LongEval: Longitudinal Evaluation of Model Performance at CLEF 2023

Rabab Alkhalifa[1,2][0000−0002−2875−5400], Iman Bilal[3][0000−0002−7972−9085], Hsuvas Borkakoty[4][0000−0003−3262−0127], Jose Camacho-Collados[4][0000−0003−1618−7239], Romain Deveaud[6][0000−0003−2676−7405], Alaa El-Ebshihy[9][0000−0001−6644−2360], Luis Espinosa-Anke[4,12][0000−0001−6830−9176], Gabriela Gonzalez-Saez[7][0000−0003−0878−5263], Petra Galuščáková[7][0000−0001−6328−7131], Lorraine Goeuriot[7][0000−0001−7491−1980], Elena Kochkina[1,5][0000−0003−0691−3647], Maria Liakata[1,3,5][0000−0001−5765−0416], Daniel Loureiro[4][0000−0001−5134−360X], Harish Tayyar Madabushi[0000−0001−5260−3653][8], Philippe Mulhem[7][0000−0002−3245−6462], Florina Piroi[9][0000−0001−7584−6439], Martin Popel[10][0000−0002−3628−8419], Christophe Servan[6,11][0000−0003−2306−7075], and Arkaitz Zubiaga[1][0000−0003−4583−3623]

[1] Queen Mary University of London, UK
[2] Imam Abdulrahman Bin Faisal University, SA
[3] University of Warwick, UK
[4] Cardiff University, UK
[5] Alan Turing Institute, UK
[6] Qwant, France
[7] Univ. Grenoble Alpes, CNRS, Grenoble INP**, LIG, Grenoble, France
[8] University of Bath, UK
[9] Research Studios Austria, Data Science Studio, Vienna, AT
[10] Charles University, Prague, Czech Republic
[11] Paris-Saclay University, CNRS, LISN, France
[12] AMPLYFI, UK

Abstract. In this paper, we describe the plans for the first LongEval CLEF 2023 shared task dedicated to evaluating the temporal persistence of Information Retrieval (IR) systems and Text Classifiers. The task is motivated by recent research showing that the performance of these models drops as the test data becomes more distant, with respect to time, from the training data. LongEval differs from traditional shared IR and classification tasks by giving special consideration to evaluating models aiming to mitigate performance drop over time. We envisage that this task will draw attention from the IR community and NLP researchers to the problem of temporal persistence of models, what enables or prevents it, potential solutions and their limitations.

Keywords: Evaluation · Temporal Persistence · Temporal Generalisability · Information Retrieval · Text Classification.

** Institute of Engineering Univ. Grenoble Alpes.

1 Introduction

Recent research demonstrates that the performance of Text Retrieval and Classification systems drops over time as patterns observed in data change, due to linguistic and societal changes [2]. In classification systems, this drop is more pronounced when the testing data is further away in time from training data [7, 5, 1], a problem we refer to as the problem of *classifier temporal persistence*. Similarly, in Information Retrieval, it has been shown that a deep neural network-based IR is dependent on the consistency between the train and test data [8]. Given that in most scenarios one has limited resources to continuously label new data to train models on, the aim of this shared task is to encourage the development of models that mitigate performance drop over time as the training data gets older. We do this by providing participants with training data distant in time from testing and un-annotated data from the testing time period. The challenges that come with such an evaluation setting are numerous, ranging from the definition and collection of the data on which the systems may be compared to the measures considered. As such, this lab focuses on two different tasks, both with a temporal axis in their design: (a) Task 1, Information Retrieval for the case in which Web documents evolve over the time, queries are not known a priori, relevance judgements are non-binary and submissions are required to provide ranked lists as results, and (b) Task 2, text classification in which the target classes are predefined while language usage associated with each class evolves rapidly over time, as in social media.

