Preface

The CLEF 2022 conference is the twenty-third 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 2022[1] was organized by the University of Bologna, Italy, from 5 to 8 September 2022.

Despite the continued outbreak of the Covid-19 pandemic, the improvement of the overall situation allowed for organizing CLEF 2022 as an in-presence event, after two editions – CLEF 2020 and 2021 – forced to be virtual only. 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 and run online.

15 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, besides the well-established criteria from the editions of previous years of CLEF such as 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 data sets and real world problems in multimodal and multilingual information access. These data sets 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 issues with peers at the workshops.

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 2022 demonstrate the maturity of the CLEF evaluation environment by creating new tasks, new and

[1] `https://clef2022.clef-initiative.eu/`

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 2022 comprised mainly labs that continued from previous editions at CLEF (ARQMath, BioASQ, CheckThat!, CheMU, eRisk, ImageCLEF, LifeCLEF, PAN, SimpleText, and Touché) and a new pilot/workshop activity (HIPE, iDPP, JOKER, and LeQUA). Below is a short summary of them.

ARQMath: Answer Retrieval for Mathematical Questions[2] aims to advance math-aware search and the semantic analysis of mathematical notation and texts. It offered the following tasks. Task 1: Answer Retrieval, given a math question post, return relevant answer posts. Task 2: Formula Retrieval; given a formula in a math question post, return relevant formulas from both question and answer posts. Task 3: Open Domain Question Answering, given a math question post, return an automatically generated answer that is comprised of excerpts from arbitrary sources, and/or machine generated.

BioASQ: Large-scale biomedical semantic indexing and question answering[3] 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: Large-Scale Online Biomedical Semantic Indexing, it classifies new PubMed documents, before PubMed curators annotate (in effect, classify) them manually into classes from the MeSH hierarchy. Task 2: Biomedical Semantic Question Answering, it uses 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 3 - DisTEMIST: Disease Text Mining and Indexing Shared Task, it focuses on the recognition and indexing of diseases in medical documents in Spanish, by posing subtasks on (1) indexing medical documents with controlled terminologies; (2) automatic detection indexing textual evidence (i.e. disease entity mentions in text); and (3) normalization of these disease mentions to terminologies. Task 4 - Task Synergy: Question Answering for developing problems, biomedical experts pose unanswered questions for the developing problem of 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 COVID-19.

CheckThat!: Lab on Fighting the COVID-19 Infodemic and Fake News Detection[4] aims at fighting misinformation and disinformation in social media, in political debates and in the news, with focus on three tasks (in seven languages: Arabic, Bulgarian, Dutch, English, German, Spanish,

[2] `https://www.cs.rit.edu/~dprl/ARQMath`

[3] `http://www.bioasq.org/workshop2022`

[4] `https://sites.google.com/view/clef2022-checkthat`

and Turkish). It offered the following tasks. Task 1: Fighting the COVID-19 Infodemic, it focuses on disinformation related to the ongoing COVID-19 infodemic and asks to identify which posts in a Twitter stream are worth fact-checking, contain a verifiable factual claim, are harmful to the society, and why. This task is offered in Arabic, Bulgarian, Dutch, English, Spanish, and Turkish. Task 2: Detecting Previously Fact-Checked Claims, given a check-worthy claim, and a set of previously-checked claims, determine whether the claim has been previously fact-checked with respect to a collection of fact-checked claims. The text can be a tweet or a sentence from a political debate. The task is offered in Arabic and English. Task 3: Fake news detection, given the text and the title of a news article, determine whether the main claim made in the article is true, partially true, false, or other (e.g., articles in dispute and unproven articles). This task is offered in English and German.

ChEMU: Cheminformatics Elsevier Melbourne University[5] focuses on information extraction in chemical patents, including five tasks ranging from document- to expression-level. It offered the following tasks. Task 1a: Named entity recognition, it aims to identify chemical compounds, their specific types, temperatures, reaction times, yields, and the label of the reaction. Task 1b: Event extraction, a chemical reaction leading to an end product often consists of a sequence of individual event steps. The task is to identify those steps which involve chemical entities recognized from Task 1a. Task 1c: Anaphora resolution, it requires the resolution of anaphoric dependencies between expressions in chemical patents. The participants are required to find five types of anaphoric relationships in chemical patents: coreference, reaction-associated, work-up, contained, and transform. Task 2a: Chemical reaction reference resolution, given a reaction description, this task requires identifying references to other reactions that the reaction relates to, and to the general conditions that it depends on. Task 2b: Table semantic classification, it is about classifying tables in chemical patents into 8 categories based on their contents .

eRisk: Early Risk Prediction on the Internet[6] explores the evaluation methodology, effectiveness metrics, and practical applications (particularly those related to health and safety) of early risk detection on the Internet. 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: 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

[5] `http://chemu2022.eng.unimelb.edu.au/`

[6] `https://erisk.irlab.org/`

written in Social Media. Task 2: Early Detection of Depression, the challenge consists of sequentially processing pieces of evidence and detect early traces of depression 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).

