Mitigating Temporal Drift in Social Media Sentiment Classification: A Comparative Study

Tamara Suárez Rodríguez
t.suarezrodriguez@se25.qmul.ac.uk
Student Number: 250160477
Project Supervisor: Professor Arkaitz Zubiaga
MSc. Artificial Intelligence
Queen Mary University of London

***Abstract*— The language used on social media changes over time, and models trained for tweet sentiment classification tend to become less accurate on later data, a phenomenon known as temporal drift. The LongEval 2023 shared task was introduced to evaluate mitigation strategies in this setting. However, no single strategy performed best across all evaluation periods, making it unclear which strategy to deploy when the target period is unknown.**

This paper compares six temporal drift mitigation strategies and a RoBERTa baseline on the LongEval benchmark to investigate why no single strategy consistently performs best across evaluation periods. A per-tweet analysis shows that the strategies rely on different mechanisms. On the earlier periods, much of the improvement comes from shifting the classifier's decision threshold, whereas on the furthest period the gains reflect changes in which tweets are classified correctly. The analysis also shows that the development data available before deployment is not sufficient to determine either the optimal decision threshold or the strongest strategy.

Building on these findings, this paper proposes PretrainedTEA, which combines continued encoder pretraining with recency-weighted temporal expert averaging. Although its macro-F1 is statistically indistinguishable from the strongest single-mechanism systems, it ranks first on a four-measure robustness composite computed for each of the three test periods. Overall, when the target period is unknown, the results support looking beyond macro-F1 and considering how consistently a strategy performs across deployment conditions.

***Index Terms*—temporal drift, sentiment classification, social media, temporal adaptation, model robustness**

I. Introduction

Sentiment classification of social media posts has become a standard way of measuring public opinion on elections, public health and consumer products. The standard pipeline trains a classifier on a labelled corpus from a fixed period and then applies it to later data. The language of those posts, however, evolves as new words appear, topics change and the sentiment of familiar words shifts. As a result, a classifier trained on one year of tweets loses accuracy on tweets from later years [Alkhalifa et al., 2023b], an effect known as temporal drift. It has been observed across a range of NLP tasks, including sentiment, topic and entity classification [Ushio and Camacho-Collados, 2024], and limits how long a deployed model can be trusted without retraining.

The LongEval 2023 shared task on temporal drift in Twitter sentiment classification made the problem concrete [Alkhalifa et al., 2023a]. The results showed that the strategies that worked best on one test period were not the same as those that worked best on another, and three of the four teams also scored higher on the long 2021 period than on the short 2018 period, which is the opposite of what a decline that grows with the gap would predict.

This study addresses three research questions:

- **RQ1** How does temporal drift vary across the three test periods?
- **RQ2** Does the measured drift indicate which strategy will help?
- **RQ3** Can these findings inform the design of a mitigation approach that performs consistently across different temporal gaps?

To answer these questions, this paper evaluates six mitigation strategies and a RoBERTa baseline on the three test sets, with three random seeds per strategy. The strategies can be grouped by mechanism into four families, each described in Section III-E. **Recency weighting** emphasises the most recent years of the training data, **encoder pretraining** continues the encoder's pretraining on a broader tweet corpus, **input tagging** prepends a year tag to each tweet, and **generative classification** produces the label as a word from a text-to-text model.

At the level of individual tweets, much of the improvement on the 2016 and 2018 periods comes from correcting the position of the baseline's decision boundary rather than from handling drift. The analysis also shows that the practice sets do not predict which strategy or threshold will work best on the test periods. With no per-period winner to pick in advance, this paper proposes PretrainedTEA, a system that combines continued encoder pretraining with recency-weighted temporal expert averaging in a single trained model.

Together, the analyses reported above and the proposed method make three contributions:

- A side-by-side comparison of six mitigation strategies from the literature and a RoBERTa baseline on the LongEval 2023 Task 2 benchmark, under a common evaluation protocol with three independent seeds per strategy.

- A per-tweet analysis that decomposes the reported gains, showing that about half of the improvement on the two earlier periods can be reproduced by changing the decision threshold, while on the long period the gains come from changes in which tweets are classified correctly. It also shows that the practice sets cannot identify the best threshold or strategy for the test periods.
- PretrainedTEA, a proposed method that combines continued encoder pretraining with recency-weighted temporal expert averaging and ranks first among the eight systems on the robustness composite in all three test periods.

II. Related Work

A. Temporal drift

Concept drift is the general problem of distribution shift between training and deployment data. Formally, it is defined as a change in the joint distribution of inputs and labels between two time points [Gama et al., 2014]:

$$p_{t_0}(X, y) \neq p_{t_1}(X, y). \tag{1}$$

The version of it studied in this paper, where the shift comes from train and test being from different periods, is known as temporal drift in the literature. A useful distinction is made between two types of drift [Gama et al., 2014]:

- *Real drift*[1]: a change in the relation between input and label $P(y \mid X)$, for instance known words whose sentiment shifts over time.
- *Virtual drift*: a change in the input distribution $P(X)$ alone that leaves the decision boundary unchanged, for instance new vocabulary appearing in later periods.

Based on this distinction, Section III-D defines the three aspects of drift examined in this paper: lexical drift and contextual semantic shift, both changes in $P(X)$, and real drift as a change in $P(y \mid X)$.

Several studies have shown that the degradation caused by drift resists general model improvements. First, scaling does not help: Lazaridou et al. [2021] find that scaling up Transformer-XL does not reduce the drop between training and a test set from two years later. Second, switching architecture offers little improvement: Alkhalifa et al. [2023b] compare many models and architectures across longitudinal datasets and find the drop in all cases, with the dataset mattering more than the model choice. And most recently, the pattern extends to modern large language models, which show the same loss when evaluated on text from after their training cutoff [Zhu et al., 2025].

Since neither scale nor architecture removes the problem, a separate body of work has developed strategies that target drift directly [Gama et al., 2014], and the question becomes which type of drift each strategy addresses and under what conditions it helps.

Social media data is particularly susceptible to this kind of drift because vocabulary and topics change quickly. News and events introduce new entities and slang (*COVID-19* and *Brexit* being recent examples), conversational patterns evolve, and the population of users itself is not stable. Twitter has been a common domain of study, with prior work asking whether pretraining on newer data closes the gap [Ushio and Camacho-Collados, 2024].

[1]The label *real* is conventional and does not imply that virtual drift is any less of a drift.

B. The LongEval task

This paper builds on the LongEval 2023 shared task at the Conference and Labs of the Evaluation Forum (CLEF) [Alkhalifa et al., 2023a]. LongEval was launched to evaluate how machine learning systems hold up when training and test data come from different time periods, and was organised into two tasks: Task 1 on information retrieval over Web search corpora, and Task 2 on binary sentiment classification of tweets, which is the one analysed in this paper. Its dataset comes from a sample of the TM-Senti corpus [Yin et al., 2021], with training data from 2014–2016 and three test sets from 2016, 2018 and 2021. Four teams participated, evaluating systems against an official RoBERTa baseline [Liu et al., 2019].

