Preface

The CLEF 2023 conference is the twenty-fourth 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 2023[1] was organized by the Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece, from 18 to 21 September 2023. CLEF 2023 was the 14th year of the CLEF Conference and the 24th year of the CLEF initiative as a forum for IR Evaluation. 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 15 lab proposals were received and evaluated in peer review based on their innovation potential and the quality of the resources created. The 13 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 2023 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 13 labs running as part of CLEF 2023 comprised mainly labs that continued from previous editions at CLEF (BioASQ, CheckThat!, eRisk, iDPP, ImageCLEF, JOKER, LifeCLEF, PAN, SimpleText, and Touché) and new pi-

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

lot/workshop activities (DocILE, EXIST, and LongEval). In the following we give a few details for each of the labs organized at CLEF 2023 (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 - MedProcNER: Medical Procedure Text Mining and Indexing Shared Task*: focuses on the recognition and indexing of medical procedures in clinical documents in Spanish posing subtasks on (1) indexing medical documents with controlled terminologies, (2) automatic detection indexing textual evidence, i.e. medical procedure entity mentions in text, and (3) normalization of these medical procedure mentions to terminologies.

CheckThat!: Check-Worthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and their Sources[3] aims at producing technology to support the fight against misinformation and disinformation in social media, in political debates and in the news with a focus on check-worthiness, subjectivity, bias, factuality, and authority of the claim. It offered the following tasks. *Task 1 - Check-worthiness in textual and multimodal content*: determine whether an item, be it a text alone or a text plus an image deserves the attention of a journalist to be fact-checked. *Task 2 - Subjectivity in News Articles*: assess whether a text snippet within a news article is subjective or objective. *Task 3 - Political Bias of News Articles and News Media*: identify the political leaning of an article or media source: left, centre or right. *Task 4 - Factuality of Reporting of News Media*: determine the level of factuality of both a document and a medium. *Task 5 - Authority Finding in Twitter*: identify authorities that should be trusted to verify a contended claim expressed in an Arabic tweet.

DocILE: Document Information Localization and Extraction[4] runs the largest benchmark for the tasks of Key Information Localization and Extraction (KILE) and Line Item Recognition (LIR) from business documents like invoices. It offered the following tasks. *Task 1 - Key Information Localization*

[2] http://www.bioasq.org/workshop2023

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

[4] https://docile.rossum.ai/

and Extraction (KILE): localize fields of each pre-defined category and read out their values. *Task 2 - Line Item Recognition (LIR)*: find all line items, e.g., a billed item in a table, and localize their corresponding fields in the document as in Task 1.

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. Early detection technologies can be employed in different areas, particularly those related to health and safety. For instance, early alerts could be sent when a predator starts interacting with a child for sexual purposes, or when a potential offender starts publishing antisocial threats on a blog, forum or social network. Our main goal is to pioneer a new interdisciplinary research area that would be potentially applicable to a wide variety of situations and to many different personal profiles. Examples include potential paedophiles, stalkers, individuals that could fall into the hands of criminal organisations, people with suicidal inclinations, or people susceptible to depression. It offered the following tasks. *Task 1 - Search for symptoms of depression*: the challenge 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. A sentence will be deemed relevant to a BDI symptom when it conveys information about the user's state concerning the symptom. That is, it may be relevant even when it indicates that the user is OK with the symptom. *Task 2 - Early Detection of Signs of Pathological Gambling*: the challenge consists of sequentially processing pieces of evidence and detect early traces of pathological gambling (also known as compulsive gambling or disordered gambling), as soon as possible. The task is mainly concerned about evaluating Text Mining solutions and, thus, it concentrates on texts written in Social Media *Task 3 - Measuring the severity of the signs of Eating Disorders*: the task 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 (based on the evidence found in the history of postings).

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*: is a binary classification tasks. The systems have to decide whether or not a given tweet contains or describes sexist expressions or behaviors (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behavior). *Task 2 - Source Intention*: aims to categorize the sexist messages

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

[6] http://nlp.uned.es/exist2023/

according to the intention of the author in one of the following categories: (i) direct sexist message, (ii) reported sexist message and (iii) judgemental message. *Task 3 - Sexism Categorization*: is a multiclass task that aims to categorize the sexist messages according to the type or types of sexism they contain (according to the categorization proposed by experts and that takes into account the different facets of women that are undermined): (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence.

iDPP: Intelligent Disease Progression Prediction[7] aims to design and develop an evaluation infrastructure for AI algorithms able to: (1) better describe mechanism of the Amyotrophic Lateral Sclerosis (ALS) disease; (2) stratify patients according to their phenotype assessed all over the disease evolution; and (3) predict ALS progression in a probabilistic, time dependent fashion. It offered the following tasks. *Task 1 – Predicting Risk of Disease Worsening (Multiple Sclerosis)*: focuses on ranking subjects based on the risk of worsening, setting the problem as a survival analysis task. More specifically the risk of worsening predicted by the algorithm should reflect how early a patient experience the event "worsening". Worsening is defined based on the Expanded Disability Status Scale (EDSS), accordingly to clinical standards. *Task 2 – Predicting Probability of Worsening (Multiple Sclerosis)*: refines Task 1 asking participants to explicitly assign a probability of worsening at different time windows (e.g. between years 4 and 6, 6 and 8, 8 and 10 etc.). *Task 3 – Impact of Exposition to Pollutants (Amyotrophic Lateral Sclerosis)*: evaluates proposals of different approaches to assess if exposure to different pollutants is a useful variable to predict time to Percutaneous Endoscopic Gastrostomy (PEG), Non-Invasive Ventilation (NIV) and death in ALS patients.

