Overview of the CLEF 2024 LongEval Lab on Longitudinal Evaluation of Model Performance

Rabab Alkhalifa[1,2], Hsuvas Borkakoty[3], Romain Deveaud[4], Alaa El-Ebshihy[5,6], Luis Espinosa-Anke[3,7], Tobias Fink[5,6], Petra Galuščáková[9], Gabriela Gonzalez-Saez[8], Lorraine Goeuriot[8], David Iommi[5], Maria Liakata[1,10,11], Harish Tayyar Madabushi[12], Pablo Medina-Alias[12], Philippe Mulhem[8], Florina Piroi[5,6], Martin Popel[13], and Arkaitz Zubiaga[1]

[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

Abstract. We describe the second edition of the LongEval CLEF 2024 shared task. This lab evaluates the temporal persistence of Information Retrieval (IR) systems and Text Classifiers. Task 1 requires IR systems to run on corpora acquired at several timestamps, and evaluates the drop in system quality (NDCG) along these timestamps. Task 2 tackles binary sentiment classification at different points in time, and evaluates the performance drop for different temporal gaps. Overall, 37 teams registered for Task 1 and 25 for Task 2. Ultimately, 14 and 4 teams participated in Task 1 and Task 2, respectively.

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

1 Introduction

Outside the strict scientific context, the European Artificial Intelligence Act[14], adopted by European Commission in 2024, stresses in Article 17, section (d),

---

** Institute of Engineering Univ. Grenoble Alpes.

[14] `https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html`

that providers must comply with “examination, test and validation procedures to be carried out before, during and after the development of the high-risk AI system, and the frequency with which they have to be carried out”. Without focusing here on the degree of risk of Information Retrieval or Classification systems, this Act clearly states that AI systems must tackle evolution. Time is a dimension that is often overlooked when conducting Information Retrieval (IR) experiments, especially when static data sets are utilized. The advantages of such datasets are that they are easily used to evaluate and test systems. Some data sets, like CORD19, contain documents collected at different points in time, showing differences in the set of documents from one collection time to another. Recent research [15] has demonstrated that models trained on data pertaining to a particular time period struggle to keep their performance levels when applied on test data that is distant in time. On the other side, [22] showed that neural systems, especially transformers-based ones, are not always very sensitive to corpus evolution.

With the aim of tackling this challenge of making models have persistent quality over time, the objective of the LongEval lab is twofold: (i) to explore the extent to which temporal differences over time, as reflected in the evolution of evaluation datasets, results in the deterioration of the performance of information retrieval and classification systems, and (ii) to propose improved methods that mitigate performance drop by making models more robust over time.

The LongEval lab [3] took place as part of the Conference and Labs of the Evaluation Forum (CLEF) 2024, and consisted in two separate tasks: (i) Task 1, described in Section 2, focused on information retrieval, and (ii) Task 2, described in Section 3, focused on text classification for sentiment analysis. Both tasks provided labeled datasets enabling analysis and evaluation of models over data evolving in time (what we call “longitudinally evolving data”).

2 Task 1 - Retrieval

The retrieval task of LongEval 2024 explores the effect of changes in datasets on retrieval of text documents. More specifically, we focus on a setup in which the datasets are evolving, as in the LongEval 2023 Retrieval Task data [3]. This means, that one dataset can be acquired from another by adding, removing (and replacing) a limited number of documents and queries. The two main scenarios considered focus on one single system or on several ones, as detailed below:

A single system in an evolving setup

We explore how one selected system behaves when evaluated on several collections, which evolve along the time. The context in which this task taked place is retrieval performances for **Web search**. When considering evolution of Web data along time, we are facing a case when the documents, the queries and also the relevance continuously evolves. We are then studying how Web search engines deal with this situation. The considered scenario is then similar to classical *ad-hoc* search, in the case of evolving data sets. The evaluation in this scenario consider both the Web search case in which the top documents are the most important elements considered, and should take into account the evolving nature of the data. Evaluation should ideally reflect the changes in the collection and especially signal substantial changes that could lead to performance drop. This would allow to re-train the search engine model then and only when it is really necessary, and enable much more efficient overall training.

As described earlier, there is no consensus about the stability of the performance of the neural networks IR systems along time, but it seems to be lower than in the case of statistical models. Moreover, the performance strongly depends on the data used for training the neural model. One objective of the task is to explore the behavior of the neural system in the evolving data scenario.

Comparison of multiple systems in an evolving setup
While in the first point, we explore a single system, comparison of this systems with multiple systems across evolving collections, should provide more information about systems stability and robustness.

2.1 Description of the task

Compared to the LongEval 2023 Dataset [3], in 2024 we take larger lags between the training and the test sets. More precisely, the task is composed of:

- One training set, that contains Web documents, actual user's queries, and assessments, acquired at timestamp t;
- Two test sets, acquired later than t at time t' and t", composed of Web documents and user's queries.

The task datasets were created over sequential time periods, which allows doing observations at different time stamps t, and most importantly, comparing the performance across different time stamps t and t'. So, the IR task aims to assess the performance difference between t' and t" when t' occurs after t', according to teh fact that training set acquired at t, takes place few months before t'.

2.2 Dataset

As for LongEval 2023, in 2024 the data for this task were provided by the French search engine Qwant. They consist of the queries issued by the users of this search engine, cleaned Web documents, which were 1) selected to correspond to the queries, and 2) to add additional noise, and relevance judgments, which were created using a click model. The dataset is fully described in [14]. We provided training data, which included 599 train queries, with corresponding 9,785 relevance assessments and 2,049,729 Web pages. All training data were collected during January 2023. The test set corpus is composed of two subsets: Lag6 acquired in June 2023 (i.e., 6 months later than the training set), and Lag8 acquired in August 2024 (i.e. acquired 8 months later than the training set). The test dataset contains 4,321,642 documents (June: 1,790,028; August: 2,531,614) and 1,925 test queries (June: 407; August: 1,518). The datasets are accessible through the lab's webpage[15] and from the TU Wien Research Data Repository[16].

The data collected from the Qwant search engine is in French. In a way to help participants, the LongEval data set for the Retrieval task also contains automatic translations into English of both queries and documents. We mention however that the translations provided by LongEval are only applied to the first 500 characters of each sentence of the initial French documents downloaded.

The document and query overlap ratios between the collections is given by Table 1 and Table 2. We see from these tables that there is a substantial overlap between the Train and the Test collection documents and (due to the larger size of the August query set) a substantial overlap between the Train / June queries and the August queries.

Table 1: Ratio of documents shared between the LongEval 2024 train and test collections, row vs. column, i.e. 0.93 means that 93% of documents in the row collection are also included in the column collection.

	Train 2024	**June (Lag6)**	**August (Lag8)**
Train 2024	1.00	0.67	0.93
June (Lag6)	0.77	1.00	0.97
August (Lag8)	0.75	0.69	1.00

Table 2: Ratio of the queries shared between the LongEval 2024 train and test collections, rows vs. columns, i.e. 0.99 means that 99% of queries in the row collection are also included in the column collection.