The results reveal a pattern that motivates the present work. The four participating teams disagree on which test period yields the highest scores. The official baseline scores highest on the 2016 and 2018 periods, and three of the four teams score higher on the furthest 2021 period than on the closer 2018 one, contrary to what a progressive drop with the time gap would predict.

C. Mitigation strategies

A range of strategies have been proposed to address temporal drift in text classifiers. Lightweight approaches incorporate temporal information directly into the input by prepending a date tag to each input instance [Ninalga, 2023]. Other approaches adapt the encoder by adopting more recent pretrained models, such as TimeLMs [Loureiro et al., 2022], or through continued pre-training on temporally extended corpora [Gururangan et al., 2020]. Alternatively, some methods leave the encoder unchanged and focus on the optimisation process, including assigning higher weights to more recent data points [Gama et al., 2014] or combining multiple year-specific experts [Liu et al., 2025].

More recently, generative text-to-text models have offered a different way of addressing the problem. Rather than adapting individual components of a discriminative classifier, these approaches reformulate classification as a text generation task, where the model is fine-tuned to generate the class label as text [Raffel et al., 2020, Medina-Alias and Şimşek, 2023].

Two of the four teams in the LongEval 2023 task published a description of their system, and both are among the strategies described above [Alkhalifa et al., 2023a]. The first-placed team used a generative text-to-text model [Medina-Alias and Şimşek, 2023], and the second-placed team used date-prefixed input [Ninalga, 2023].

This paper compares these strategies side by side on the same dataset and against the same baseline, which makes the results directly comparable across strategies.

TABLE I
THE DATASET SPLITS.

Split	Period	Size	Labels
Train	2014–2016	49,608	distant
Practice (2016)	2016	1,344	distant
Practice (2018)	2018	1,344	distant
Test (within)	2016	908	manual
Test (short)	2018	908	manual
Test (long)	2021	908	manual

D. Class-prior and threshold adaptation

A separate line of work examines how classifiers can be adapted when the balance between positive and negative examples changes between training and deployment. These methods adjust the predicted probabilities of a fixed classifier rather than retraining the model. Alexandari et al. [2020] combine maximum-likelihood adjustment with bias-corrected calibration, Garg et al. [2023] combine data rebalancing with an adjustment based on the estimated target-label distribution, and Esuli et al. [2021] revisit a related probability-adjustment method and find that it works best for well-calibrated classifiers with few classes.

These findings raise the question of whether some of the gains attributed to temporal mitigation may instead come from the position of the decision threshold.

III. METHODOLOGY

A. Dataset

The experiments use the LongEval 2023 Task 2 dataset [Alkhalifa et al., 2023a], a sampled and re-annotated subset of the TM-Senti corpus [Yin et al., 2021], where the task is binary, with positive and negative labels. The data was preprocessed by removing retweets and replies, dropping posts containing URLs, excluding posts shorter than five words or longer than 140 characters, anonymising user mentions as `@MENTION`, and stripping emojis and emoticons before sampling.

The data is divided by time into six splits, summarised in Table I. The training set covers February 2014 to December 2016, two practice sets cover 2016 and 2018, and three test sets cover 2016, 2018 and 2021. The training and practice sets are labelled by distant supervision based on emojis, whereas the test sets are balanced between the two classes and manually annotated through Amazon Mechanical Turk, with five workers labelling each tweet and the final label taken as the mode of their annotations.

In addition to the labelled splits, LongEval also distributes an unlabelled tweet corpus covering 2013 to 2021, with 120,000 tweets per year.

B. Evaluation metrics

The primary metric is the **macro-averaged F1 score**, which weights both classes equally,

$$\mathrm{F1}_{\text{macro}} = \tfrac{1}{2}\left(\mathrm{F1}_{\text{pos}} + \mathrm{F1}_{\text{neg}}\right), \tag{2}$$

where each class F1 is the harmonic mean of its precision and recall. Macro-F1 is the metric used to rank the participating systems in the LongEval 2023 task [Alkhalifa et al., 2023a], which keeps the results directly comparable with them.

For each system, predictions are obtained by averaging the positive-class probabilities across the three seeds and applying a threshold of 0.5.

To examine whether strategy gains vary with how different a tweet is from the training data, we calculate a **drift score** using two signals. It combines the proportion of its words not seen in training with its distance from the training average in MiniLM embedding space. For a test tweet t, the score is defined as

$$\begin{aligned}\mathrm{drift}(t) = \mathrm{OOV}(t) + \\ \left[1 - \cos\big(\mathrm{MiniLM}(t),\ \overline{\mathrm{MiniLM}(\mathrm{train})}\big)\right],\end{aligned} \tag{3}$$

Here, $\mathrm{OOV}(t)$ is the share of words in t not seen in training. MiniLM (`all-MiniLM-L6-v2`), a pretrained sentence encoder, represents each tweet as a vector, and the second term measures the cosine distance between the test tweet and the average vector of all training tweets. A higher score means that the tweet contains more unseen vocabulary, is represented further from the training data in embedding space, or both. The score is used to rank tweets within each test set and divide them into drift quintiles, not to quantify an absolute amount of drift.

The analysis also considers three aspects of robustness: ranking quality independently of the decision threshold, stability across training seeds, and performance on the most drifted tweets as identified by the drift score. Section IV-G evaluates these aspects using the following measures:

AUROC measures the probability that a randomly selected positive tweet receives a higher predicted positive-class score than a randomly selected negative tweet [Fawcett, 2006].

Seed disagreement is the share of tweets on which the three trained seeds do not all predict the same label [Dodge et al., 2020].

Macro-F1 on the highest-drift quintile is computed on the twenty per cent of tweets with the largest drift scores.

C. Baseline

A fine-tuned RoBERTa model [Liu et al., 2019] serves as the reference against which every mitigation strategy is compared. This baseline follows the official LongEval setup: a binary classification head trained with a learning rate of 2×10^{-5}, a batch size of 32, weight decay of 0.01, a warmup ratio of 0.1, and a maximum of 25 epochs.

The only change from the official LongEval baseline is the early stopping rule. The official baseline stops on training loss. However, Medina-Alias and Şimşek [2023] report that additional fine-tuning steps degrade the long-period score, which they attribute to catastrophic forgetting. Following their observation, this paper stops on the loss of the 2016 practice set, which comes from the same period as the training data.

D. Drift characterisation

To capture different ways in which the data may change over time, we analyse vocabulary, word usage and the relationship

between words and sentiment labels in the training and test data.

Lexical drift is measured by two related quantities, the OOV rate, which is the share of test word occurrences the model never saw during training, and the new-type rate, which is the same idea but counts each distinct word only once.