ImageCLEF: Multimedia Retrieval[8] is set to promote 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. It offered the following tasks. *Task 1 - ImageCLEFmedical*: continues the tradition of bringing together several initiatives for medical applications fostering cross-exchanges, namely: medical concept detection and caption prediction, synthetic medical images generated with GANs, Visual Question Answering and generation, and doctor-patient conversation summarization. *Task 2 - ImageCLEFaware*: the images available on social networks can be exploited in ways users are unaware of when initially shared, including situations that have serious consequences for the users' real lives. The task addresses the development of algorithms which raise the users' awareness about real-life impact of online image sharing. *Task 3 - ImageCLEFfusion*: despite the current advances in knowledge discovery, single learners do not produce satisfactory performances when dealing with complex data, such as class imbalance,

[7] https://brainteaser.health/open-evaluation-challenges/idpp-2023/

[8] https://www.imageclef.org/2023

high-dimensionality, concept drift, noise, multimodality, subjective annotations, etc. This task aims to fill this gap by exploiting novel and innovative late fusion techniques for producing a powerful learner based on the expertise of a pool of classifiers. *Task 4 - ImageCLEFrecommendation*: focuses on content-recommendation for cultural heritage content in 15 broad themes that have been curated by experts in the Europeana Platform. Despite current advances, there is limited understanding how well these perform and how relevant they are for the final end-users.

JOKER: Automatic Wordplay Analysis[9] aims to create reusable test collections for benchmarking and to explore new methods and evaluation metrics for the automatic processing of wordplay. It offered the following tasks. *Task 1 - Pun detection*: detection of puns in English, French, and Spanish. *Task 2 - Pun interpretation*: interpretation of puns in English, French, and Spanish. *Task 3 - Pun translation*: translation of puns from English to French and Spanish.

LifeCLEF: Multimedia Retrieval in Nature[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 - Cross-Discourse Type Authorship Verification*: focuses on (cross-discourse type)

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

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

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

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

authorship verification where both written (e.g., essays, emails) and oral language (e.g., interviews, speech transcriptions) are represented in the set of discourse types. *Task 2 - Profiling Cryptocurrency Influencers with Few-shot Learning*: aims to profile cryptocurrency influencers in social media (Twitter) from a low-resource perspective. *Task 3 - Multi-Author Writing Style Analysis*: addresses multi-authored documents whose authorship cannot be easily determined by exploiting topic changes alone. *Task 4 - Trigger Detection*: addresses the task of assigning a single trigger warning label (violence) to narratives in a corpus of fanfiction.

SimpleText: Automatic Simplification of Scientific Texts[13] aims to create a simplified summary of multiple scientific documents based on a popular science query which provides a user with an instant accessible overview on this specific topic. It offered the following tasks. *Task 1 - What is in, or out?*: selecting passages to include in a simplified summary. *Task 2 - What is unclear?*: difficult concept identification and explanation. *ask 3 - Rewrite this!*: rewriting scientific text.

Touché: Argument and Causal Retrieval[14] aims to foster and support the development of technologies for argument and causal retrieval and analysis that includes argument quality estimation, stance detection, image retrieval, and causal evidence retrieval. It offered the following tasks. *Task 1 - Argument Retrieval for Controversial Questions*: given a controversial topic and a collection of web documents, the task is to retrieve and rank documents by relevance to the topic, by argument quality, and to detect the document stance. *Task 2 - Evidence Retrieval for Causal Questions*: given a causality-related topic and a collection of web documents, the task is to retrieve and rank documents by relevance to the topic and detect the document "causal" stance (i.e., whether a causal relationship from the topic0s title holds). *Task 3 - Image Retrieval for Arguments*: given a controversial topic, the task is to retrieve images (from web pages) for each stance (pro/con) that show support for that stance. *Task 4 - Intra-Multilingual Multi-Target Stance Classification*: given a proposal on a socially important issue, its title, and topic in different languages, the task is to classify whether a comment is in favor, against, or neutral towards the proposal.

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 organisation 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[15], a non-profit legal entity

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

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

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

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:
Bogdan Ionescu, University "Politehnica" of Bucharest, Romania;
Henning Müller, University of Applied Sciences Western Switzerland (HES-SO), Switzerland.

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:
Alexander Bondarenko, Martin-Luther-Universit at Halle-Wittenberg, Germany;
Martin Braschler, ZHAW Zurich University of Applied Sciences, Switzerland;
Donna Harman, National Institute of Standards and Technology (NIST), USA;
Aldo Lipani, University College London, UK;
Maria Maistro, University of Copenhagen, Denmark;
Paolo Rosso, Universitat Politècnica de València, Spain;
Efstathios Stamatatos, University of the Aegean, Greece.

We thank the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH) in Thessaloniki, Greece (with special mention to the Multimodal Data Fusion & Analytics Group of the Multimedia Knowledge and Social Media Analytics Laboratory) for their invaluable support. We also 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, 2023

Mohammad Aliannejadi,
Guglielmo Faggioli,
Nicola Ferro,
Michalis Vlachos

Organization

CLEF 2023, *Conference and Labs of the Evaluation Forum – Experimental IR meets Multilinguality, Multimodality, and Interaction*, was hosted by the Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Greece.

General Chairs

Avi Arampatzis, Democritus University of Thrace, Greece

Evangelos Kanoulas, University of Amsterdam, The Netherlands

Theodora Tsikrika, Information Technologies Institute, CERTH, Greece

Stefanos Vrochidis, Information Technologies Institute, CERTH, Greece

Program Chairs

Anastasia Giachanou, Utrecht University, The Netherlands

Dan Li, Elsevier, The Netherlands

Lab Chairs

Mohammad Aliannejadi, University of Amsterdam, The Netherlands

Michalis Vlachos, University of Lausanne, Switzerland

Lab Mentorship Chair

Jian-Yun Nie, University of Montreal, Canada

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

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

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