	Train 2024	**June (Lag6)**	**August (Lag8)**
Train 2024	1.00	0.22	0.42
June (Lag6)	0.32	1.00	0.56
August (Lag8)	0.17	0.15	1.00

To evaluate the submissions we use one set of relevance judgments: the judgments acquired by the Qwant click model. For the evaluation, we use the NDCG measure (calculated for each dataset) at 10, as well as the drop between the Lag8 and Lag6 collection. This allows us to check to which extend the IR system face the evolution of the data. We also plan to use manual assessments, acquired through the interface described in section 2.8.

2.3 Submissions

14 teams submitted their systems to the Retrieval task. Each team was allowed to submit up to 10 systems. Together, this a overall of 73 runs submitted. Two teams submitted their runs on the wrong test data set, so we do not include their submission results in our further analysis.

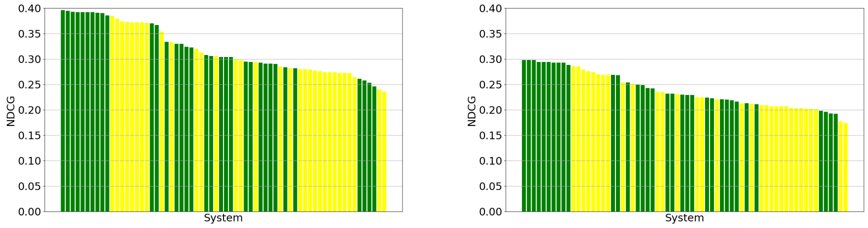
---

[15] `https://clef-longeval.github.io/`

[16] `https://doi.org/10.48436/xr350-79683`

2.4 Absolute Scores

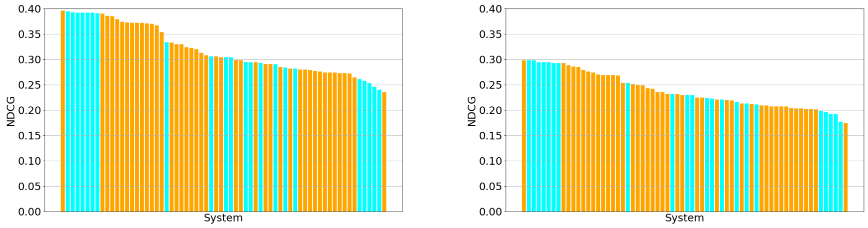
For the Retrieval task of the LongEval lab, we computed two sets of scores for each of the lags in the test collection, namely NDCG and MAP. Table 3 gives the overview of them for each run on the Lag6 and Lag8 datasets. For each run, the columns of the table indicate which language was used (English, French, or both), whether neural approaches were involved (values yes/no), and whether a single or a combination of several approaches was used (values yes/no). In addition, we show NDCG score histograms for these runs, in decreasing order, for each dataset, showing whether a run uses any neural approach (green for yes, yellow for no) in Figure 1, and whether the run uses a combination of more than a single approach (orange for yes, cyan for no) in Figure 2. This information was acquired from the participants through a questionnaire the participants had to fill for each submitted run. Figure 3 shows which language each made use of.



(a) Lag6 Dataset

(b) Lag8 Dataset

Fig. 1: Overview of the systems using a neural approach (green) vs. other (yellow).



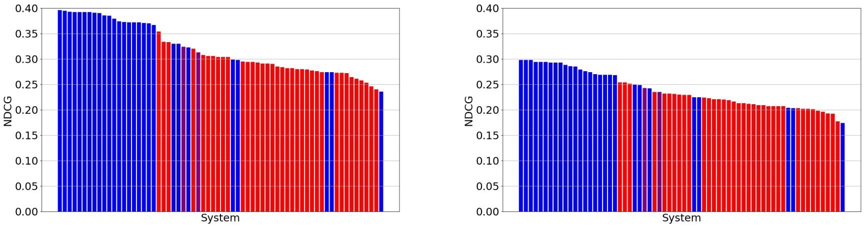
(a) Lag6 Dataset

(b) Lag8 Dataset

Fig. 2: Overview of the systems which use a single approach (orange) and which use a combination of multiple approaches (cyan)

From Table 3 we see that the systems which did best for the Lag6 data are also among the top for the Lag8, where the first ranked nine systems scores are comparable to each other. For instance, the best system on Lag6, according to the

NDCG measure, (dam_run_4), is ranked the second best also on Lag8. Similarity, the best system on Lag8, according to the NDCG measure, (mouse_run_8), is ranked the second best also on Lag6. This finding holds for the MAP measure as well.



(a) Lag6 Dataset

(b) Lag8 Dataset

Fig. 3: Overview of the systems which use French (blue), which use English translations (red), and which use both (purple).

Most of the solutions chosen by the participants to the LongEval Retrieval task apply a multi-stage retrieval approach. Often, the first stage involves a lexical-based retrieval (e.g., BM25), and query expansion methods like PL2 or BO1. Query expansion is also done by employing Large Language Models, like Mistral or Llama 3. Reranking is done either using neural-based methods or sentence based transformers. Listwise rerankers and fusing have also been used in reranking of retrieved results. Notably, the temporal aspect of the LongEval test collection has been used by some participants to include past query relevance information into query reformulation either from clicklogs or from the documents deemed relevant in the previous

Considering the Figures 1, 2 and 3, we see that the shape of the distribution of the NDCG values are similar for the Lag6 and Lag8 datasets. However, the systems have higher performances on Lag6 than on Lag8, with maximum 0.4 value for the NDCG on the Lag6 versus 0.3 for the Lag8.

Table 3: NDCG and MAP scores for Lag6, Lag8. Results are sorted according to the NDCG scores on the Lag6.

				NDCG		**MAP**	
Run Id	**Neural**	**Comb.**	**Language**	**Lag6**	**Lag8**	**Lag6**	**Lag8**
dam_run_4	yes	no	French	0.396	0.294	0.249	0.171
mouse_run_8	yes	yes	French	0.395	0.298	0.248	0.174
mouse_run_10	yes	yes	French	0.393	0.298	0.246	0.175
iris_run_4	yes	yes	French	0.392	0.293	0.244	0.171
mouse_run_9	yes	yes	French	0.392	0.298	0.245	0.175
iris_run_1	yes	yes	French	0.392	0.294	0.244	0.171