Contextual semantic shift is measured at the per-word level, following the idea of Hamilton et al. [2016] that if the contexts in which a word appears change over time, its usage has shifted. Every tweet is encoded with the same MiniLM sentence encoder. For each target word, the average of the MiniLM vectors of all tweets containing it in a given year is its prototype for that year, and the semantic shift between two years is the cosine distance between the two prototypes. To account for the variation that comes from sampling rather than from drift, a noise floor is estimated by splitting the 2016 sample into two random halves and measuring the per-word shift between them, and shifts that exceed this floor are treated as genuine.

Real drift is examined at the word level. Using within as the reference, we compare it separately with short and long. For each word that occurs at least fifteen times in both test sets, we use a two-sided z-test to compare the share of tweets labelled positive. We then apply a Bonferroni correction to adjust the significance threshold for the number of words tested. This tests whether the relationship between words and sentiment labels changes across test sets.

E. Mitigation strategies

This paper evaluates six mitigation strategies for temporal drift, all taken from the existing literature. Five of them modify a single aspect of the baseline (its input format, its backbone, its loss, or its training procedure) while keeping every other setting unchanged. The sixth replaces the classifier with a generative model and is therefore trained under its own configuration. Every strategy is trained from three independent random seeds, and the strategies are grouped below by the approach they take.

Recency weighting. This group contains two strategies, RecencyWeight [Gama et al., 2014] and TEA [Liu et al., 2025]. Both emphasise the most recent years of the training data. *RecencyWeight* weights training examples by year in the loss, with higher weights on more recent years. *TEA* (temporal experts averaging) trains a separate classifier on each year of the training data, then averages their weights, with the more recent experts contributing more to the average.[2]

Encoder pretraining. This group contains two strategies, MLMPretrain [Gururangan et al., 2020] and TimeLMs [Loureiro et al., 2022]. Both pre-condition the encoder on a broader corpus that includes words seen after 2016. *MLMPretrain* runs the masked-language-model objective on the LongEval unlabelled corpus, which contains over one million tweets from 2013 to 2021. *TimeLMs* replaces the general RoBERTa backbone with a Twitter-specific RoBERTa model (`cardiffnlp/twitter-roberta-base-2021-124M`), pre-trained on tweets from 2018 to 2021.

Input tagging. A single strategy, *DatePrefix* [Ninalga, 2023], prepends a tag of the form `year: YYYY` to each tweet, with no other changes to the backbone, loss or training procedure. This is the second-placed LongEval 2023 entry and is the only strategy that acts on the input alone.

Generative classification. A single strategy, *T5* [Medina-Alias and Şimşek, 2023], treats sentiment classification as a text-to-text task. In place of a classification head on RoBERTa, a T5 model [Raffel et al., 2020] is fine-tuned to generate the label word, and the predicted class is the one it assigns the higher generation likelihood. This is the first-placed LongEval 2023 entry, and the only strategy that replaces the underlying model rather than adjusting a component of it.

F. The proposed method: PretrainedTEA

Building on these strategies, this paper proposes PretrainedTEA, which combines MLMPretrain's continued pretraining with TEA's recency-weighted expert averaging. The two components act at different stages: continued pretraining adapts the encoder's representations, while recency weighting determines how much each yearly expert contributes.

We train one expert per year (2014, 2015 and 2016) from the continued-pretrained encoder and average their parameters using TEA's fixed recency weights of 0.5, 0.75 and 1.0.

G. Per-tweet analysis

Alongside the macro-F1 numbers per period, this paper keeps a per-tweet record for every strategy. For each of the 2,724 test tweets, the record stores the strategy's predicted class and its mean positive-class probability across the three seeds, along with the tweet's OOV fraction, its semantic distance to the training distribution (Section III-B), and a polarity-mismatch flag that marks tweets whose surface polarity (the sentiment suggested by the words taken literally) contradicts their gold label. From the stored predictions, the record also marks the always-fail set, the tweets that every system misclassifies under all three seeds.

To compute the polarity-mismatch flag, we count each tweet's positive and negative words using the opinion lexicon of Hu and Liu [2004]. The flag identifies tweets whose lexicon-based surface polarity differs from the gold label. This may indicate non-literal wording, such as sarcasm or a complaint softened by a joke. This interpretation is a simplification, as the flag may also reflect limitations of the lexicon.

Code, checkpoints, data access and the interactive analysis interface are documented in Appendix A.

IV. Results

A. Each period drifts in a different way

Drift metrics for the three test sets appear in Table II. The two lexical metrics, the OOV rate and the new-type rate, both grow monotonically with the gap from the training period,

[2]This is a simplified version of TEA. The original method uses more elaborate weighting schemes, while this paper uses fixed recency weights.

TABLE II
LEXICAL AND CONTEXTUAL DRIFT ACROSS THE TEST PERIODS.

Metric	within	short	long
OOV rate (%)	3.47	4.39	5.20
New types (%)	14.14	16.50	19.27
Per-word semantic shift (mean)	—	0.033	0.045
Share above noise floor (%)	—	1.3	3.1

making lexical drift a predictable function of the time gap. On the long period these new types include words absent from the 2014–2016 training vocabulary such as *covid*, *lockdown*, *tiktok* and *airpods*, tied to events and products that postdate the training period.

Contextual semantic shift behaves differently. Its average value is small on both later periods and stays below the 2016 self-split noise floor, indicating that words are on average used in similar contexts to before. What does change is the share of words whose shift clearly exceeds the noise floor, which grows from 1.3 to 3.1 per cent between short and long. These observations show that some contextual shift is present in the corpus, but it is concentrated in a small group of words rather than spread across the whole vocabulary, and the size of that group grows with the time gap.

The word-level analysis finds no statistically reliable evidence of real drift. The initial tests identify a possible change in sentiment association for 3 of 46 words in the within–short comparison and 5 of 47 words in the within–long comparison. However, after the Bonferroni correction, none of these differences provides strong enough evidence of a change.

Taken together, the drift measures document changes in vocabulary and contexts of use, while the word-level analysis does not find reliable evidence that the relationship between words and sentiment has changed.

B. No single strategy leads on all three periods

The macro-F1 results in Table III show that no single strategy leads on all three periods. T5 achieves the top macro-F1 on within (0.748), MLMPretrain and PretrainedTEA share the top on short (0.694 each), and PretrainedTEA leads on long (0.717), ahead of TEA (0.713) and RecencyWeight (0.711).

Every system scores lower on short than on long, despite the additional three-year gap between long and the training period. The official LongEval 2023 evaluation showed the same pattern, with three of the four participating teams also performing better on long than on short.

Part of what makes short harder appears to be its composition rather than its distance from training. The polarity-mismatch flag, which marks tweets whose words carry a sentiment opposite to their gold label, is highest on short at 17 per cent, against 15 on long and 11 on within, and as the next subsection shows, the same trend also appears in the always-fail set.