We encourage the development of novel approaches that can automatically adapt to possible temporal dynamics in textual data so as to progress towards time-insensitive computational methods. As such, the expected outcomes from this lab are threefold:

- to draw a deeper understanding of how time impacts IR and classification systems;
- to assess the effectiveness of different retrieval and classification approaches in achieving temporal persistence;
- to propose computational methods to leverage ageing labelled datasets, while minimising performance drop over time.

Given the prevalence of text classification in IR and NLP research across CLEF labs, as well as our objective to rank top models that provide high temporal persistence for NLP tasks, we propose our evolving sets which analyse how natural language use changes overtime sets,(either over the short term or long term) compared to testing data from the same time frame (within-time). LongEval is built on a common framework that adds the temporal gap that defines the distance between training and testing as a time-sensitivity measure. As shown in Figure 1, we compare the retrieval or classification temporal generalisability of a given IR or classification system when operating on data acquired at time t from same time as training, its persistence when operating on data acquired at time t' (occurring a short period after time t), and its persistence when operating on data acquired at time t" (occurring a long period after time t). The system's ability to cope with dynamic data is thus evaluated using longi-

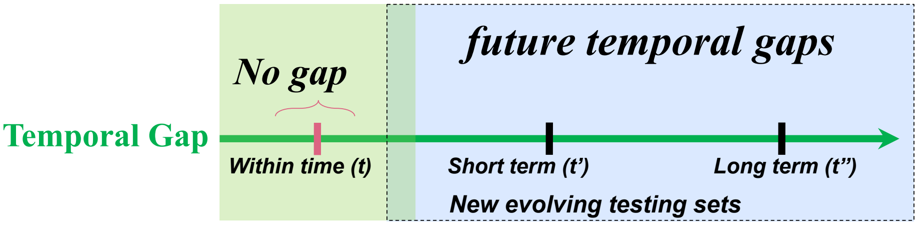

Fig. 1. Global framework for the LongEval Tasks.

tudinal datasets split at different temporal granularities, i.e. within-time, short and long time distances from the training data.

The remainder of the paper is structured as follows: LongEval-Retrieval is covered in Section 2.1, while LongEval-Classification is covered in Section 2.2. Both sections propose tasks and provide additional information about the data and the baseline to be used. Section 3 contains additional information and guidelines for participants.

2 Tasks

2.1 Task 1: LongEval-Retrieval

The goal of the Task 1 is to support the development of Information Retrieval systems that cope with temporal evolution. The retrieval systems evaluated in this task are expected to be persistent in their retrieval efficiency over time, as Web documents and Web queries evolve. To evaluate such features of systems, we rely on collections of documents and queries, corresponding to real data acquired from an actual Web search engine.
The ***LongEval-Retrieval*** organizes two sub-tasks.

Sub-task 1.A: Short-term persistence. In this task, participants will be asked to examine the retrieval effectiveness when the test documents are acquired shortly (typically within a range of few months distance) after the documents available in the train collection.

Sub-task 1.B: Long-term persistence. Here, participants will be asked to examine retrieval effectiveness on the documents published after a long period, at least 4 months after the documents in the train collection were published.

As mentioned above, **any participation in the sub-tasks 1.A or 1.B necessitate a "within time" run submission**.

Data The data for this task is a sequence of Web document collections and queries, each containing a few million documents (e.g. 2.8 M for the training data) and hundreds of queries (e.g. almost 700 for the training data), provided by Qwant[1]. Each document set will have a release time stamp, with the first set (in

[1] Qwant search engine: https://www.qwant.com/

chronological order) being the training data. Discrete relevance assessments are computed using a simplified Dynamic Bayesian Network (sDBN) Click Model [3, 4] acquired from real users of the Qwant search engine. As the initial corpus contains only French documents, an automatic translation into English will be released. The organisers will provide the following data:

1. A training set (queries, documents, qrels) created over a time interval t. Such data should be used by the participants to train their models.
2. One "within time" test set (queries, documents) acquired at the same time frame as the training set. This test set will be used to assess the initial performance of the trained models, and will not be used to directly assess submissions;
3. Two test sets: one test set (queries, documents) acquired during a time interval t' occurring shortly after t (with no intersection between t and t') dedicated to evaluate **short-term persistence** sub-task 1.A, and one test set acquired long after t during a time interval t'' (with no intersection between t and t''), for **long-term persistence** evaluation, sub-task 1.B.