iris_run_2	yes	yes	French	0.392	0.293	0.242	0.170
iris_run_3	yes	yes	French	0.391	0.293	0.243	0.171
iris_run_5	yes		French	0.390	0.294	0.240	0.171
mouse_run_7	yes	no	French	0.386	0.288	0.236	0.163
dam_run_3	no	no	French	0.385	0.285	0.235	0.162
quokkas_run_2	no	no	French	0.379	0.276	0.225	0.150
quokkas_run_1	no	no	French	0.374	0.274	0.221	0.148
lfzzo_run_7	no	no	French	0.373	0.269	0.221	0.145
lfzzo_run_8	no	no	French	0.372	0.269	0.221	0.144
lfzzo_run_9	no	no	French	0.372	0.268	0.221	0.143
lfzzo_run_10	no	no	French	0.372	0.269	0.219	0.145
lfzzo_run_6	no	no	French	0.371	0.270	0.218	0.145
dam_run_5	yes	no	French	0.370	0.279	0.220	0.156
mouse_run_6	yes	no	French	0.367	0.286	0.215	0.162
cir_run_3	no	no	English	0.354	0.242	0.226	0.136
snu_run_1	yes	yes	English	0.334	0.251	0.197	0.142
ows_run_1	no	no	English	0.333	0.243	0.199	0.139
kalu_run_2	yes	no	French	0.330	0.254	0.192	0.143
kalu_run_3	yes	no	French	0.330	0.254	0.192	0.143
kalu_run_5	yes	no	Frencg	0.324	0.249	0.188	0.140
kalu_run_4	yes	no	French	0.323	0.250	0.186	0.140
cir_run_4	no	no	English	0.320	0.229	0.172	0.117
wonder_run_3	no	no	French,English	0.313	0.235	0.163	0.116
cir_run_2	yes	no	English	0.308	0.230	0.173	0.123
mouse_run_3	yes	yes	English	0.306	0.235	0.171	0.126
ows_run_2	no	no	English	0.306	0.229	0.197	0.140
dam_run_2	yes	no	English	0.304	0.231	0.169	0.121
mouse_run_4	yes	yes	English	0.304	0.232	0.167	0.124
mouse_run_5	yes	yes	English	0.304	0.232	0.166	0.124
wonder_run_4	no	no	French	0.299	0.223	0.155	0.107
kalu_run_1	no	no	French	0.298	0.219	0.158	0.107
galapagos_run_4	yes	yes	English	0.295	0.220	0.189	0.131
ows_run_3	yes	yes	English	0.294	0.224	0.188	0.135
dam_run_1	no	no	English	0.294	0.221	0.156	0.112
galapagos_run_5	yes	yes	English	0.293	0.221	0.187	0.132
mouse_run_2	yes	no	English	0.291	0.225	0.152	0.115
mouse_run_1	yes	no	English	0.291	0.225	0.153	0.114
ows_run_7	yes	yes	English	0.290	0.213	0.180	0.123
cir_run_5	no	no	English	0.285	0.212	0.148	0.104
ows_run_6	yes	yes	English	0.284	0.216	0.173	0.126
cir_run_1	no	no	English	0.282	0.211	0.145	0.103
snu_run_2	yes	yes	English	0.282	0.213	0.177	0.127
lfzzo_run_4	no	no	English	0.280	0.209	0.142	0.102
lfzzo_run_2	no	no	English	0.280	0.207	0.142	0.099
wonder_run_2	no	no	English	0.279	0.207	0.137	0.099

lfzzo_run_3	no	no	English	0.277	0.209	0.139	0.102
lfzzo_run_1	no	no	English	0.276	0.207	0.140	0.100
lfzzo_run_5	no	no	English	0.274	0.207	0.137	0.101
seekx_run_1	no	no	French	0.274	0.201	0.145	0.095
seekx_run_2	no	no	French	0.274	0.202	0.144	0.096
seekx_run_4	no	no	English	0.273	0.202	0.139	0.098
wonder_run_5	no	no	English	0.273	0.203	0.137	0.098
wonder_run_1	no	no	English	0.272	0.203	0.136	0.098
seekx_run_5	no	no	English	0.264	0.193	0.133	0.091
galapagos_run_2	yes	yes	English	0.261	0.198	0.162	0.115
galapagos_run_1	yes	yes	English	0.258	0.196	0.157	0.111
galapagos_run_3	yes	yes	English	0.253	0.192	0.151	0.107
ows_run_4	yes	yes	English	0.246	0.204	0.128	0.114
ows_run_5	no	yes	English	0.240	0.177	0.124	0.085
seekx_run_3	no	no	French	0.236	0.174	0.120	0.079
AVERAGE				**0.318**	**0.238**	**0.183**	**0.129**

2.5 Changes in the Scores

The main part of the retrieval task is to study the changes in the performance scores between the collections. The collections were created using the same approach and procedure have a relatively high overlap in terms of both queries and documents (see Tables 1 and 2), we thus provide the Relative NDCG Drop (RND) values of systems between the collections Lag8 and Lag6. RnD(r) for a system r, is defined as as:

$$RND(r) = \frac{\text{NDCG}_{Lag6}(r) - \text{NDCG}_{Lag8}(r)}{\text{NDCG}_{Lag6}(r)}$$

With such definition, small RND values man more robust systems against changes, and large RND values mean that the systems are not able to generalize well between lag6 and lag8. What we see in Table 4 is that the systems which are more robust to the evolution of the test collections (low values on RND) are not the best ones: for instance, ows_run_4 is the more robust system but the third worse one in table 3. The best systems in term of NDCG values in lag6, dam_run_4 and $mouse_run$_8, have an RND of 0.245, which means that they quite robust, but much less than the most robut ones. This shows that the very best systems do cope with some extend to the evolution of the corpus, but that their is room for improving best systems against robustness. We also see that the worse robust system against changes, cir_run_3, is a system that does not rely on neural IR models: such finding shows that neural models are also likely to be more robust against changes than non-neural ones.

Table 4: Changes in the NDCG scores (RND). Lines are ordered by descending RND values. Due to the space, only the most robust run per team in terms of RND is shown.

	NDCG		**RND**
System	**Lag6**	**Lag8**	
ows_run_4 [1]	0.246	0.204	0.169
mouse_run_6 [9]	0.367	0.286	0.220
kalu_run_4 [19]	0.323	0.250	0.224
galapagos_run_1 [17]	0.258	0.196	0.239
dam_run_2 [8]	0.304	0.231	0.241
lfzzo_run_3	0.277	0.209	0.243
snu_run_2 [24]	0.282	0.213	0.245
wonder_run_3	0.313	0.235	0.247
iris_run_5 [13]	0.390	0.294	0.248
cir_run_2 [18]	0.308	0.230	0.252
seekx_run_4	0.273	0.202	0.260
quokkas_run_1	0.374	0.274	0.268

2.6 Run Rankings

Another point of view studied is how the submitted runs compare to each other, either in terms of the absolute NDCG scores achieved on the collections, or in terms of NDCG changes between the collections. We also calculated the Pearson correlation between the runs (now shown here), with high correlation in terms of NDCG scores, 0.99, and similarly high, 0.98, with respect to ranking order. This corresponds to the relatively high overlaps of the documents and also the queries between Lag6 and Lag8 collections (Table 1 and Table 2). This observation does not hold for the correlation between the ranking according to the NDCG score achieved and the ranking of the performance change, which is relatively low. The Pearson correlation is 0.07 for the Lag6 dataset and -0.05 on the Lag8 dataset.

Last, we calculated a combination of both rankings (ranking in terms of absolute values and ranking in terms of change). For this, we first calculated a Borda count of the ranking in terms of absolute values and Borda count of the ranking in terms of relative change and then we simply summed these two Borda counts: this result is displayed in the last column in the Table 5. We see that in terms of this measure the top performing systems (on Lag6 and Lag8 datasets) are ranked higher, although they have lower rank in terms of the rank of the NDCG change.

2.7 Queries Overview

We further investigate performance on the provided queries. Due to the space reason, we only investigate a selected subset of queries from each collection.