C. A shared residual that drift cannot explain

Across all three random seeds, all eight evaluated systems misclassify a common subset of tweets in each test set. We refer to this shared residual as the *always-fail* set. It covers 10.9 per cent of within, 15.7 per cent of short and 14.0 per cent of long.

TABLE III
MACRO-F1 BY STRATEGY.

Method	within	short	long
Baseline (val-loss)	0.713	0.671	0.688
Recency weighting			
RecencyWeight	0.734	0.679	0.711
TEA	0.736	0.687	0.713
Encoder pretraining			
MLMPretrain	0.718	**0.694**	0.699
TimeLMs	0.726	0.683	0.695
Input tagging			
DatePrefix	0.724	0.675	0.691
Generative classification			
T5	**0.748**	0.681	0.692
Proposed method			
PretrainedTEA	0.735	**0.694**	**0.717**

Two thirds of these tweets carry a positive gold label, while each full test set contains equal numbers of positive and negative examples. The baseline's mean confidence on these tweets is almost identical to that on the tweets that every strategy classifies correctly (0.884 and 0.886, respectively), suggesting that the model is just as confident on these errors. The always-fail set also contains about twice the rate of polarity-mismatch tweets found in the test sets overall.

The following examples illustrate the two directions of the polarity mismatch:

- *"wait people ACTUALLY have tweet notifications turned on for me omg pls i m crying"* carries a positive gold label but a baseline positive probability of only 0.05. Taken literally, "crying" suggests sadness, although the tweet expresses a positive reaction to the number of people following the writer.
- *"I now have another platform to spend hours on avoiding my actual responsibilities and physical wellbeing love that for me"* carries a negative gold label but a baseline positive probability of 0.92, its surface enthusiasm masking a sarcastic complaint.

Further examples from the shared residual are provided in Appendix C-A.

D. Drift does not predict the best strategy

Given that the strategies win on different periods and each period drifts in its own way, we test whether a tweet's drift score (Section III-B) is related to which strategy works best for it. With only three periods, this can only be checked at the tweet level by comparing strategy gains across drift quintiles within each test set, focusing on the highest-drift quintile, where a drift-mitigation effect should be most visible.

The highest-drift quintile favours a different strategy in each period: DatePrefix on within ($+2.66$ macro-F1 points over the baseline), MLMPretrain on short ($+6.04$) and RecencyWeight

on long (+2.17). However, the bootstrap intervals do not identify a clear winner in any period. On within and long, no strategy clearly improves over the baseline. On short, four strategies do, but the highest-scoring strategy, MLMPretrain, is not clearly better than most of the alternatives. Overall, these results show that some strategies can improve performance on high-drift tweets, but the drift score does not consistently identify which strategy will work best.

Since a tweet's drift does not point to a strategy, and no single strategy achieves the best macro-F1 across the three periods, this paper proposes a single model for settings where the target period is unknown. Section IV-F evaluates whether it remains competitive across all three periods and across multiple robustness measures.

E. Boundary placement, and why the winner cannot be picked in advance

A mitigation strategy's gain over the baseline may lie partly in better predictions on individual tweets and partly in a shift in the decision threshold that turns those predictions into class labels. This separation reveals how much of each strategy's gain is boundary placement rather than a change in the per-tweet predictions.

To measure this, on each test period we try every decision threshold from 0.30 to 0.70 in steps of 0.01 on the baseline's positive-class probability, and keep the one that maximises macro-F1. We then compare that maximum to the F1 of the best mitigation strategy on the same period, which gives the share of the strategy's gain that boundary movement alone can explain. We leave T5 out of this comparison, as it generates its label directly rather than thresholding the baseline's probability.

Figure 1 shows that the per-period optimal threshold recovers 51 per cent of the top strategy's gain on within and 49 per cent on short. On long it recovers only 16 per cent, and the rest is a change in which tweets the model classifies correctly. However, this test-tuned threshold needs the manually annotated test labels, which are not available in deployment. To test whether the same shift can be picked from data available before deployment, we repeat the sweep on the distant-labelled 2016 practice set and apply the resulting threshold to each of the three test splits.

The results show that the practice-tuned threshold does not transfer well. On long it lands below the baseline's default, and on short it costs nearly a full percentage point of F1. A threshold-tuning approach struggles here because it needs a labelled calibration set from the target period, which is exactly what is missing at deployment. The full threshold results are reported in Appendix B-E.

Even though the strategies need no deployment-time tuning, the practice set cannot pick the best one either. Ranking the eight systems by their practice-set F1 would deploy TimeLMs, which leads both practice sets but does not lead any test period, while TEA, one of the strongest systems on the test periods, ranks near the bottom of both practice sets. As a result, deploying the practice-set winner would cost between

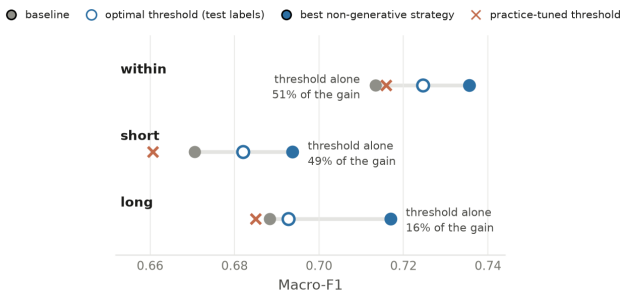


Fig. 1. Baseline macro-F1 before and after threshold adjustment, compared with the best non-generative strategy.

1.1 and 2.2 points of test macro-F1 compared to the true best system. The complete practice-set comparison is reported in Appendix B-D.

Both attempts rely on the practice set, which shares the same emoji-based labelling as the training data. One possible explanation is that a high practice score reflects agreement with that labelling more than with the manual annotation used on the test.

F. The strongest systems are closely matched

Given that neither the threshold nor the strategy can be picked for the target period in advance, the practical question is not which strategy leads in a single period, but which remains reliable when the target period is unknown. PretrainedTEA was proposed with this objective in mind.

Macro-F1 alone does not clearly separate the strongest systems on the long period, where PretrainedTEA, TEA and RecencyWeight lie within about one percentage point of one another (Table III). To understand what the combination contributes, we examine how the two mechanisms interact.

To isolate the effect of the continued-pretrained backbone and the year-expert averaging, we run the two-by-two factorial in Table IV, with TEA as the averaging-only cell and MLM-Pretrain as the backbone-only cell. Each helps on its own, the backbone most on the short period and the averaging most on long. The combination remains close to the best individual cell in every period, but adds less than the sum of the two individual gains. This negative interaction is largest on short.

The correction overlap varies across periods: the Jaccard overlap between the baseline errors corrected by MLMPretrain and TEA is 0.30 on within, 0.24 on short and 0.49 on long. Taking a closer look at the long period, where the overlap is highest, we find that the two mechanisms correct many of the same tweets rather than complementary ones. As a result, combining them adds only a small number of additional corrections: the backbone alone fixes 40 of the baseline's 281 errors, the averaging alone fixes 48, and PretrainedTEA fixes 52.

