo sessions and laboratory evaluation workshops. These are proposed and operated by groups of organizers volunteering their time and effort to define, promote, administrate and run an evaluation activity.

CLEF 2024[1] 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. The conference format remained the same as in past 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 in presence but also allowing for remote participation for those who were not able to attend physically.

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. 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 the 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.

We continued the mentorship program to support the preparation of lab proposals for newcomers to CLEF. The CLEF newcomers mentoring program offered help, guidance, and feedback on the writing of draft lab proposals by assigning a mentor to proponents, who helped them in preparing and maturing the lab proposal for submission. If the lab proposal fell into the scope of an already existing CLEF lab, the mentor helped proponents to get in touch with those lab organizers and team up forces.

Building on previous experience, the Labs at CLEF 2024 demonstrate the maturity of the CLEF evaluation environment by creating new tasks, new and larger data sets, new ways of evaluation or more languages. Details of the individual Labs are described by the Lab organizers in these proceedings.

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

[1] `https://clef2024.clef-initiative.eu/`

new pilot/workshop activities (ELOQUENT and qCLEF). In the following we give a few details for each of the labs organized at CLEF 2024 (presented in alphabetical order):

BioASQ: Large-scale biomedical semantic indexing and question answering[2] 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.

CheckThat! Lab on Checkworthiness, Subjectivity, Persuasion, Roles, Authorities and Adversarial Robustness[3] 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*

[2] `http://www.bioasq.org/workshop2024`
[3] `http://checkthat.gitlab.io/`

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[4] 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[5] 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 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.

[4] `https://eloquent-lab.github.io/`
[5] `https://erisk.irlab.org/`

EXIST: sEXism Identification in Social neTworks[6] 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[7] 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 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

[6] `http://nlp.uned.es/exist2024/`
[7] `https://brainteaser.health/open-evaluation-challenges/idpp-2024/`

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[8] 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[9] 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, Incon-

[8] `https://www.imageclef.org/2024`
[9] `http://joker-project.com/`

gruity, 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[10] 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[11] 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[12] 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 rewrite 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[13] 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

[10] `http://www.lifeclef.org/`
[11] `https://clef-longeval.github.io/`
[12] `http://pan.webis.de/`
[13] `https://qclef.dei.unipd.it/`

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 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[14] 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[15] 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 follow-

[14] `http://simpletext-project.com/`
[15] `https://touche.webis.de/`

ing 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.

CLEF has always been backed by European projects that complement the incredible amount of volunteering work performed by Lab Organizers and the CLEF community with the resources needed for its necessary central coordination, in a similar manner to the other major international evaluation initiatives such as TREC, NTCIR, FIRE and MediaEval. Since 2014, the organization of CLEF no longer has direct support from European projects and are working to transform itself into a self-sustainable activity. This is being made possible thanks to the establishment of the CLEF Association[16], a non-profit legal entity in late 2013, which, through the support of its members, ensures the resources needed to smoothly run and coordinate CLEF.

Acknowledgments

We would like to thank the mentors who helped in shepherding the preparation of lab proposals by newcomers:
Liana Ermakova, Université de Bretagne Occidentale, France;
Florina Piroi, TU Wien, Austria.

We would like to thank the members of CLEF-LOC (the CLEF Lab Organization Committee) for their thoughtful and elaborate contributions to assessing the proposals during the selection process:
Vito Walter Anelli, Politecnico di Bari, Italy
Luis Alberto Barron Cedeno, University of Bologna, Italy
Alex Brandsen, Leiden University, The Netherlands
Timo Breuer, TH Köln, Germany
Paul D. Clough, University of Sheffielf, UK
Fabio Crestani, Università della Svizzera Italiana, Switzerland
Claudia Hauff, Spotify, The Netherlands
Bogdan Ionescu, Politehnica University of Bucharest, Romania
Alexis Joly, INRIA, France
Jaap Kamps, University of Amsterdam, The Netherlands
Johannes Kiesel, Bauhaus-Universität Weimar, Germany
Birger Larsen, Aalborg University, Denmark
Henning Müller, HES-SO, Switzerland

[16] `https://www.clef-initiative.eu/#association`

Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
Jian-Yun Nie, University de Montreal, Canada
Doug Oard, University of Maryland, USA
Pavel Pecina. Charles University, Czech Republic
Martin Potthast, University of Kassel, Spain
Paolo Rosso, Universitat Politècnica de València, Spain
Eric Sanjuan, Université d'Avignon, France
Ian Soboroff, NIST, USA
Christa Womser-Hacker, University of Hildesheim, Germany

We thank the Friends of SIGIR program 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 proposal authors, the organizers of the selected labs and workshops, the colleagues and friends involved in running them, and the participants who contribute their time to making the labs and workshops a success, the CLEF labs would not be possible.