Table 5: Ranking of the submitted systems by NDCG scores (columns 2-3), changes in NDCG scores between Lag6 and Lag8 dataset (column 4). Column 4 shows the sum of the Borda count applied to ranking on Lag6 and Lag8 datasets and Borda count of ranking change between Lag8 and Lag6 dataset. The darker color means better performance. Here we show the ranking of the selected runs in Table 4.

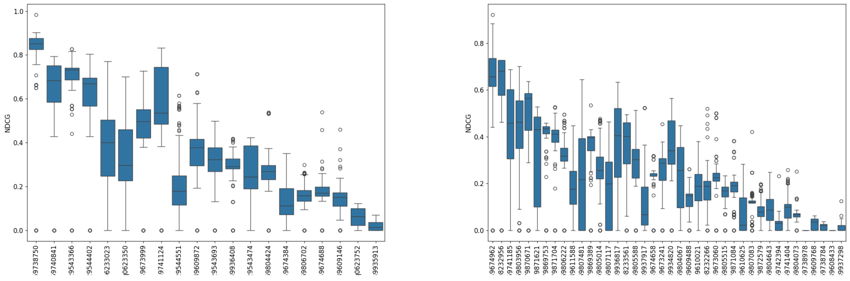
System	**NDCG Lag6**	**NDCG Lag8**	**RND**	**Borda**
iris_run_5	9	6	26	160
quokkas_run_1	13	15	57	116
mouse_run_6	20	11	2	168
kalu_run_4	27	24	3	147
wonder_run_3	29	29	25	118
cir_run_2	30	33	37	101
dam_run_2	33	32	17	119
snu_run_2	48	46	21	86
lfzzo_run_3	52	50	20	79
seekx_run_4	57	58	48	38
galapagos_run_1	62	62	13	64
ows_run_4	64	55	1	81

We used a pooling strategy to select these queries to be used for the manual assessment process (described in Section 2.8). We first selected the top five performing runs on the average NDCG performance on both collections. We then calculated the performance of these runs per queries for each collection (i.e. Lag6 and Lag8) and sorted the queries based on their NDCG performance for the five runs. Then, we divided the query set in each collection to four sets and randomly selected from each set: five and 10 queries from Lag 6 and Lag8, respectively. We selected in total 20 queries from Lag6 collection and 40 Lag8 collection. We selected more queries from Lag8 collection since, as shown in Table 2, the number of Lag8 collection is higher than Lag6 collection.

Overview of the scores achieved for the selected queries in each collection is displayed in Figure 4. The figure shows minimum performance (by any submitted run), 25%, quantile, 75% quantile and the maximum achieved NDCG score. Due to a relatively large number of runs, the range of the scores achieved is typically quite large and for some of the queries it even ranges between 0 and 0.8. It can be also noticed that the variation (corresponding to the size of the boxplot) of the query performance for the Lag8 collection is higher than Lag6 collection.

Some of the worst performing queries are very general (“birdsong”, “taxes”, and “used car” for instance) and can thus be expected to be ambiguous. This is in contrast with the top performing queries (e.g. “camping concarneau”, “Prune rabbit”, and “point bordeaux vision”) which refer to more specific information need. Some other top performing queries have high variation in the results, e.g.

the query “origami bird” for which it is not specified if the user focuses about about "origami bird" or looks for tutorials to make them.



(a) Lag6 Dataset

(b) Lag8 Dataset

Fig. 4: Selected queries performance from Lag6 and Lag8 datasets.

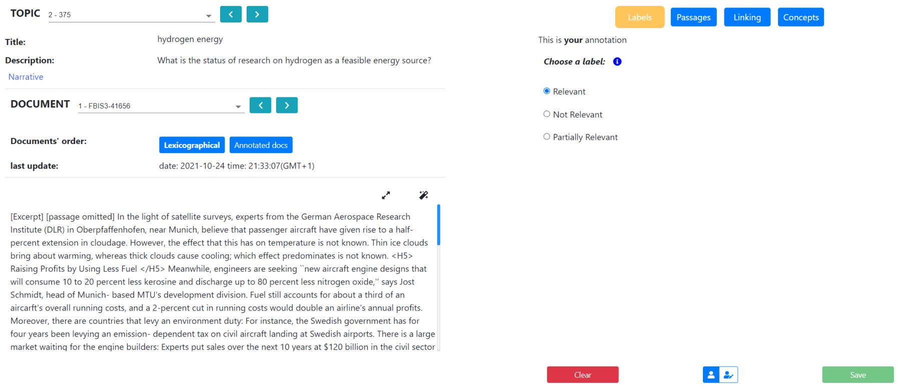
2.8 Manual relevance judgments acquisition

The evaluation results of LongEval IR task presented above rely on automatic assessments generated from click models [14]. In addition to these click-based relevance assessments, we have set up an annotation tool to acquire further relevance assessments by humans. For that, we used the open source annotation tool, Doctag [16], on a sample of the queries selected in section 2.7 (60 queries in total). Doctag provides a customizable and portable platform specifically designed for Information Retrieval (IR) evaluation. To perform manual relevance judgments using Doctag, annotators utilize its web-based interface. They access the tool and interact with its annotation functionalities, including the assignment of labels to indicate document relevance to specific queries. Annotators view the documents and associate appropriate relevance labels (Fig. 5). We set up dedicated online servers where Doctag is deployed, through their use we have acquired over 25K manual assessments which we intend to use in further evaluations.

2.9 Discussion and conclusion

This task was the second attempt to collectively investigate the impact of the evolution of the data on search system’s performances. Having 14 participating teams submitting runs confirmed that this topic was of interest to the community.

The dataset released for this task consisted in a sequence of test collections corresponding to different times. The collections were composed of documents and queries coming from Qwant, and relevance judgment coming from a click model and manual assessment. While the manual assessment is ongoing at the



TOPIC 2 - 375

Title: hydrogen energy

Description: What is the status of research on hydrogen as a feasible energy source?

DOCUMENT 1 - FBIS3-41656

Documents' order: Lexicographical Annotated docs

last update: date: 2021-10-24 time: 21:33:07(GMT+1)

[Excerpt] [passage omitted] In the light of satellite surveys, experts from the German Aerospace Research Institute (DLR) in Oberpfaffenhofen, near Munich, believe that passenger aircraft have given rise to a half-percent extension in cloudage. However, the effect that this has on temperature is not known. Thin ice clouds bring about warming, whereas thick clouds cause cooling; which effect predominates is not known. <H5> Raising Profits by Using Less Fuel </H5> Meanwhile, engineers are seeking ``new aircraft engine designs that will consume 10 to 20 percent less kerosine and discharge up to 80 percent less nitrogen oxide,'' says Jost Schmidt, head of Munich- based MTU's development division. Fuel still accounts for about a third of an aircarft's overall running costs, and a 2-percent cut in running costs would double an airline's annual profits. Moreover, there are countries that levy an environment duty: For instance, the Swedish government has for four years been levying an emission- dependent tax on civil aircraft landing at Swedish airports. There is a large market waiting for the engine builders: Experts put sales over the next 10 years at $120 billion in the civil sector

This is **your** annotation

Choose a label:

● Relevant
○ Not Relevant
○ Partially Relevant

Fig. 5: Screenshot from Doctag main page. Labels annotation is done associating to each document one label that expresses the relevance of that document for that topic.

time of the paper's publication, performances of participants' submitted runs were measured using the click logs.