The remaining question is whether these small differences are statistically meaningful. We run a paired bootstrap with 5,000 resamples of each test set. PretrainedTEA improves over the baseline on all three periods, with every 95 per cent

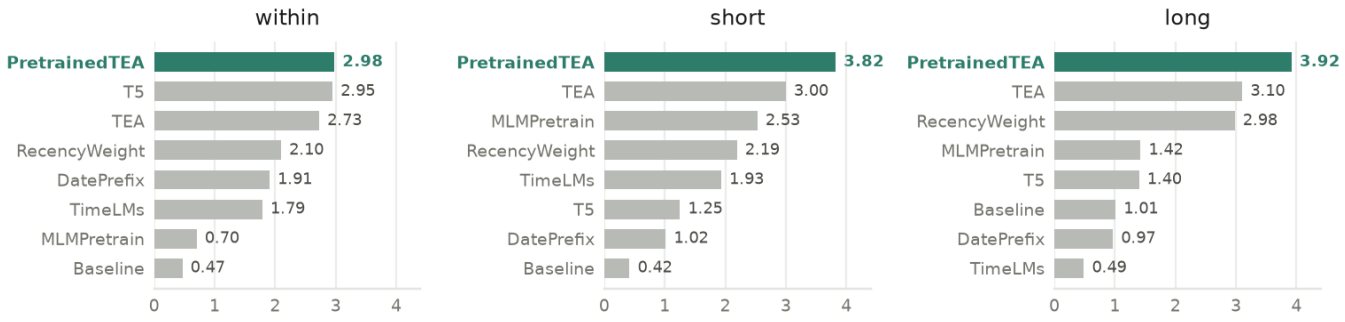


Fig. 2. Robustness composite scores for all eight systems.

TABLE IV
FACTORIAL COMPARISON OF THE TWO MECHANISMS COMBINED IN PRETRAINEDTEA. SCORES ARE MACRO-F1, EFFECTS AND INTERACTIONS ARE PERCENTAGE POINTS.

	within	short	long
Baseline	0.713	0.671	0.688
Backbone only (MLMPretrain)	0.718	0.694	0.699
Averaging only (TEA)	0.736	0.687	0.713
PretrainedTEA (both)	0.735	0.694	0.717
Backbone effect (pp)	+0.4	+2.3	+1.1
Averaging effect (pp)	+2.2	+1.7	+2.4
Interaction (pp)	−0.5	−1.7	−0.7

confidence interval excluding zero. Against MLMPretrain, the interval excludes zero on within and long but includes zero on short. Against TEA, however, all three intervals include zero, leaving PretrainedTEA and TEA statistically indistinguishable on macro-F1 in every period. Full paired-bootstrap comparisons are reported in Appendix B-A.

G. A robustness view beyond macro-F1

The robustness analysis focuses on three aspects relevant to deployment across all three test periods: ranking quality without target-period threshold tuning, stability across training seeds, and performance on the most drifted tweets. Each one targets a factor the model cannot control at deployment:

- AUROC, since the decision threshold cannot be tuned to the target period.
- Seed disagreement, since predictions should remain stable across independent training runs.
- Macro-F1 on the highest-drift quintile, since these are the tweets with the highest measured drift.

Although macro-F1 leaves the strongest systems closely matched, Figure 2 shows clearer differences when the four measures are considered together. To calculate the composite, we rescale the four measures to $[0, 1]$ across the eight systems within each period and sum them. PretrainedTEA ranks first in the robustness composite on within (2.98), short (3.82) and long (3.92). On short and long, it also ranks in the top three on all four measures, while its rankings vary more on within: third on macro-F1, second on AUROC, first on stability and tied fifth on highest-drift performance. This leaves a narrow composite lead over T5 of only 0.03 points. Full measure values appear in Appendix B-B.

In a setting where the best strategy cannot be identified in advance, this result supports considering consistency across measures alongside macro-F1.

V. CONCLUSION

This paper extends the LongEval 2023 observation that the relative performance of mitigation strategies changes across evaluation periods and shows that, across the three test periods, temporal drift does not follow a simple pattern in which a longer gap produces worse performance. It also shows that threshold adjustment can reproduce a larger share of the improvement on within and short, while improvement on long depends more on changes in which tweets are classified correctly.

The relationship between drift and temporal distance depends on the type of drift: lexical drift increases with temporal distance, while average contextual shift remains below the noise floor in both the short and long periods. For real drift, the analysis finds no statistically reliable changes in the association between individual words and sentiment labels. Despite its shorter temporal gap, short is the hardest period for every system and has the highest rate of polarity-mismatch tweets. These contrasts reflect differences in both the amount and the type of drift across the three test sets (RQ1).

After characterising the drift in each test set, we tested whether tweet-level drift could indicate which strategy would help. We compared strategy gains across drift quintiles in all three periods. The highest-drift group favours a different strategy on within, short and long. Bootstrap intervals show no reliable gain over the baseline for any strategy on within or long. Several strategies improve over the baseline on the highest-drift tweets in short, but the highest-scoring strategy is not clearly separated from most of the alternatives. These results leave no consistent mapping between measured drift and a mitigation strategy (RQ2).

The practice sets also fail to identify either a transferable threshold or the strongest test strategy. Selecting their top-

ranked strategy would reduce test macro-F1 by 1.1 to 2.2 points relative to the best system on each test set. This motivated the development of PretrainedTEA, which combines continued encoder pretraining with recency-weighted temporal expert averaging. PretrainedTEA remains within 1.3 macro-F1 points of the best system on every period and leads on long, but its differences from TEA remain within resampling noise. This small advantage on long is consistent with the two mechanisms having their highest correction overlap in that period. However, PretrainedTEA ranks first in the four-measure robustness composite on all three periods. Its contribution is not a clear F1 improvement, but a consistently competitive profile across temporal gaps and deployment conditions (RQ3). This suggests that, when the target period is unknown, a strategy designed for consistency can be a practical alternative to selecting a model on macro-F1 alone.

VI. Limitations

The scope of the evaluation limits how far the findings can be generalised. The study considers binary tweet sentiment classification on a single benchmark, with training data from 2014 to 2016 and test sets from 2016, 2018 and 2021. The patterns reported here, including the boundary-placement decomposition and the robustness ranking of PretrainedTEA, may be specific to this domain and to the years covered by the dataset. Establishing whether they also hold in other settings requires repeating the analysis on other tasks and datasets covering different years.

To compare the systems beyond macro-F1, the composite combines four measures, but its ranking depends on design choices. The measures receive equal weight and are normalised over the eight systems evaluated, and different measures, weights or comparison sets may produce a different ranking.

