



LongEval: Longitudinal Evaluation of Model Performance at CLEF 2024

Rabab Alkhalifa, Hsuvas Borkakoty, Romain Deveaud, Alaa El-Ebshihy, Luis Espinosa-Anke, Tobias Fink, Petra Galuščáková, Gabriela Gonzalez-Saez, Lorraine Goeuriot, David Iommi, et al.

► To cite this version:

Rabab Alkhalifa, Hsuvas Borkakoty, Romain Deveaud, Alaa El-Ebshihy, Luis Espinosa-Anke, et al.. LongEval: Longitudinal Evaluation of Model Performance at CLEF 2024. Advances In Information Retrieval (ECIR 2024), Mar 2024, Glasgow (Ecosse), United Kingdom. pp.60-66, 10.1007/978-3-031-56072-9_8 . hal-04577466

HAL Id: hal-04577466

<https://hal.science/hal-04577466v1>

Submitted on 8 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

LongEval: Longitudinal Evaluation of Model Performance at CLEF 2024

Rabab Alkhalifa^{1,2}, Hsuvas Borkakoty³, Romain Deveaud⁴, Alaa El-Ebshihy^{5,6}, Luis Espinosa-Anke^{3,7}, Tobias Fink^{5,6}, Petra Galuščáková⁹, Gabriela Gonzalez-Saez⁸, Lorraine Goeuriot⁸, David Iommi⁵, Maria Liakata^{1,10,11}, Harish Tayyar Madabushi¹², Pablo Medina-Alias¹², Philippe Mulhem⁸, Florina Piroi^{5,6}, Martin Popel¹³, Christophe Servan^{4,14}, and Arkaitz Zubiaga¹

¹ Queen Mary University of London, UK

² Imam Abdulrahman Bin Faisal University, SA

³ Cardiff University, UK

⁴ Qwant, France

⁵ Research Studios Austria, Data Science Studio, Vienna, AT

⁶ TU Wien, Austria

⁷ AMPLYFI, UK

⁸ Univ. Grenoble Alpes, CNRS, Grenoble INP**, LIG, Grenoble, France

⁹ University of Stavanger, Stavanger, Norway

¹⁰ Alan Turing Institute, UK

¹¹ University of Warwick, UK

¹² University of Bath, UK

¹³ Charles University, Prague, Czech Republic

¹⁴ Paris-Saclay University, CNRS, LISN, France

Abstract. This paper introduces the planned second LongEval Lab, part of the CLEF 2024 conference. The aim of the lab’s two tasks is to give researchers test data for addressing temporal effectiveness persistence challenges in both information retrieval and text classification, motivated by the fact that model performance degrades as the test data becomes temporally distant from the training data. LongEval distinguishes itself from traditional IR and classification tasks by emphasizing the evaluation of models designed to mitigate performance drop over time using evolving data. The second LongEval edition will further engage the IR community and NLP researchers in addressing the crucial challenge of temporal persistence in models, exploring the factors that enable or hinder it, and identifying potential solutions along with their limitations.

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

** Institute of Engineering Univ. Grenoble Alpes.

1 Introduction

The second edition of the LongEval CLEF 2024 shared task continues its exploration of the temporal persistence of Information Retrieval (IR) systems and Text Classifiers, building upon, and aiming to further, the insights from the first edition [1]. We extend its focus on evaluating system performance degradation over time using evolving data, consistent with prior research [4,10,8,3,12].

The previous edition of LongEval reinforced the evidence that the performance of information retrieval and classification systems is indeed influenced by the temporal evolution of data. In this year’s edition, the two tasks, retrieval and classification, are once again proposed. Task 1, related to Information Retrieval, deals with the scenario where web documents evolve over time, queries are not known in advance, relevance judgments are non-binary, and submissions must provide ranked lists as results. The Task 2 focuses on text classifiers in which the target classes are predefined while language usage associated with each class evolves rapidly over time, as in social media.

The goal of the LongEval 2024 lab is to promote the proposal of novel approaches that can automatically adapt to possible temporal dynamics in textual data. In doing so, we expect that new approaches will be able to foster time-insensitive computational retrieval and text classifiers methods. As such, the expected outcomes from this lab are threefold:

- to draw a deeper understanding of how time impacts IR and classification systems. The LongEval 2023 results need to be extended to a longer timeline to be more useful to the research community;
- to assess the effectiveness of different retrieval and classification approaches in achieving temporal persistence;
- to enable the advancement of computational methods that leverage ageing labelled datasets, while minimising performance drop over time.

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 present tasks and provide additional information about the data and the baselines to be used. Section 3 contains additional information and guidelines for participants.

