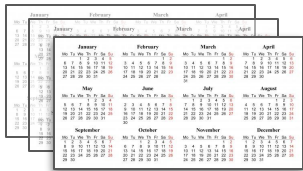


# LongEval: Longitudinal Evaluation of Model Performance



<https://clef-longeval.github.io>



1<sup>st</sup> edition

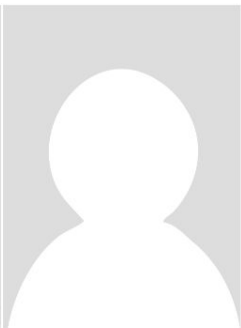


Kodicare

# Organizers



**Rabab Alkhalifa**  
Queen Mary  
University of London



**Iman Bilal**  
University of  
Warwick



**Hsuvas Borkakoty**  
Cardiff University



**Jose Camacho-  
Collados**  
Cardiff University



**Romain Deveaud**  
Qwant



**Luis Espinosa-Anke**  
Cardiff University



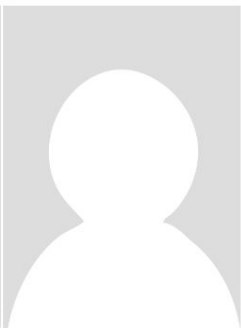
**Petra Galuščáková**  
Université Grenoble  
Alpes



**Lorraine Goeriot**  
Université Grenoble  
Alpes



**Elena Kochkina**  
Queen Mary  
University of London



**Maria Liakata**  
Queen Mary  
University of London



**Daniel Loureiro**  
Cardiff University



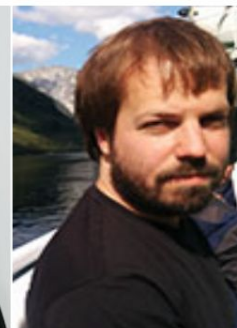
**Harish Tayyar  
Madabushi**  
University of Bath



**Philippe Mulhem**  
CNRS



**Florina Piroi**  
RSA



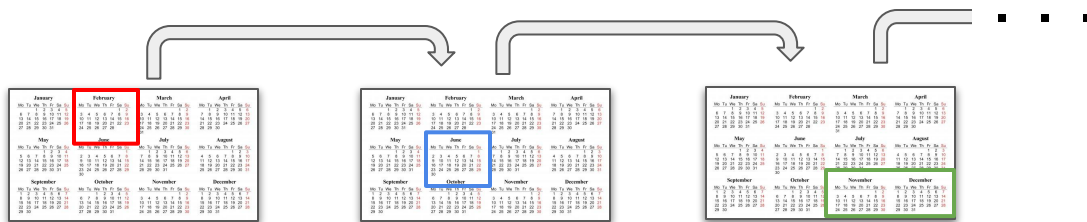
**Arkaitz Zubiaga**  
Queen Mary  
University of London



**Christophe Servan**  
Qwant


# Motivations for the CLEF 2023 LongEval Lab

Temporal persistence evaluation: How do systems face temporal evolution?



Creating benchmarks for a continuous evaluation for two major data collections:  
sentiment analysis on social network and Web search:

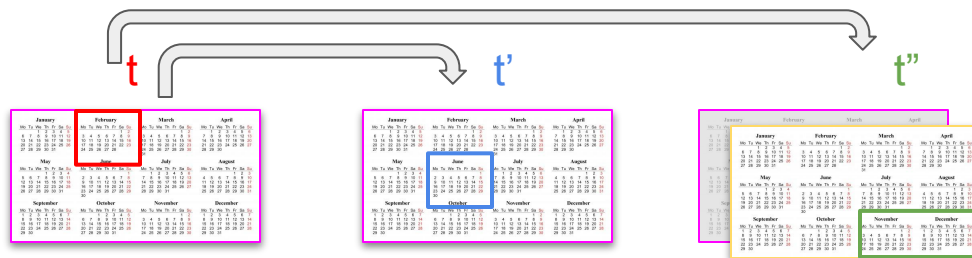
 Classification: Sentiment (pos / neg) of tweets

 Information Retrieval: Quality of retrieval of Web documents and queries

# Research question for the CLEF 2023 LongEval Lab

To quantify the drop of quality of Classification and IR systems

Global Framework



1. Training set from data acquired at time  $t$
2. Test set *within-time* from data acquired at time  $t$  (reference)
3. Test set with data acquired in  $(t, t']$  (short term) (sub-task A)
4. Test set with data acquired in  $(t', t'']$  (sub-task B)
5. Evaluate the drop of quality between  $t$   $t'$  and  $t$   $t''$

# Impact for the CLEF 2023 LongEval Lab

No existing shared tasks dedicated specifically to this important question.

Build a first overview of the persistence of the state of the art Classification and Information Retrieval Systems evolution over the time.

# Task.1 LongEval-Retrieval

## Temporal Information Retrieval

### Dataset



- All Web documents and queries from **Qwant** search engine
  - overall: ~ 500k docs (fr); ~ 10k queries (fr); assessments from a click model
  - queries acquired using their popularity
  - documents from SERPs + background
- 4 subsets:
  - one training set at time  $t$ ,
  - three test sets:
    - one within-time at  $t$
    - one short-term in  $(t, t']$  (sub-task A)
    - one long-term in  $(t', t'']$  (sub-task B).
    - $t' = t + 3$  months,  $t'' = t + 6$  months.

### Evaluation measures

- Absolute evaluation: nDCG
- Relative nDCG Drop with a respect to within-time

# Task.2 LongEval-Classification

## Longitudinal Text Classification

### Dataset

- TM-Senti: Temporal multilingual sentiment dataset
    - a general **large-scale tweets** sentiment dataset in the **English** language
    - **spanning a 9-year period ranging from 2013 to 2021.**
    - Tweets are **binary labelled** for sentiment as either “positive” or “negative”
  - 4 subsets:
    - one training set at time  $t$
    - three evaluation sets
      - within-time test set at  $t$
      - short-term set in  $(t, t']$  (sub-task A)
      - long-term set in  $(t', t'']$  (sub-task B).
- where  $t' = t + 1 \text{ year}$  ,  $t'' = t + n \text{ years where } n > 1$*

### Evaluation measures

- Macro-averaged F1-score
- Relative Performance Drop

# Timeline

## Same for the two tasks:

- **December 2022:** Training data release
- **End of April 2023:** Participants' submissions
- **May 2023:** Participants' papers submissions
- **June 2023:** Evaluation results release
- **July 2023:** Camera ready paper submissions
- **September 2023:**



LongEval Lab Web site: <https://clef-longeval.github.io>