The analysis does not fully explain why the same tweets are missed by every system. The shared error set is present on within, where there is no temporal gap, which suggests that factors other than temporal drift are involved. Additional experiments would be needed to test the role of distant supervision and the use of different annotation methods across splits.

VII. Future work

Evaluating the same strategies on datasets from other domains and years would test whether the findings extend beyond LongEval. More temporal checkpoints, larger annotated test sets and additional training seeds would provide stronger evidence about consistency across temporal gaps and the relationship between drift and strategy performance.

The composite could be recalculated with different measures, weights and comparison systems to determine whether PretrainedTEA remains first under alternative definitions of robustness.

The factorial analysis showed that continued pretraining and temporal expert averaging often correct the same errors. This suggests another way to choose which methods to combine. With additional development periods, future work could identify methods that tend to correct different tweets and select the combination before evaluating it on unseen periods. This would test whether lower correction overlap leads to a larger combined improvement without using target labels for selection.

Re-annotating a sample of the shared residual with guidelines that allow uncertainty could separate label disagreement from genuinely difficult language. This would help determine whether these errors require temporal adaptation, better modelling of non-literal language or a different definition of the target label.

Acknowledgement

I would like to thank Professor Arkaitz Zubiaga for his guidance and valuable feedback during this project. The research leading to the results presented in this publication received funding from the Uruguayan National Agency for Research and Innovation (ANII) under grant POS_EXT_2024_1_185222.

References

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 222–232. PMLR, 2020.

Rabab Alkhalifa, Iman M. Bilal, Hsuvas Borkakoty, Jose Camacho-Collados, Romain Deveaud, Alaa El-Ebshihy, Luis Espinosa-Anke, Gabriela Gonzalez-Saez, Petra Galuščáková, Lorraine Goeuriot, Elena Kochkina, Maria Liakata, Daniel Loureiro, Philippe Mulhem, Florina Piroi, Martin Popel, Christophe Servan, Harish Tayyar Madabushi, and Arkaitz Zubiaga. Overview of the CLEF-2023 LongEval lab on longitudinal evaluation of model performance. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2023)*, volume 14163 of *Lecture Notes in Computer Science*, pages 440–458. Springer Nature Switzerland, 2023a. doi: 10.1007/978-3-031-42448-9_28.

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management*, 60(2):103200, 2023b. doi: 10.1016/j.ipm.2022.103200.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020. doi: 10.48550/arXiv.2002.06305.

Andrea Esuli, Alessio Molinari, and Fabrizio Sebastiani. A critical reassessment of the Saerens-Latinne-Decaestecker algorithm for posterior probability adjustment. *ACM Transactions on Information Systems*, 39(2): 19:1–19:34, 2021. doi: 10.1145/3433164.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, and Zachary Chase Lipton. RLSbench: Domain adaptation under relaxed label shift. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10879–10928. PMLR, 2023.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1141.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. Association for Computing Machinery, 2004. doi: 10.1145/1014052.1014073.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liška, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363. Curran Associates, Inc., 2021.

Aoming Liu, Kevin Miller, Venkatesh Saligrama, Kate Saenko, Boqing Gong, Ser-Nam Lim, and Bryan A. Plummer. Scaling up temporal domain generalization via temporal experts averaging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28208–28231. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1432.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. doi: 10.48550/arXiv.1907.11692.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa-Anke, and Jose Camacho-Collados. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-demo.25.

Pablo Medina-Alias and Özgür Şimşek. The temporal persistence of generative language models in sentiment analysis. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, volume 3497 of *CEUR Workshop Proceedings*, pages 2458–2468. CEUR-WS.org, 2023.

Dean Ninalga. Keeping in time: Adding temporal context to sentiment analysis models. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, volume 3497 of *CEUR Workshop Proceedings*, pages 2469–2475. CEUR-WS.org, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Asahi Ushio and Jose Camacho-Collados. A systematic analysis on the temporal generalization of language models in social media. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 52–62. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.wassa-1.5.

Wenjie Yin, Rabab Alkhalifa, and Arkaitz Zubiaga. The emojification of sentiment on social media: Collection and analysis of a longitudinal Twitter sentiment dataset. *arXiv preprint arXiv:2108.13898*, 2021. doi: 10.48550/arXiv.2108.13898.

Chenghao Zhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang. Is your LLM outdated? a deep look at temporal generalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7433–7457. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.381.

APPENDIX A
REPRODUCIBILITY

A. Code, checkpoints and data

The code, reproduction pipeline and instructions to regenerate every table and figure are available at https://github.com/tamarasuarezrod/temporal-drift-sentiment-msc. The trained checkpoints for the baseline and each strategy are archived on Zenodo at https://doi.org/10.5281/zenodo.21850584. Every system is trained on the 2014 to 2016 split and evaluated on the 2016, 2018 and 2021 test sets, with three random seeds (42, 1 and 2) per system. Reported macro-F1 results are computed after averaging the positive-class probabilities across the three seeds and applying a threshold of 0.5.

The experiments use the LongEval 2023 Task 2 data, available at https://clef-longeval-2023.github.io/data/.

B. Interactive analysis interface

An interactive dashboard was developed to support exploration of the results and is available at https://temporal-drift-sentiment-dashboard.vercel.app/. The dashboard has two sections, a tweet explorer for inspecting individual predictions and filtering by test set, drift, annotation flags and system disagreement, and an experiments section that presents the main analyses interactively. Access is password-protected because the explorer contains the complete test sets.

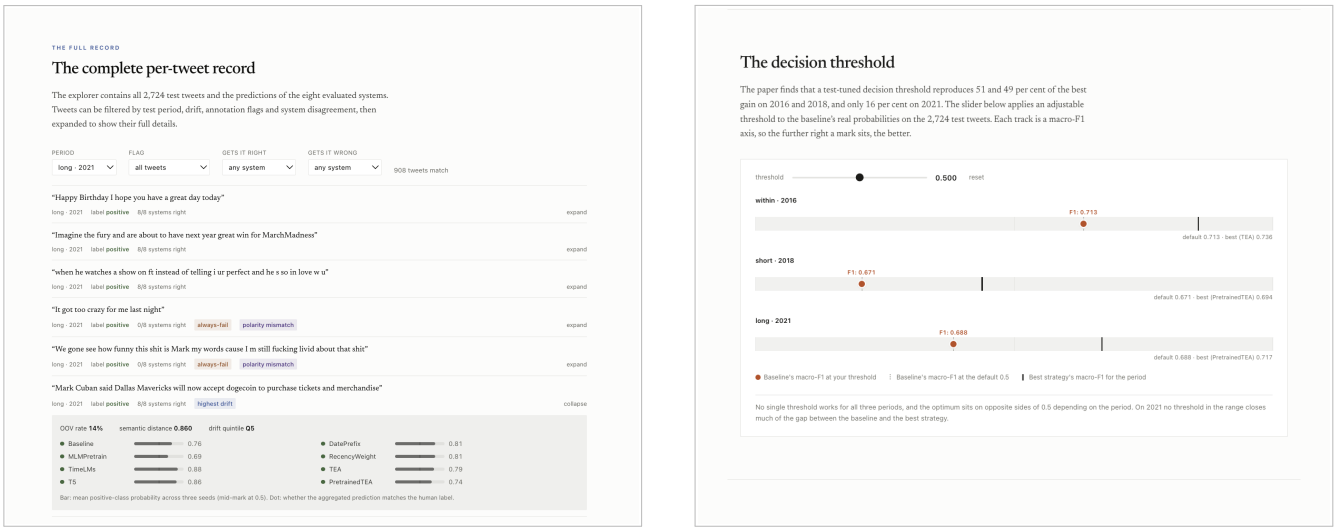


Fig. 3. Dashboard used during the analysis. Left: per-tweet explorer. Right: one of the interactive experiments, showing the decision-threshold analysis.