Most of submitted runs rely on multi-stage retrieval approaches. In addition to the usage of Large Language Models in Query expansion. The effect of the translation of the documents and queries provided by the lab has a clear impact: the best results were obtained on the original French data.

Since each subset had substantial overlaps, the correlations between systems rankings was pretty high. As for the robustness of the systems towards dataset changes, we observed that the systems that are the more robust to the evolution of test collection were not the best performing ones.

Further evaluations will be carried out in the near future with the manual assessment of the pooled sets. A thorough analysis of the results will be necessary to study the impact of queries on the results (their nature, topic, difficulty, etc.). Further analysis work will be necessary to fully establish the robustness of the systems and the specific impact of dataset evolution on the performances.

3 Task 2 - Classification

Stance detection, an essential task in natural language processing (NLP), involves identifying an author's position or attitude towards a particular topic or statement. This task goes beyond simple sentiment analysis by requiring models to discern not just positive or negative sentiments but also the specific stance (supporting/believer, opposing/denier, or neutral) towards a given target [20,23].

Comprehending the evolution of social media stances over time poses a significant challenge, 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 [7].

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 [2]. Our primary goal is to assess the persistence of stance detection models in the dynamic landscape of social media posts.

The evolving nature of language and social opinions adds an additional layer of complexity to the challenges faced by text classifiers. Language undergoes continuous changes, reflecting shifts in societal norms and opinions and the emergence of novel concepts and words. For instance, consider the evolution of public opinion on climate change over the past two decades:

- **Sentence from 2000:** “Global warming is a theory that needs more proof; it’s not urgent.”
- **Sentence from 2010:** “Evidence for climate change is mounting, and we need to start taking action.”
- **Sentence from 2020:** “Climate change is an undeniable crisis that requires immediate global action.”

The context over two decades in the above example shows that language and urgency surrounding climate change have evolved from skepticism to an accepted crisis. Models not updated with recent discussions and policy changes might fail to accurately capture the critical tone and terminology used in current dialogues about the environment. Similarly, the rapid emergence of new vocabulary, as witnessed with terms like COVID-19 [6], highlights the dynamic nature of language, presenting unique challenges for text classifiers.

3.1 Description of the task

To assess the extent of the performance drop of models over shorter and longer temporal gaps, we provided a comprehensive training dataset along with five testing sets. These testing sets include two practice sets and three development sets. The shared competition aimed to stimulate the development of classifiers that can effectively handle temporal variations and maintain performance persistence over different time distances. Participants were expected to submit solutions for two sub-tasks, showcasing their ability to address the challenges of temporal variations in performance. The shared task was in turn divided into two sub-tasks:

Sub-Task 1: Short-Term Persistence: In this sub-task, participants were tasked with developing models that demonstrated performance persistence over short periods. Specifically, the models needed to maintain their performance over a temporal gap between the within datasets and the short-term datasets. This involved comparing the performance from the **within-practice** data (January 2010 to December 2010) to the **short-practice** data (January 2014 to December 2014), a time gap of 4 years, and from the **within-dev** data (January 2011 to December 2011) to the **short-dev** data (January 2015 to December 2015), a time gap of 4 years

Sub-Task 2: Long-Term Persistence: This sub-task required participants to develop models that maintained performance persistence over a longer period of time. The classifiers were expected to mitigate performance drops over a temporal gap between the within time datasets and the long-term datasets. This involved comparing the performance from the **within-dev** data (January 2011 to December 2011) to the **long-dev** data (January 2018 to September 2019), a time gap of approximately 7 to 8 years.

In addition to the main sub-tasks, participants were also asked to work on models that maintained performance within the same temporal year of the training set, with the **practice-within** data covering January 2010 to December 2010 and the **within-dev** data covering January 2011 to December 2011, with no gap between them and the training set (time gap 0).

3.2 Dataset

In this section, we present the process of constructing our final annotated corpus for the task. The large-scale Climate Change Twitter dataset was originally described in [11], 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 fall into 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 the entire 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 [17] and, thus, can be easily manually annotated to improve its precision using human in the loop.

Data sampling. The dataset is first downsampled to ensure an equal number of instances for each stance (neutral, denier, believer) within a specified date range, using the minimum stance count across all selected months and years to avoid bias. This involves randomly sampling the same number of rows for each stance, year, and month combination, ensuring balanced representation. The downsampled data is then shuffled and split into training, development, and practice sets, including short- and long-term coverage, with any intersecting IDs between these sets being removed to maintain data integrity and prevent data leakage. Finally, a summary of the downsampled data is generated, detailing the number of rows, date and time of sampling, and statistics per year and month.

Test set annotation. We annotate our test data using Prolific[18], which is a high quality data collection and annotation platform. The forms that contain

---

[17] https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset

[18] `https://www.prolific.com/`

data to annotate are created using Qualtrics[19]. We run the annotation in several batches, and provide the annotation guideline stating the task details and guidelines for the participants to follow. We add several filters, automatic and manual to select the optimal demographic and qualified annotators. Additionally, a manual annotation is also enforced which contains 5 tweets from the training set, which the organisers first annotate and then using the majority annotation is released as qualification task. The participant have to correctly answer 4 out of 5 questions to access the actual annotation task. We also provide fields in our form for every annotator to give their feedback and to point out if any tweet is inappropriate or contains explicit content in it. We collect responses from 5 annotators for each tweet, and select the majority annotation from the five annotation. In some cases, we find equal agreement among the annotators, and for those cases, we run an extra round of annotation to finalise the agreement. Finally after cleanup and majority annotation finding process, we manually check the data and divide into their respective splits.

The resulting distribution of data is shown in Table 6. table Dataset statistics summary of training, practice and testing sets.

Table 6: Dataset statistics summary of training, practice and testing sets.

Dataset	**Time Period**	**Size**
train	January 2009 to December 2011	35739
within-practice	January 2010 to December 2010	450
short-practice	January 2014 to December 2014	450
dev-within	January 2011 to December 2011	1074
dev-short	January 2015 to December 2015	1074
dev-long	January 2018 to September 2019	1074

In the Practice phase, participants undertake Pre-Evaluation tasks with datasets from 2010 and 2014, sampled from CC-SD, allowing them to practice within a recent time frame and over a short duration. These datasets are manually verified. Additionally, human-annotated "within time" and "short time" practice sets are provided, also sampled from CC-SD, to refine model development before formal evaluation.

Subsequently, the Evaluation phase assesses models using datasets from 2011, 2015, and the longer period of 2018-2019, all sampled from CC-SD. These datasets undergo manual verification and encompass within-timeframe assessments, short-term predictions, and long-term predictions, offering a holistic evaluation of model performance across various temporal contexts. By incorporating datasets covering different years, the evaluation ensures thorough testing and understanding of models' temporal persistence and performance.