Evaluation The submitted systems will be evaluated in two ways:

1. **nDCG** scores calculated on test set provided for the sub-tasks. Such a classical evaluation measure is consistent with Web search, for which the discount emphasises the ordering of the top results.
2. **Relative nDCG Drop (RnD)** measured by computing the difference between nDCG on "within time" test data vs short- or long-term testing sets. This measure relies on the "within time" test data, and supports the evaluation of the impact of the data changes on the system's results.
3. **Relative nDCG Drop (RnD)** measured by computing the difference between nDCG on "within time" test data vs short- or long-term testing sets. This measure relies on the "within time" test data, and supports the evaluation of the impact of the data changes on the system's results.

These measures will be used to assess the extent to which systems provide good results, but also the extent to which they are robust against the changes within the data (queries/documents) along time. Using these evaluation measures, a system that has good results using nDCG, and also good results according to the RnD measure is considered to be able to cope with the evolution over time of the Information Retrieval collection.

2.2 Task 2: LongEval-Classification

The first LongEval-Classification challenge focuses on systems that perform social media sentiment analysis, which is expressed as a binary classification task in English. By addressing this critical and widely studied task, we hope to attract attention and participation from the broader AI/NLP communities in order to better understand this emerging field and develop novel temporally persistent approaches.

The ***LongEval-Classification*** organizes two sub-tasks.

Sub-task 2.A: Short-term persistence. In this task participants will be asked to develop models which demonstrate performance persistence over short periods of time (test set within 1 year from the training data).

Sub-task 2.B: Long-term persistence. In this task participants will be asked to develop models which demonstrate performance persistence over longer period of time (test set over 1 year apart from the training data).

Data The training data to be provided to the task participants will consist of the TM-Senti dataset[2] extended with a development set and three human-annotated novel test sets for submission evaluation. TM-Senti is a general large-scale Twitter sentiment dataset in English language, spanning a 9-year period from 2013 to 2021. Tweets are labelled for sentiment as either "positive" or "negative". The annotation is performed using distant supervision based on a manually curated list of emojis and emoticons [9] and, thus, can be easily extended to cover more recent years. We plan to release data in two phases:

1. In the **development phase**, participants will be given **(1) a distantly annotated training set** (tweet, label) created over a time interval t. Such data is dedicated model training, as well as **(2) human-annotated "within time" development set** (tweet, label) from the same time period t. This development set is intended to allow participants to develop their systems before the following phase, and will not be used to rank their submissions. For participant interested in data-centric approaches, we provide **(3) an un-labelled corpora** (timestamp, tweet) covering all periods of training, development and testing. All these resources, including python-based baseline code, evaluation scripts, and un-labelled temporal data, will be made available to participating teams upon data release in December 2022.
2. In the **evaluation phase**, participants will be provided with three human-annotated testing sets without their labels (id, tweet): **(1) "within time"** acquired during time period t, **(2) short-term** acquired during a time interval t' occurring shortly after t (with no intersection between t and t') dedicated to evaluate ***short-term persistence (sub-task 2.A)***, and **(3) long-term** acquired long after t during a time interval t'' (with no intersection between t and t'') dedicated to evaluate ***long-term persistence (sub-task 2.B)***. Similarly to Task 1, participating teams are required to provide a performance score for the "within time" test set, even if they are only interested in one of the sub-tasks to calculate persistence metrics, i.e. RPD.