2 Description of the Tasks

2.1 Task 1: LongEval-Retrieval

The LongEval Task 1 aims to support the development of Information Retrieval systems able to face temporal evolution. This task makes use of evolving Web data, in a way to evaluate retrieval systems longitudinally: the systems are expected to be persistent in their retrieval efficiency over time. The systems are evaluated on several collections of documents and queries, corresponding to

real data acquired from a French Web search engine, Qwant¹. The *LongEval-Retrieval* evaluation relies on the evaluation of **the same IR systems** on three test collections:

- Lag-3 (respectively Lag-12 and Lag-14 collection(s): a test collection acquired three (respectively twelve and fourteen) months after the last sample from the train collection

Lessons learned from 2023 37 teams registered for the first edition of the LongEval Retrieval task, and 14 teams submitted their runs. This number is quite large for a first edition. Several insights were learned [1]:

- No real proposal was specifically dedicated to cope with the evolution of the data
- The best approaches rely on large language model query expansion techniques
- The correlation between ranking of systems is similar for short and long lags
- The systems that are the more robust to the evolution of test collection were not the best performing ones

From these lessons, this year task enlarges the lags between the train and test collections, and provides three test collections in a way to provide a deeper understanding of the impact of the data evolution.

Data Globally, the dataset for 2024 is twice the size of 2023, as we use the train+test data of 2023 as the 2024 train set.

1. The train dataset is composed of 4M documents (in French, translated to English), as well as 3,000 of queries with associated computed relevance assessments from a simplified Dynamic Bayesian Network (sDBN) Click Model [5,6] acquired from real users of the French Qwant search engine. We will require the participants to not make any use of the assessments provided on the 2023 collections, but only the data of the 2023 train set.
2. The test collection is composed of three sub-collections: Lag-3, Lag-12 and Lag-14, corresponding to data acquired at several lags (3, 12 and 14 months from the last train data). Each of these test collections is similar to the train set, except that they do not contain any relevance assessments. The participants are expected to submit runs for each of the three lag collections, using the same system, i.e. a system trained only on the train dataset.

The total data for this task will be composed of 8 million documents and 6,000 queries, provided by Qwant². Each document set will have a release time stamp, with the first set (in chronological order) being the training data.

¹ Qwant being mostly used by French speaker, it explains why it is easier to gather data (user queries and documents) in this language rather than English.

² Qwant search engine: <https://www.qwant.com/>

Evaluation The submitted systems will be evaluated in two ways:

1. **nDCG** scores calculated on each lag 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 values between Lag-3 and Lag-12 datasets, Lag-3 and Lag-14 datasets as well as between Lag-12 and Lag-14 datasets. Such values will allow to check the robustness of systems against the evolution of the data.

These measures assess the extent to which systems provide good results, but also the extent to which they are robust against the data (queries/documents) evolution 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

Detecting the stance in social media posts is essential [9,11]. Yet, comprehending the evolution of social media stances over time poses a significant challenge [4,2], a topic that has gained recent interest in the AI and NLP communities but remains relatively unexplored. The performance of social media stance classifiers is intricately linked to temporal shifts in language and evolving societal attitudes toward the subject matter. In LongEval 2024, social media stance detection, a multi-label English classification task, takes center stage, surpassing the complexity of the binary sentiment task in LongEval 2023 [1]. Its primary aim is to assess the persistence of stance detection models in the dynamic landscape of social media posts.

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

Sub-task 2.A: Short-term persistence. In this sub-task participants will develop models which demonstrate performance persistence over short periods of time, i.e. using test set within 2-3 years apart from the training data.

Sub-task 2.B: Long-term persistence. In this sub-task participants will develop models which demonstrate performance persistence over longer period of time, i.e. test set within 4-5 years apart from the training data and also distant from the short-term test set.

The insights from the first edition of LongEval shed light on text classifiers' performance drop and highlighted its importance to the research community. To boost engagement and participation, we will release the collaboration (colab) platform, data, and starting kit ahead of time. This streamlined access encourages wider involvement from the research community. In the 2024 edition, we have expanded the dataset, focusing on a three-way classification task for stance detection. This expansion enriches research opportunities and addresses nuanced challenges in temporal persistence for text classification and opinion dynamics. The evolving nature of the task continues to uncover new dimensions in understanding temporal persistence and adaptability of text classifiers over time.