Thank you all very much!

July, 2024

Guglielmo Faggioli,
Nicola Ferro,
Petra Galuščáková,
Alba García Seco de Herrera

Organization

CLEF 2024, *Conference and Labs of the Evaluation Forum – Experimental IR meets Multilinguality, Multimodality, and Interaction*, was hosted by the University of Grenoble Alpes, France.

General Chairs

Lorraine Goeuriot, Université Grenoble Alpes, France
Philippe Mulhem, Université Grenoble Alpes, France
Georges Quénot, Université Grenoble Alpes, France
Didier Schwab, Université Grenoble Alpes, France

Program Chairs

Giorgio Maria Di Nunzio, University of Padua, Italy
Laure Soulier, Sorbonne Université, France

Lab Chairs

Petra Galuščáková, University of Stavanger, Norway
Alba García Seco de Herrera, University of Essex, UK

Lab Mentorship Chair

Liana Ermakova, Université de Bretagne Occidentale, France
Florina Piroi, TU Wien, Austria

Proceedings Chairs

Guglielmo Faggioli, University of Padua, Italy
Nicola Ferro, University of Padua, Italy

CLEF Steering Committee

Steering Committee Chair

Nicola Ferro, University of Padua, Italy

Deputy Steering Committee Chair for the Conference

Paolo Rosso, Universitat Politècnica de València, Spain

Deputy Steering Committee Chair for the Evaluation Labs

Martin Braschler, Zurich University of Applied Sciences, Switzerland

Members

Avi Arampatzis, Democritus University of Thrace, Greece

Alberto Barrón-Cedeño, University of Bologna, Italy

Khalid Choukri, Evaluations and Language resources Distribution Agency (ELDA), France

Fabio Crestani, Università della Svizzera italiana, Switzerland

Carsten Eickhoff, University of T ubingen, Germany

Norbert Fuhr, University of Duisburg-Essen, Germany

Anastasia Giachanou, Utrecht University, The Netherlands

Lorraine Goeuriot, Université Grenoble Alpes, France

Julio Gonzalo, National Distance Education University (UNED), Spain

Donna Harman, National Institute for Standards and Technology (NIST), USA

Bogdan Ionescu, University "Politehnica" of Bucharest, Romania

Evangelos Kanoulas, University of Amsterdam, The Netherlands

Birger Larsen, University of Aalborg, Denmark

David E. Losada, Universidade de Santiago de Compostela, Spain

Mihai Lupu, Vienna University of Technology, Austria

Maria Maistro, University of Copenhagen, Denmark

Josiane Mothe, IRIT, Université de Toulouse, France

Henning Müller, University of Applied Sciences Western Switzerland (HES-SO),

Switzerland

Jian-Yun Nie, Université de Montréal, Canada

Gabriella Pasi, University of Milano-Bicocca, Italy

Eric SanJuan, University of Avignon, France

Giuseppe Santucci, Sapienza University of Rome, Italy

Laure Soulier, Pierre and Marie Curie University (Paris 6), France

Theodora Tsikrika, Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Greece

Christa Womser-Hacker, University of Hildesheim, Germany

Past Members

Paul Clough, University of Sheffield, United Kingdom

Djoerd Hiemstra, Radboud University, The Netherlands

Jaana Kekäläinen, University of Tampere, Finland

Séamus Lawless, Trinity College Dublin, Ireland

Carol Peters, ISTI, National Council of Research (CNR), Italy
(Steering Committee Chair 2000–2009)

Emanuele Pianta, Centre for the Evaluation of Language and Communication Technologies (CELCT), Italy

Maarten de Rijke, University of Amsterdam UvA, The Netherlands

Jacques Savoy, University of Neuchêtel, Switzerland

Alan Smeaton, Dublin City University, Ireland