---

[19] `https://www.qualtrics.com/`

3.3 Evaluation

Evaluation metrics for this edition of the task remain consistent with the previous version [3,4]. All submissions were assessed using two key metrics: the **macro-averaged F1-score** on the corresponding sub-task's development set and the **Relative Performance Drop (RPD)**, calculated by comparing performance on "within time" data against results from short- or long-term distant development sets. Submissions for each sub-task were ranked primarily based on the macro-averaged F1-score. Additionally, a unified score, **the weighted-F1**, was 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.

Participants were expected to design an experimental architecture to enhance a text classifier's temporal performance. In such, the performance of the submissions was evaluated in two ways:

1. **Macro-averaged F1-score:** This metric measured the overall F1-score on the testing set for the sentiment classification sub-task. The F1-score combines precision and recall to provide a balanced measure of model performance. A higher F1-score indicated better performance in terms of both positive and negative sentiment classification.

$$F_{\text{macro}} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

2. **Relative Performance Drop (RPD):** This metric quantified the difference in performance between the "within-period" data and the short- or long-term distant testing sets. RPD was computed as the difference in performance scores between two sets. A negative RPD value indicated a drop in performance compared to the "within-period" data, while a positive value suggested an improvement.

$$RPD = \frac{f_{\text{score}_{t_j}} - f_{\text{score}_{t_0}}}{f_{\text{score}_{t_0}}} \tag{2}$$

   Where t_0 represents performance when the time gap is 0, and t_j represents performance when the time gap is short or long, as introduced in previous work [5].

The submissions were ranked primarily based on the macro-averaged F1-score, emphasizing the overall performance of the stance detection model on the testing sets. The higher the macro-averaged F1-score, the higher the ranking of the submission.

3.4 Models

In our study, we evaluated several baseline classifiers to assess their performance and temporal persistence when exposed to evolving data. The models we focused on include **bert-base-uncased**, **roberta-base**, and their respective variations with additional continual incremental pretraining from the climate change corpus.

To address the challenges posed by evolving data, we implemented continual incremental pretraining for both **bert-base-uncased** and **roberta-base** models. These variations, referred to as *++MLM 2019*, were further pretrained on a climate change corpus that covers data from the initial training year up to 2019 using masked language modeling. This approach aimed to incorporate recent linguistic trends and contextual information, enhancing the models' ability to adapt to new and evolving data.

The dataset is segmented by years, starting from 2006 to various end years (2011, 2013, 2015, 2017, 2019). For each end year, data from all preceding years up to that point is aggregated and preprocessed. Preprocessing includes filling missing values with the most frequent value in each column, removing rows with missing values in the 'text' or 'stance' columns, and eliminating duplicate entries. Text data is normalized to lowercase, and entries with fewer than six words are excluded. Post-processing, the data is merged into a single dataset for each end year, resulting in five datasets representing different temporal spans. These datasets are subsequently balanced by downsampling to ensure uniform representation for incremental training.

Using a masked language modeling strategy, the textual data without its label is fed into the models incrementally in their chronological order, starting with the 2011 sample and ending with the 2019 sample. This approach ensures a balanced and clean dataset, facilitating robust analysis and model training. Each model was incrementally tested to evaluate its persistence over time, and the best performance was reported in the results section.

- **bert-base-uncased** (Bidirectional Encoder Representations from Transformers) [10] is a foundational model in NLP that introduced the concept of bidirectional training of transformers for language modeling. The bert-base-uncased model is a version of BERT that ignores case sensitivity, which helps in learning case-independent features. It also consists of 12 transformer layers, 768 hidden units, and 12 attention heads. BERT uses a static masked language modeling objective during pretraining, which involves predicting masked words in a sentence based on their context.
- **roberta-base** (Robustly optimized BERT approach) [21] is a variant of the BERT model designed to improve performance by optimizing the pretraining process. It uses dynamic masking, a larger batch size, and more data to enhance the training of transformer-based models. The roberta-base model consists of 12 transformer layers, 768 hidden units, and 12 attention heads. It is pretrained on a diverse range of data to capture rich contextual representations, making it effective for various NLP tasks.
- *++MLM 2019*: A masked language modeling strategy used to adapt a language model to new data by incrementally pretraining with an unlabeled corpus up to 2019. This method leverages recent linguistic trends and contextual updates to improve model adaptation and performance over time.

This systematic approach allowed us to evaluate and enhance the models' temporal persistence and robustness baselines, ensuring they remain effective in the face of evolving language patterns.

3.5 Results

This section presents the results obtained during both the practice and evaluation phases of task 2.

3.6 Practice phase

In this subsection, we present the results of the practice phase of task 2. This practice dataset was provided to participants to allow them to practice and initiate their text classifiers. Since we did not get any submissions and to understand the initial performance of our practice sets, we compared several baseline classifiers. The models evaluated include **roberta-base**, **bert-base-uncased**, and their respective variations with additional continual incremental pretraining from the climate change corpus from the initial year of training up to 2019 using masked lanague modeling. The results are summarized in Table 7.

Table 7: Performance of baseline models on practice data. The columns represent: **f-Within** - performance within the same time period, **f-Short** - performance over short temporal gaps, **f-Avg** - average performance across all temporal gaps, and **RPD** - relative performance drop when applied to temporally distant data.

Model	**f-Within**	**f-Short**	**f-Avg**	**RPD**
roberta-base	0.586	0.523	0.555	-10.80%
++MLM 2019	0.612	0.525	0.569	-14.36%
bert-base-uncased	0.577	0.536	0.557	-7.19%
++MLM 2019	0.586	0.542	0.564	-7.59%

As it can be seen from Table 7, the results indicate that the *++MLM 2019* variations of both **roberta-base** and **bert-base-uncased** demonstrate improved f-Within and f-Avg scores compared to their original counterparts. This suggests that additional continual pretraining based on recent data, incrementally over time, contributes to better performance persistence. Notably, **bert-base-uncased** *++MLM 2019* achieved the lowest RPD, highlighting its resilience to temporal changes.

3.7 Evaluation phase

In this subsection, we present the results of the evaluation phase of task 2. Using the development dataset provided to participants, we evaluated the final performance of the text classifier models. To understand the performance of our development sets, we compared several baseline classifiers due to the lack of submissions. The models evaluated include **roberta-base**, **bert-base-uncased**, and their respective variations with additional continual incremental pretraining from the climate change corpus up to 2019 using masked language modeling. The results are summarized in Table 8.

Table 8: Performance of baseline models on development sets. The columns represent: **f-Within** - performance within the same time period, **f-Short** - performance over short temporal gaps, **f-Long** - performance over long temporal gaps, **f-Avg** - average performance across all temporal gaps, **RPD-Short** - relative performance drop over short temporal gaps, **RPD-Long** - relative performance drop over long temporal gaps, and **RPD-Avg** - average relative performance drop.