Evaluation The performance of the submissions will be evaluated using the following metrics:

1. **Macro-averaged F1-score** on the testing set of the corresponding sub-task

[2] https://figshare.com/articles/dataset/TM-Senti/16438281

2. **Relative Performance Drop (RPD)** measured by computing the difference between performance on “within time” data vs short- or long-term distant testing sets.

The submissions for each sub-task will be ranked based on the first metric of macro-averaged F1. In order to identify the best submission, we will also calculate a unified score between the two sub-tasks as a **weighted average between the scores obtained for each sub-task (weighted-F1)**. This will encourage participants to contribute to both sub-tasks in order to be correctly placed on a joint leader board, as well as to enable better analysis of their system performance in both settings.

Baseline Participants are expected to propose temporally persistent classifiers based on state-of-the-art data-centric or architecture-centric computational methods. The goal is to achieve high weighted-F1 performance across short and long temporally distant test sets while maintaining a reasonable RPD when compared to a test set from the same time period as training. We intend to use **RoBERTa**[3] [6] as a baseline classifier for our task because it has been demonstrated to be persistent over time [1].

3 LongEval Timeline

Information and updates about the LongEval Lab, the data and training / submission guidelines will be communicated mainly through the lab’s website https://clef-longeval.github.io. The training data for both tasks will be released in December 2022, and the test data in February 2023. Participant submission deadline is planned for the end of April 2023, with the evaluation results to be released in June 2023.

During the CLEF 2023 conference, LongEval will organize a one-day workshop, with participant presentations as well as 2-3 invited speakers. The workshop will welcome other submissions on the topic of temporal persistence that were not part of the shared task.

Acknowledgements

This work is supported by the ANR Kodicare bi-lateral project, grant ANR-19-CE23-0029 of the French Agence Nationale de la Recherche, and by the Austrian Science Fund (FWF, grant I4471-N). This work is also supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant no. EP/V030302/1) and The Alan Turing Institute (grant no. EP/N510129/1) through project funding and its Enrichment PhD Scheme for Iman Bilal. This work has been using services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports

[3] https://huggingface.co/roberta-base

of the Czech Republic (Project No. LM2018101) and has been also supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ.

References

1. Alkhalifa, R., Kochkina, E., Zubiaga, A.: Building for tomorrow: Assessing the temporal persistence of text classifiers. arXiv preprint arXiv:2205.05435 (2022)
2. Alkhalifa, R., Zubiaga, A.: Capturing stance dynamics in social media: open challenges and research directions. International Journal of Digital Humanities pp. 1–21 (2022)
3. Chapelle, O., Zhang, Y.: A dynamic bayesian network click model for web search ranking. In: Proceedings of the 18th international conference on World wide web. pp. 1–10. WWW '09, Association for Computing Machinery, New York, NY, USA (Apr 2009), https://doi.org/10.1145/1526709.1526711
4. Chuklin, A., Markov, I., Rijke, M.d.: Click models for web search. Synthesis Lectures on Information Concepts, Retrieval, and Services **7**(3), 1–115 (Jul 2015), https://doi.org/10.2200/S00654ED1V01Y201507ICR043
5. Florio, K., Basile, V., Polignano, M., Basile, P., Patti, V.: Time of your hate: The challenge of time in hate speech detection on social media. Applied Sciences **10**(12), 4180 (2020)
6. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
7. Lukes, J., Søgaard, A.: Sentiment analysis under temporal shift. In: Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis. pp. 65–71 (2018)
8. Ren, R., Qu, Y., Liu, J., Zhao, W.X., Wu, Q., Ding, Y., Wu, H., Wang, H., Wen, J.R.: A thorough examination on zero-shot dense retrieval (2022, arxiv:220412755). https://doi.org/10.48550/ARXIV.2204.12755, https://arxiv.org/abs/2204.12755
9. Yin, W., Alkhalifa, R., Zubiaga, A.: The emojification of sentiment on social media: Collection and analysis of a longitudinal twitter sentiment dataset. arXiv preprint arXiv:2108.13898 (2021)