Data In this task, we will make use of, and extend with new annotations, the Climate Change Twitter dataset [7]. Our primary focus will be on climate change stance, time of the post (created at), and the textual content of the tweets, which we will refer to as the **CC-SD** dataset. This **CC-SD** is large-scale, covering a span of 13 years and containing a diverse set of more than 15 million tweets from various years. Using the BERT model to annotated tweets, the **CC-SD** stance labels in three categories: those that express support for the belief in man-made climate change (believer), those that dispute it (denier), and those that remain neutral on the topic. The total sum of the categorized tweets over all time span are as follows: 11,292,424 tweets as believers, 1,191,386 as deniers, and 3,305,601 as neutral, distributed across the timeline. The annotation is performed using transfer learning with BERT as distant supervision based on another sentiment climate change dataset ³ and, thus, can be easily manually annotated to improve its precision using manual annotation. We plan to release data in two phases:

1. In the **practice phase**, participants will be given **(1) a distantly annotated training set sampled from CC-SD** (tweet, label) created over a time interval t . Such data is dedicated for model training, as well as **(2) human-annotated “within time” practice set** (tweet, label) from the same time period t . **(3) human-annotated “short time” practice set** (tweet, label) distant from time period t . These two human practice sets are intended to allow participants to develop their systems before the following evaluation phase, and will not be used to rank their submissions. All these resources, including python-based baseline code and evaluation scripts, will be made available to participating teams upon data release.
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 interested in one of the sub-tasks to calculate persistence metrics, i.e. RPD.

Evaluation Metrics Evaluation metrics for this edition of the task remain consistent with the previous version. All submissions will be assessed using two key metrics: the **macro-averaged F1-score** on the corresponding sub-task’s testing set and the **Relative Performance Drop (RPD)**, calculated by comparing performance on “within time” data against results from short- or long-term distant testing sets. Submissions for each sub-task will be ranked primarily based on the macro-averaged F1-score. Additionally, a unified score, **the weighted-F1**,

³ <https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset>

will be computed between the two sub-tasks, encouraging participants to contribute to both for accurate placement on a collective leaderboard and a deeper analysis of their system’s performance in various settings.

Baseline Participants are expected to propose temporally persistent classifiers based on state-of-the-art 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 **BERT**⁴ [7] as a baseline classifier.

3 LongEval Timeline

Information and updates about the LongEval Lab, and the submission guidelines, will be communicated mainly through the lab’s website⁵. The training data for both tasks will be released in December 2023, and the test data in February 2024. Participant submission deadline is planned for May 2024, with the evaluation results to be released in June 2024. During the CLEF 2024 conference, LongEval will organize a workshop, with participant presentations as well as 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). 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 of the Czech Republic (Project No. LM2023062) and has been also supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

References

1. Alkhalifa, R., Bilal, I.M., Borkakoty, H., Romain, Deveaud, El-Ebshihiy, A., Luis, Espinosa-Anke, Gabriela, Gonzalez-Saez, Galuscáková, P., Goeuriot, L., Kochkina, E., Liakata, M., Loureiro, D., Mulhem, P., Piroi, F., Popel, M., Servan, C., Madabushi, H.T., Arkaitz, Zubiaga: Extended overview of the clef-2023 longeval lab on longitudinal evaluation of model performance (2023), <https://api.semanticscholar.org/CorpusID:259953335>

⁴ <https://huggingface.co/bert-base-uncased>

⁵ <https://clef-longeval.github.io>

2. Alkhalifa, R., Kochkina, E., Zubiaga, A.: Opinions are made to be changed: Temporally adaptive stance classification. In: Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks. pp. 27–32 (2021)
3. Alkhalifa, R., Kochkina, E., Zubiaga, A.: Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management* **60**(2), 103200 (2023)
4. Alkhalifa, R., Zubiaga, A.: Capturing stance dynamics in social media: open challenges and research directions. *International Journal of Digital Humanities* pp. 1–21 (2022)
5. 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>
6. 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>
7. Effrosynidis, D., Karasakalidis, A.I., Sylaios, G., Arampatzis, A.: The climate change twitter dataset. *Expert Systems with Applications* **204**, 117541 (2022). <https://doi.org/https://doi.org/10.1016/j.eswa.2022.117541>, <https://www.sciencedirect.com/science/article/pii/S0957417422008624>
8. 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)
9. Küçük, D., Can, F.: Stance detection: A survey. *ACM Comput. Surv.* **53**(1) (feb 2020). <https://doi.org/10.1145/3369026>, <https://doi.org/10.1145/3369026>
10. Lukes, J., Søgaaard, 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)
11. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in Tweets. *ACM Transactions on Internet Technology* **17**(3) (2017). <https://doi.org/10.1145/3003433>, <http://alt.qcri.org/semeval2016/task6/>
12. 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>