Model	**f-Within**	**f-Short**	**f-Long**	**f-Avg**	**RPD-Short**	**RPD-Long**	**RPD-Avg**
roberta-base	0.626	0.558	0.529	0.571	-10.81%	-15.46%	-26.26%
++MLM 2019	0.623	0.594	0.552	0.590	-4.74%	-11.46%	-16.20%
bert-base-uncased	0.614	0.569	0.536	0.573	-7.26%	-12.64%	-19.89%
++MLM 2019	0.600	0.571	0.540	0.570	-4.94%	-10.01%	-14.94%

As shown in Table 8, the *++MLM 2019* variations of both **roberta-base** and **bert-base-uncased** models exhibit notable improvements in the **f-Short** and **f-Long** scores, as well as reduced **RPD** values compared to their standard counterparts. The *++MLM 2019* variation of **roberta-base** achieved an f-Avg score of (0.590), an improvement over the original model's score of (0.571). It also showed a significantly lower RPD-Short of (-4.74%) and RPD-Long of (-11.46%), indicating better resilience to temporal changes over both short and long gaps. Similarly, the *++MLM 2019* variation of **bert-base-uncased** achieved an **f-Avg** score of (0.570), slightly lower than the original model's 0.573. However, it exhibited a lower **RPD-Long** of (-10.01%) and **RPD-Avg** of (-14.94%), demonstrating improved performance persistence over time.

These results reinforce the value of continual incremental pretraining with recent data to maintain and improve model performance in dynamic environments. The *++MLM 2019* variations consistently showed enhanced performance metrics and reduced performance degradation over time, validating the effectiveness of this approach in enhancing temporal persistence.

3.8 Discussion and conclusion

This section discusses the results of our study on temporally adaptive classification methods, highlighting the significance of incorporating temporal information into text classification models to mitigate performance drops over time and the use of an outdated language model. These results reveal that classifiers trained on older data exhibit significant performance drops when applied to newer data. This is evident from the relative performance drops (RPD) reported, where the *++MLM 2019* variations showed a marked improvement in mitigating this drop.

Previous work by Alkhalifa et al. [5] introduced the *Incremental Temporal Alignment (ITA)* method as a superior approach for enhancing temporal persistence of static word embedding. This method aligns closely with the continual incremental pretraining approach evaluated in our results, where *++MLM 2019* variations of both **roberta-base** and **bert-base-uncased** demonstrated improved **f-Within**, **f-Avg** scores, and lower **RPD** values. The *ITA* method's emphasis on leveraging incremental updates to word embeddings aligns with the improvements seen in the *++MLM 2019* models, showcasing their resilience to evolving data and enhancing their persistence as text classifiers as context updated overtime.

The results reinforce several best practices for designing temporally robust and persistent text classifiers. Methods relying on incremental updates generally outperform static embeddings, as corroborated by the superior performance of the *++MLM 2019* models. Additionally, it is crucial to select robust baseline models and incrementally update them to accommodate evolving language patterns over time.

The practical implications of our findings are significant for real-world NLP applications. In dynamic environments such as stance posts on social media, language evolves rapidly, making temporal adaptation through an incremental pretraining approach substantially enhance the longevity and persistence of text classifiers. These results provide empirical evidence supporting the implementation of temporally adaptive classification methods in real-world scenarios.

Acknowledgement. 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. Alexander, D., Fröbe, M., Hendriksen, G., Schlatt, F., Hagen, M., Hiemstra, D., Potthast, M., de Vries, A.P.: Team openwebsearch at clef 2024: Longeval. In: Faggioli et al. [12]
2. Alkhalifa, R., Bilal, I., Borkakoty, H., Camacho-Collados, J., Deveaud, R., El-Ebshihy, A., Espinosa-Anke, L., Gonzalez-Saez, G., Galuščáková, P., Goeuriot, L., Kochkina, E., Liakata, M., Loureiro, D., Tayyar Madabushi, H., Mulhem, P., Piroi,

F., Popel, M., Servan, C., Zubiaga, A.: Longeval: Longitudinal evaluation of model performance at clef 2023. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval. Springer Nature Switzerland, Cham (2023)

3. Alkhalifa, R., Bilal, I., Borkakoty, H., Camacho-Collados, J., Deveaud, R., El-Ebshihy, A., Espinosa-Anke, L., Gonzalez-Saez, G., Galuščáková, P., Goeuriot, L., Kochkina, E., Liakata, M., Loureiro, D., Madabushi, H.T., Mulhem, P., Piroi, F., Popel, M., Servan, C., Zubiaga, A.: Overview of the clef-2023 longeval lab on longitudinal evaluation of model performance. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023). Lecture Notes in Computer Science (LNCS), Springer, Thessaliniki, Greece (18-23 September 2023)
4. Alkhalifa, R., Bilal, I.M., Borkakoty, H., Romain, Deveaud, El-Ebshihy, 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`
5. 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)
6. Alkhalifa, R., Yoong, T., Kochkina, E., Zubiaga, A., Liakata, M.: QMUL-SDS at checkthat! 2020: Determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions. CoRR **abs/2008.13160** (2020), `https://arxiv.org/abs/2008.13160`
7. Alkhalifa, R., Zubiaga, A.: Capturing stance dynamics in social media: open challenges and research directions. International Journal of Digital Humanities pp. 1–21 (2022)
8. Basaglia, A., Stocco, A., Popović, M., Ferro, N.: Seupd@clef: Team dam on reranking using sentence embedders. In: Faggioli et al. [12]
9. Cazzador, L., Faveri, F.L.D., Franceschini, F., Pamio, L., Piron, S., Ferro, N.: Seupd@clef: Team mouse on enhancing search engines effectiveness with large language models. In: Faggioli et al. [12]
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
11. 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`
12. Faggioli, G., Ferro, N., Galuščáková, P., de Herrera, A.G.S. (eds.): Proceedings of Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, Aachen (2024)
13. Galli, F., Rigobello, M., Schibuola, M., Zuech, R., Ferro, N.: Seupd@clef: Team iris on temporal evolution of query expansion and rank fusion techniques applied to cross-encoder re-rankers. In: Faggioli et al. [12]
14. Galuščáková, P., Deveaud, R., Gonzalez-Saez, G., Mulhem, P., Goeuriot, L., Piroi, F., Popel, M.: Longeval-retrieval: French-english dynamic test collection for continuous web search evaluation (2023)

15. Gangi Reddy, R., Iyer, B., Sultan, M.A., Zhang, R., Sil, A., Castelli, V., Florian, R., Roukos, S.: Synthetic target domain supervision for open retrieval qa. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1793–1797. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). `https://doi.org/10.1145/3404835.3463085`, `https://doi.org/10.1145/3404835.3463085`
16. Giachelle, F., Irrera, O., Silvello, G.: Doctag: A customizable annotation tool for ground truth creation. In: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II. Lecture Notes in Computer Science, vol. 13186, pp. 288–293. Springer (2022)
17. Gründel, M., Weber, M., Franke, J., Reimer, J.H.: Team galápagos tortoise at longeval 2024: Neural re-ranking and rank fusion for temporal stability. In: Faggioli et al. [12]
18. Keller, J., Breuer, T., Schaer, P.: Leveraging prior relevance signals in web search. In: Faggioli et al. [12]
19. Kimia, A., Akan, A., Arwa, F., Ferro, N.: Seupd@clef: Team kalu on improving search engine performance with query expansion and re-ranking approach. In: Faggioli et al. [12]
20. 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`
21. 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)
22. Lovón-Melgarejo, J., Soulier, L., Pinel-Sauvagnat, K., Tamine, L.: Studying catastrophic forgetting in neural ranking models. p. 375–390. Springer-Verlag, Berlin, Heidelberg (2021). `https://doi.org/10.1007/978-3-030-72113-8_25`, `https://doi.org/10.1007/978-3-030-72113-8_25`
23. 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/`
24. Yoon, S., Kim, J., won Hwang, S.: Analyzing the effectiveness of listwise reranking with positional invariance on temporal generalizability. In: Faggioli et al. [12]

A Runs submitted to the IR Task

Table 9: The original name of the submitted runs for the IR task are shown in the second column while the Runs Ids used assigned to the systems and used in the paper are shown in the first column.