HIPE: Named Entity Recognition and Linking in Multilingual Historical Documents[7] focuses on named entity recognition and linking in historical documents, with the objective of assessing and advancing the development of robust, adaptable, and transferable named entity processing systems. Compared to the first HIPE edition in 2020, HIPE 2022 will confront systems with the challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation schemas. It offered the following tasks. Task 1: Named Entity Recognition and Classification (NERC), with two subtasks: NERC-coarse on high-level entity types, for all languages and NERC-fine on finer-grained entity types, for English, French, and German only. Task 2: Named Entity Linking (EL), Or the linking of named entity mentions to a unique referent in a knowledge base (Wikidata) or to a NIL node if the mention does not have a referent in the KB.

iDPP: Intelligent Disease Progression Prediction[8] 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: Ranking Risk of Impairment, it focuses on ranking of patients based on the risk of impairment in specific domains. We will use the ALSFRS-R scale to monitor speech, swallowing, handwriting, dressing/hygiene, walking and respiratory ability in time and asks participants to rank patients based on time to event risk of experiencing impairment in each specific domain. Task 2: Predicting Time of Impairment, it refines Task 1 asking participants to predict when specific impairments will occur (i.e. in the correct time-window) by assessing model calibration in terms of the ability of the proposed algorithms to estimate a probability of an event close to the true probability within a specified time-window. Task 3: Explainability of AI algorithms, it calls for position papers to start a discussion on AI explainability including proposals on how the single patient data can be visualized in a multivariate fashion contextualizing its dynamic nature and the model predictions together with information on the predictive variables that most influence the prediction.

[7] `https://hipe-eval.github.io/HIPE-2022/`

[8] `https://brainteaser.health/open-evaluation-challenges/idpp-2022/`

ImageCLEF: Multimedia Retrieval[9] 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 of images in various usage scenarios and domains. It offered the following tasks. Task 1: ImageCLEFmedical, it focuses on interpreting and summarizing the insights gained from radiology images, i.e. develop systems that are able to predict the UMLS concepts from visual image content, and implementing models to predict captions for given radiology images. The tuberculosis task fosters systems that are expected to detect cavern regions localization rather than simply provide a label for the CT images. Task 2: ImageCLEFcoral, it fosters tools for creating 3-dimensional models of underwater coral environments. It requires participants to label coral underwater images with types of benthic substrate together with their bounding box, and to segment and parse each coral image into different image regions associated with benthic substrate types. Task 3: ImageCLEFaware, the online disclosure of personal data often has effects which go beyond the initial context in which data were shared. Participants are required to provide automatic rankings of photographic user profiles in a series of real-life situations such as searching for a bank loan, an accommodation, a waiter job or a job in IT. The ranking will be based on an automatic analysis of profile images and the aggregation of individual results. Task 4: ImageCLEFfusion, dystem fusion allows to exploit the complementary nature of individual systems to boost performance. Participants will be tasked with creating novel ensembling methods that are able to significantly increase the performance of precomputed inducers in various use-case scenarios, such as visual interestingness and video memorability prediction.

JokeR: Automatic Wordplay and Humour Translation Workshop[10] aims to bring together translators and computer scientists to work on an evaluation framework for creative language, including data and metric development, and to foster work on automatic methods for wordplay translation. It offered the following tasks. Pilot task 1: Classify and interpret wordplay, classify single words containing wordplay according to a given typology, and provide lexical-semantic interpretations. Pilot task 2: Translate single term wordplay, translate single words containing wordplay. Pilot task 3: Translate phrase wordplay, translate entire phrases that subsume or contain wordplay. Task 4: Unshared Task, we welcome submissions that use our data in other ways.