APPENDIX B
SUPPLEMENTARY EVALUATION RESULTS

A. Statistical uncertainty

Table V gives paired bootstrap differences from 5,000 resamples. All comparisons with the baseline have 95 per cent confidence intervals that exclude zero. However, several intervals comparing the top strategies include zero. On long, PretrainedTEA improves over the baseline and MLMPretrain, but its difference from TEA also includes zero.

TABLE V
PAIRED BOOTSTRAP DIFFERENCES IN MACRO-F1, IN PERCENTAGE POINTS, WITH 95% CONFIDENCE INTERVALS.

Comparison	split	diff (pp)	95% CI
TEA vs Baseline	within	+2.20	[+0.58, +3.89]
TEA vs RecencyWeight	within	+0.14	[-1.46, +1.72]
MLMPretrain vs Baseline	short	+2.32	[+0.77, +3.97]
MLMPretrain vs TEA	short	+0.64	[-1.11, +2.40]
TEA vs Baseline	long	+2.44	[+0.59, +4.37]
TEA vs RecencyWeight	long	+0.13	[-1.56, +1.84]
PretrainedTEA vs baseline	within	+2.19	[+0.20, +4.20]
PretrainedTEA vs MLMPretrain	within	+1.75	[+0.20, +3.30]
PretrainedTEA vs TEA	within	-0.03	[-1.60, +1.50]
PretrainedTEA vs baseline	short	+2.31	[+0.60, +4.00]
PretrainedTEA vs MLMPretrain	short	+0.00	[-1.30, +1.30]
PretrainedTEA vs TEA	short	+0.67	[-0.90, +2.30]
PretrainedTEA vs baseline	long	+2.86	[+1.00, +4.90]
PretrainedTEA vs MLMPretrain	long	+1.76	[+0.30, +3.20]
PretrainedTEA vs TEA	long	+0.43	[-1.20, +2.10]

B. Robustness measures

Table VI reports the four inputs to the robustness composite for each period.

TABLE VI
ROBUSTNESS MEASURES AND COMPOSITES FOR ALL SYSTEMS AND TEST PERIODS.

Period	System	Macro-F1	AUROC	Seed disag. (%)	Q5 gain (pp)	Composite
within	Baseline	0.713	0.791	13.8	0.00	0.47
	RecencyWeight	0.734	0.804	14.1	+1.63	2.10
	TEA	0.736	0.812	10.5	+1.63	2.73
	MLMPretrain	0.718	0.804	19.1	+0.36	0.70
	TimeLMs	0.726	0.798	14.0	+1.92	1.79
	DatePrefix	0.724	0.801	16.1	**+2.66**	1.91
	T5	**0.748**	**0.820**	14.3	+1.41	2.95
	PretrainedTEA	0.735	0.815	**7.7**	+1.41	**2.98**
short	Baseline	0.671	0.730	14.0	0.00	0.42
	RecencyWeight	0.679	0.737	13.0	+3.32	2.19
	TEA	0.687	0.738	9.5	+3.86	3.00
	MLMPretrain	**0.694**	0.734	17.4	**+6.04**	2.53
	TimeLMs	0.683	0.735	12.6	+2.21	1.93
	DatePrefix	0.675	0.731	16.1	+2.71	1.02
	T5	0.681	0.729	13.7	+2.74	1.25
	PretrainedTEA	**0.694**	**0.739**	**7.1**	+4.95	**3.82**
long	Baseline	0.688	0.741	14.9	0.00	1.01
	RecencyWeight	0.711	0.754	15.1	**+2.17**	2.98
	TEA	0.713	0.757	10.5	+0.32	3.10
	MLMPretrain	0.699	0.749	19.1	0.00	1.42
	TimeLMs	0.695	0.738	16.0	-1.98	0.49
	DatePrefix	0.691	0.742	18.6	+0.50	0.97
	T5	0.692	0.749	14.2	-0.86	1.40
	PretrainedTEA	**0.717**	**0.757**	**7.1**	+1.82	**3.92**

C. Highest-drift uncertainty

Table VII reports paired-bootstrap intervals from 5,000 resamples of the highest-drift quintile in each test period. Each cell compares a strategy with the baseline in macro-F1 percentage points.

TABLE VII
MACRO-F1 DIFFERENCES FROM THE BASELINE ON THE HIGHEST-DRIFT QUINTILE, WITH 95% PAIRED-BOOTSTRAP CONFIDENCE INTERVALS.

Strategy	within	short	long
DatePrefix	+2.66 [−0.86, +6.52]	+2.71 [−0.46, +6.05]	+0.50 [−3.54, +4.68]
MLMPretrain	+0.36 [−3.70, +4.63]	+6.04 [+2.75, +9.83]	+0.00 [−4.54, +4.56]
RecencyWeight	+1.63 [−2.15, +5.74]	+3.32 [+0.07, +6.70]	+2.17 [−1.89, +6.45]
TimeLMs	+1.92 [−2.18, +6.32]	+2.21 [−1.62, +6.06]	−1.98 [−8.37, +4.34]
TEA	+1.63 [−1.93, +5.55]	+3.86 [+1.03, +7.17]	+0.32 [−4.40, +5.27]
PretrainedTEA	+1.41 [−2.95, +6.12]	+4.95 [+0.62, +9.37]	+1.82 [−3.42, +7.14]
T5	+1.41 [−2.61, +5.54]	+2.74 [−1.04, +6.73]	−0.86 [−5.67, +4.14]

D. Model selection

Table VIII shows each system's macro-F1 on the two practice sets against its scores on the three test periods. The practice sets do not identify the test winners, meaning that choosing their highest-scoring strategy would not select the best system on any test set.

TABLE VIII
MACRO-F1 ON THE PRACTICE AND TEST SETS.