Run Id	**Submitted System**
abyss_run_1	ABYSS_BM25-French-Stop50_40FR_10EN-SnowStem-Dict-Fuzzy-Phrase-Start-Synonyms-RR
abyss_run_2	ABYSS_BM25-French-Stop50_40FR_10EN-SnowStem-Fuzzy-Phrase-Start
abyss_run_3	ABYSS_BM25-French-Stop50_40FR_10EN-SnowStem-Fuzzy-Phrase-Start-RR
cir_run_1	CIR_BM25
cir_run_2	CIR_BM25+monoT5
cir_run_3	CIR_BM25+qrel_boost
cir_run_4	CIR_BM25+RF
cir_run_5	CIR_BM25+time_boost
galapagos_run_1	galapagos-tortoise-bm25-bo1-pl2-monot5-kmax-avg-k-4
galapagos_run_2	galapagos-tortoise-bm25-bo1-pl2-monot5-max
galapagos_run_3	galapagos-tortoise-bm25-bo1-pl2-monot5-mean
galapagos_run_4	galapagos-tortoise-rank-zephyr
galapagos_run_5	galapagos-tortoise-wsum
kalu_run_1	KALU_MISTRAL_FRENCH
kalu_run_2	KALU_RERANK_HARMONIC_MISTRAL_FRENCH
kalu_run_3	KALU_RERANK_HARMONIC_MISTRAL_FRENCH_SHOULD
kalu_run_4	KALU_RERANK_SIMPLE_FRENCH_LLAMA
kalu_run_5	KALU_RERANK_SIMPLE_MISTRAL_FRENCH
ows_run_1	ows_bm25_bo1_keyqueries
ows_run_2	ows_bm25_reverted_index
ows_run_3	ows_ltr_all
ows_run_4	ows_ltr_wows_all_rerank
ows_run_5	ows_ltr_wows_base_rerank
ows_run_6	ows_ltr_wows_rerank_and_keyquery
ows_run_7	ows_ltr_wows_rerank_and_reverted_index
quokkas_run_1	Quokkas_french-letter-lightstem
quokkas_run_2	Quokkas_french-standard-lightstem
dam_run_1	seupd2324-dam_EN-Stop-SnowBall-Poss-Prox(50)
dam_run_2	seupd2324-dam_EN-Stop-SnowBall-Poss-Prox(50)-Reranking(200)
dam_run_3	seupd2324-dam_FR-Stop-FrenchLight-Elision-ICU-Prox(50)
dam_run_4	seupd2324-dam_FR-Stop-FrenchLight-Elision-ICU-Prox(50)-Reranking(150)
dam_run_5	seupd2324-dam_FR-Stop-FrenchLight-Elision-ICU-Shingles-Prox(50)-Reranking(150)
iris_run_1	seupd2324-iris_FR_GFF@12_w0.162_MMARCO@1000_ADD_w5
iris_run_2	seupd2324-iris_FR_GFF@12_w0.162_MMARCO@1000_MAXMIN_ADD_w5
iris_run_3	seupd2324-iris_FR_MMARCO@1000_ADD_w5
iris_run_4	seupd2324-iris_FR_url_w1.4_GFF@12_w0.162_MMARCO@1000_ADD_w5
iris_run_5	seupd2324-iris-FR_Q2K@1_w0.16_MMARCO@1000_MAXMIN_ADD_w5
lfzzo_run_1	seupd2324-lfzzo-englishSystem1
lfzzo_run_2	seupd2324-lfzzo-englishSystem2
lfzzo_run_3	seupd2324-lfzzo-englishSystem3
lfzzo_run_4	seupd2324-lfzzo-englishSystem4
lfzzo_run_5	seupd2324-lfzzo-englishSystem5
lfzzo_run_6	seupd2324-lfzzo-frenchSystem1
lfzzo_run_7	seupd2324-lfzzo-frenchSystem2
lfzzo_run_8	seupd2324-lfzzo-frenchSystem3
lfzzo_run_9	seupd2324-lfzzo-frenchSystem4
lfzzo_run_10	seupd2324-lfzzo-frenchSystem5
mouse_run_1	seupd2324-mouse_English_Porter_Standard_NoStop_Mixtral-8x7b_NoRerank
mouse_run_2	seupd2324-mouse_English_Porter_Standard_stopwords-en_LLama3-70b_NoRerank
mouse_run_3	seupd2324-mouse_English_Porter_Standard_top125_LLama3-70b_Cohere-100-w06
mouse_run_4	seupd2324-mouse_English_Porter_Standard_top125_LLama3-70b_Pygaggle-Luyu-20-w06
mouse_run_5	seupd2324-mouse_English_Porter_Standard_top125_Mixtral-8x7b_Pygaggle-Luyu-20-w06
mouse_run_6	seupd2324-mouse_French_FrenchLight_Standard_NoStop_Mixtral-8x7b_NoRerank
mouse_run_7	seupd2324-mouse_French_FrenchLight_Standard_stopwords-fr_LLama3-70b_NoRerank
mouse_run_8	seupd2324-mouse_French_FrenchLight_Standard_top125_LLama3-70b_Cohere-100-w06
mouse_run_9	seupd2324-mouse_French_FrenchLight_Standard_top125_LLama3-70b_Pygaggle-Luyu-20-w06
mouse_run_10	seupd2324-mouse_French_FrenchLight_Standard_top125_Mixtral-8x7b_Pygaggle-Luyu-20-w06
seekx_run_1	seupd2324-seekx_LetLightFR
seekx_run_2	seupd2324-seekx_LetLightStopFR
seekx_run_3	seupd2324-seekx_LetLightStopSynFR
seekx_run_4	seupd2324-seekx_StanMinEN
seekx_run_5	seupd2324-seekx_StanMinSynEN
snu_run_1	SNU_LDI_listt5
snu_run_2	SNU_LDI_monot5
wonder_run_1	WONDER_BASELINE
wonder_run_2	WONDER_ENGLISH
wonder_run_3	WONDER_ENGLISH_FRENCH
wonder_run_4	WONDER_FRENCH
wonder_run_5	WONDER_TWOPHASE
xplore_run_1	XPLORE_French-BM25-FrenchLight-Stop
xplore_run_2	XPLORE_French-BM25-FrenchLight-Stop-SynonymMapper
xplore_run_3	XPLORE_French-BM25Default-FrenchLight-Stop
xplore_run_4	XPLORE_French-LMDirichlet-FrenchLight-Stop