LeQua: Learning to Quantify[11] aims to allow the comparative evaluation of methods for “learning to quantify” in textual datasets; i.e. methods for training predictors of the relative frequencies of the classes of interest in sets of unlabelled textual documents. These predictors (called “quantifiers”) will be required to issue predictions for several such sets, some of them

[9] https://www.imageclef.org/2022

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

[11] https://lequa2022.github.io/

characterized by class frequencies radically different from the ones of the training set. It offered the following tasks. Task 1: participants are provided with documents already converted into vector form; the task is thus suitable for participants who do not wish to engage in generating representations for the textual documents, but want instead to concentrate on optimizing the methods for learning to quantify. Task 2: participants are provided with the raw text of the documents; the task is thus suitable for participants who also wish to engage in generating suitable representations for the textual documents, or to train end-to-end systems.

LifeCLEF: Biodiversity identification and prediction[12] aims to stimulate research in data science and machine learning for biodiversity monitoring. It offered the following tasks. Task 1: BirdCLEF, bird species recognition in audio soundscapes. Task 2: PlantCLEF, image-based plant identification at global scale (300K classes). Task 3: GeoLifeCLEF, location-based prediction of species based on environmental and occurrence data. Task 4: SnakeCLEF, snake species identification in medically important scenarios. Task 5: FungiCLEF, fungi Recognition from image and metadata.

PAN: Digital Text Forensics and Stylometry[13] focuses on digital text forensics and stylometry, studying how to quantify writing style and improve authorship technology. It offered the following tasks. Task 1: Authorship Verification, given two texts, determine if they are written by the same author. Task 2: IROSTEREO, profiling Irony and Stereotype Spreaders on Twitter, given a Twitter feed, determine whether its author spreads Irony and Stereotypes. Task 3: Style Change Detection, given a document, determine the number of authors and at which positions the author changes. Task 4: Trigger Warning Prediction, given a document, determine whether its content warrants a warning of potential negative emotional responses in readers..

SimpleText: Automatic Simplification of Scientific Texts[14] addresses the challenges of text simplification approaches in the context of promoting scientific information access, by providing appropriate data and benchmarks, and creating a community of NLP and IR researchers working together to resolve one of the greatest challenges of today. It offered the following tasks. Task 1: What is in (or out)? Select passages to include in a simplified summary, given a query. Task 2: What is unclear? Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications, ...). Task 3: Rewrite this! Given a query, simplify passages from scientific abstracts. Task 4: Unshared task, we welcome any submission that uses our data.

Touché: Argument Retrieval[15] focuses on decision making processes, be it at the societal or at the personal level, often come to a point where one

[12] https://www.imageclef.org/LifeCLEF2022

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

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

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

side challenges the other with a why-question, which is a prompt to justify some stance based on arguments. Since technologies for argument mining are maturing at a rapid pace, also ad-hoc argument retrieval becomes a feasible task in reach. It offered the following tasks. Task 1: Argument Retrieval for Controversial Questions, given a controversial topic and a collection of argumentative documents, the task is to retrieve and rank sentences (the main claim and its most important premise in the document) that convey key points pertinent to the controversial topic. Task 2: Argument Retrieval for Comparative Questions, given a comparative topic and a collection of documents, the task is to retrieve relevant argumentative passages for either compared object or for both and to detect their respective stances with respect to the object they talk about. 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.

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[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 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.

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.

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

Thank you all very much!

July, 2022

Guglielmo Faggioli,
Nicola Ferro,
Allan Hanbury,
Martin Potthast

Organization

CLEF 2022, *Conference and Labs of the Evaluation Forum – Experimental IR meets Multilinguality, Multimodality, and Interaction*, was hosted by University of Bologna, Italy.

General Chairs

Alberto Barrón-Cedeño, University of Bologna, Italy

Giovanni Da San Martino, University of Padua, Italy

Mirko Degli Esposti, University of Bologna, Italy

Fabrizio Sebastiani, National Council of Research, ISTI CNR, Italy

Program Chairs

Craig Macdonald, University of Glasgow, UK

Gabriella Pasi, University of Milan Bicocca, Italy

Lab Chairs

Allan Hanbury, Vienna University of Technology, Austria

Martin Potthast, Leipzig University, Germany

Lab Mentorship Chair

Paolo Rosso, Universitat Politécnica de Valencia, Spain

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

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

Fabio Crestani, Università della Svizzera italiana, Switzerland

Carsten Eickhoff, Brown University, USA

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

Eric SanJuan, University of Avignon, France

Giuseppe Santucci, Sapienza University of Rome, Italy

Jacques Savoy, University of Neuchêtel, Switzerland

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

Alan Smeaton, Dublin City University, Ireland