System	prac. 2016	prac. 2018	within	short	long
Baseline	0.794	0.777	0.713	0.671	0.688
DatePrefix	0.797	0.781	0.724	0.675	0.691
MLMPretrain	0.808	0.788	0.718	**0.694**	0.699
RecencyWeight	0.810	0.789	0.734	0.679	0.711
TimeLMs	**0.826**	**0.795**	0.726	0.683	0.695
TEA	0.792	0.769	0.736	0.687	0.713
T5	0.801	0.773	**0.748**	0.681	0.692
PretrainedTEA	0.802	0.769	0.735	**0.694**	**0.717**

E. Threshold analysis

Table IX shows how much of the best non-generative strategy's improvement over the baseline can be reproduced by changing only the baseline's decision threshold. The *in-sample optimal* threshold is selected using the labels from each test set and serves as a diagnostic comparison rather than a deployable method. The *practice-tuned* threshold is selected using the 2016 practice set and then applied to the test sets. The "recovered" columns report the percentage of the best strategy's improvement reproduced by each threshold.

TABLE IX
BASELINE THRESHOLD ADJUSTMENT USING TEST-SET AND PRACTICE-SET LABELS.

			in-sample optimal		practice-tuned	
split	baseline F1	best non-gen.	thresholded F1	recovered	thresholded F1	recovered
within	0.713	TEA (0.736)	0.725	51%	0.716	11%
short	0.671	PretrainedTEA (0.694)	0.682	49%	0.661	-43%
long	0.688	PretrainedTEA (0.717)	0.693	16%	0.685	-12%

APPENDIX C
PER-TWEET EXAMPLES

A. Always-fail

Table X lists further tweets from the always-fail set of Section IV-C, with examples from both gold labels. The baseline assigns low positive-class probabilities to the positive-labelled examples and high probabilities to the negative-labelled examples.

TABLE X
A SAMPLE OF ALWAYS-FAIL TWEETS. $P(\text{POS})$ IS THE BASELINE MEAN POSITIVE-CLASS PROBABILITY.

Tweet	split	gold	$P(\text{pos})$
I wish my mom was here I already miss the dinner she makes	within	pos.	0.03
Hay I miss Lee Min Ho amp Kim Go Eun The drought is real	long	pos.	0.03
all of the new snow miku designs are so cute how am I supposed to decide which to vote for pos	long	pos.	0.08
I made dinner for my friends but they didn t show up	short	pos.	0.06
not my bf trying to join me working out and gave up after 10 minutes	long	pos.	0.10
I now have another platform to spend hours on avoiding my actual responsibilities and physical wellbeing love that for me	short	neg.	0.92
Happy happy birthday Have a fab weekend	long	neg.	0.98
Thanks mom for telling me how stupid I am for wanting to move schools to better myself Makes me feel great	within	neg.	0.85
The Greatest Showman is one of the best movies I have ever seen	short	neg.	0.66
I was good yesterday I m good today and Ima be good tomorrow Just in case you were wondering	within	neg.	0.91
Cute little lights on the lake	short	neg.	0.91

B. High-drift

Table XI shows six tweets from the highest-drift quintile on long. The final column reports how many systems classify each tweet correctly, and words not seen in training are shown in blue.

TABLE XI
EXAMPLES FROM THE HIGHEST-DRIFT QUINTILE ON LONG, WITH OOV WORDS SHOWN IN BLUE.

Tweet	OOV	sem. dist.	drift	correct (/8)
Every lockdown Gwil and I have completed a series of Married At First Sight Australia New lockdown and a new series YAS	0.136	0.629	0.765	0
Girls want a lot of things from one guy Conversely guys want one thing from a lot of girls RandomThoughts	0.050	0.790	0.840	4
sorry not sorry for my spam	0.000	0.766	0.766	5
its my first time making a carrd like this	0.111	0.728	0.839	6
everyone on tiktok calling ten stans gay i mean as part of ten s cult they re not wrong tho	0.050	0.772	0.822	7
I got new AirPods and every single song sounds literally 1000 times better I swear Everything is so much clearer I m so so happy	0.040	0.744	0.784	8

APPENDIX D

ECS750P/ECS753P/ECS754P – MSc Project Module

Generative AI Tools – Student Accountability Statement

Student Name: Tamara Suárez Rodríguez
Project Title: Mitigating Temporal Drift in Social Media Sentiment Classification: A Comparative Study
Supervisor Name: Arkaitz Zubiaga

1. Have you used Generative AI tools for your project?
YES ☒ NO ☐

2. Please provide details of all the AI tools that you used in your project work.

AI Tool	Version	Publisher	Access URL
ChatGPT	GPT-5.6 Sol	OpenAI	https://chatgpt.com/
Claude	Opus 4.8	Anthropic	https://claude.ai/

3. If you have used Artificial Intelligence (AI) tools in any part of your project please include in table format how, where, and why these tools were used.

Project Section	Description	Justification
Background and literature review	I used both tools to explain unfamiliar concepts, suggest search terms and identify potentially relevant publications.	This helped me understand the research area and focus the literature search. I checked the original publications before using them.
Prior-work search for PretrainedTEA	I used Claude's deep-research feature to complement my search of ACL Anthology and arXiv publications from 2020 to 2026 for work combining continued MLM pretraining with temporal-expert averaging.	This helped me check whether describing PretrainedTEA as a proposed strategy was appropriate. I found no reviewed paper combining both components.
Experimental design and analysis	I used both tools to brainstorm analyses, discuss experimental comparisons, question interpretations and identify possible inconsistencies between draft claims and result files.	This helped me decide which checks to run. I ran the analyses on the project data and verified the reported values against the pipeline outputs.
Software, tests and reproducibility	I used Claude to discuss implementation approaches, identify possible errors, and assist with tests, documentation and LaTeX formatting.	This helped me debug and test my implementation. I reviewed, adapted and ran every suggestion before including it in the project.
Dissertation paper and reflective essay	I used both tools for structural feedback, language review, grammar checks and wording alternatives.	This helped me improve clarity and organisation in English. I chose and revised the final wording.
Interactive interface	I used Claude to discuss the interface structure, identify implementation errors and review how the experiment and per-tweet results were presented.	This helped me refine and test the interface. I reviewed and adapted the suggestions and checked the displayed values against the same generated files used by the paper.
Project presentation	I used Claude to help build the visual template and layout of the slides.	This helped with the presentation design and clarity. I selected and wrote the content of the slides and script.

4. Generative AI Student Declaration

- I have declared all uses of AI tools (e.g., coding assistants, text generators, data analysis tools) in this project.
- I have reviewed and verified all AI-assisted content to ensure accuracy and suitability.
- I take full responsibility for the originality and integrity of my project submission.
- I have referenced all sources, including those suggested or produced with AI support.
- I have not used AI in ways that undermine academic honesty (e.g., plagiarism, unverified code submission).
- My use of AI was supplementary and did not replace my own independent learning and problem-solving.

Signed: 

Date: 15 August 2026