A Systematic Analysis on the Temporal Generalization of Language Models in Social Media

Asahi Ushio[*]
Amazon
asahiu@amazon.com

Jose Camacho-Collados
Cardiff NLP, Cardiff University, UK
camachocolladosj@cardiff.ac.uk

Abstract

In machine learning, temporal shifts occur when there are differences between training and test splits in terms of time. For streaming data such as news or social media, models are commonly trained on a fixed corpus from a certain period of time, and they can become obsolete due to the dynamism and evolving nature of online content. This paper focuses on temporal shifts in social media and, in particular, Twitter. We propose a unified evaluation scheme to assess the performance of language models (LMs) under temporal shift on standard social media tasks. LMs are tested on five diverse social media NLP tasks under different temporal settings, which revealed two important findings: (i) the decrease in performance under temporal shift is consistent across different models for entity-focused tasks such as named entity recognition or disambiguation, and hate speech detection, but not significant in the other tasks analysed (i.e., topic and sentiment classification); and (ii) continuous pre-training on the test period does not improve the temporal adaptability of LMs.

1 Introduction

Modern natural language processing (NLP) is centered on language models (LMs) (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019; Min et al., 2023). The versatility of LMs has enabled many real world applications, including chatbot[1], text-guided image generation (Aditya et al., 2021), and text-to-speech (Paul K. et al., 2023). One of the well-known issues of LMs, however, is that the capabilities of LMs can not be fully analyzed due to their blackbox nature. To overcome such limitations to understand LMs' true capability, methodologies and datasets to inspect LMs have been proposed in the context of model probing study, which uncovered various features such as syntax (Hewitt and Manning, 2019; Goldberg, 2019), factual knowledge (Petroni et al., 2019; Ushio et al., 2021), semantics (Ettinger, 2020; Tenney et al., 2019), and emergent ability (Jason et al., 2022).

Besides such studies of LM probing, there is another line of research that focuses on the adaptability of LMs under settings incurring changing conditions, including *temporal shifts* (Lazaridou et al., 2021; Loureiro et al., 2022a). In this paper, we refer to temporal shifts when discussing settings in which the time period of the test set is different from that of the training set (with the test set period being generally *after*, reassembling real-world settings.). These settings have been empirically known to lead a non-trivial decrease in performance on some tasks (Liska et al., 2022; Jungo et al., 2022). Needless to say, temporal shifts are more important in more dynamic streaming data with frequent meaning changes and evolving entities, such as social media (Antypas et al., 2022; Ushio et al., 2022).

In this paper, we focus on temporal shifts on Twitter, one of the major social media platforms, and propose a unified evaluation scheme to assess the adaptability of LMs toward temporal shift on Twitter. In particular, we are interested in answering the following two research questions:

- **RQ1.** Are temporal shifts in social media detrimental for LM performance in NLP tasks?
- **RQ2.** If so, what are the causes of this temporal shift and can it be mitigated (by e.g. using LMs pre-trained on recent data)?

For the evaluation, we selected five diverse social media NLP tasks for which there are datasets with temporal information available: hate speech detection, topic classification, sentiment classification, named entity disambiguation (NED), and

[*]Work done while at Cardiff NLP
[1]https://openai.com/blog/chatgpt

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 52–62

named entity recognition (NER) ranging over different time periods. The temporal shifts considered are relatively short compared to those studied in other sources of streaming data such as news and scientific papers. We test both LMs specialized on social media and other general-purpose trained on encyclopedic and web-crawled corpus.

Our study shows that tasks driven by named entities or events (i.e., hate speech, NED, and NER) present consistent decrease across model under temporal shift settings, while it is less prominent in the other tasks. Crucially, our results show that the decrease caused by temporal shift cannot be mitigated by considering a more recent corpus to the pre-training dataset. Finally, qualitative analysis highlights that the common mistakes made by LMs are indeed instances that require to understand the named entities in the tweet. All the datasets and the scripts to reproduce our experiments are made publicly available online[2].

2 Related Work

LMs on Social Media. Major LMs are commonly pre-trained on encyclopedic and web-crawled corpora (Lewis et al., 2020; Raffel et al., 2020; Aakanksha et al., 2023; Rohan et al., 2023; Hugo et al., 2023b,a; Tom B. et al., 2020), while the adaptation of such LMs to social media has led new LMs pre-trained on corpus curated over social media (Nguyen et al., 2020; Loureiro et al., 2022a; DeLucia et al., 2022; Barbieri et al., 2022), which present better performance on social media NLP tasks than standard LMs (Barbieri et al., 2020; Antypas et al., 2023). However, such studies on NLP tasks in social media mainly focus on static datasets without temporal shift. A few of them associate timestamps to the dataset and provide basic temporal analysis (Antypas et al., 2022; Ushio et al., 2022), but these are limited to a single task. Finally, related to the temporal aspect of this work, short-term meaning shift has also been studied in the context of social media and LMs (Loureiro et al., 2022b).

Temporal Generalization. Importantly, this work aligns to the research on the temporal or diachronic generalization of LMs. In this area, however, most previous works focus on relatively long term (over 10 years) (Lazaridou et al., 2021) or formal source of text such as news and scientific papers (Liska et al., 2022; Jungo et al., 2022). In the context of short-term temporal analysis, there are three studies that are most similar to ours. Luu et al. (2022) analyse the temporal performance degradation of LMs in NLP tasks in relatively short time periods. While social media is included as one of the domains, the evaluation is limited to the classification task and to general-domain models. Agarwal and Nenkova (2022) performed a similar general analysis for different tasks, while also analysing the effect of self-labeling as a mitigation to temporal misalignment, which we do not analyse in this work. The main difference between these works in ours is our focus on social media, where we carry out a targeted comprehensive analysis on short-term temporal effects. When it comes to social media, temporal shifts are especially relevant given the real-time nature of the platform and their focus on current events. In the context of Italian Twitter, Florio et al. (2020) analysed the temporal sensitivity of models for hate speech detection, which is one of the tasks included in this paper.

Temporal-aware LMs. To enhance adaptability of LMs for temporal shift, there are a few works that explicitly ingests the temporal information to the model by specific attention mechanism (Rosin and Radinsky, 2022), augmenting the input with timestamp (Rosin et al., 2022), joint modeling of temporal information (Dhingra et al., 2022), and self-labeling (Agarwal and Nenkova, 2022). In this paper, we do not include any temporal-aware LMs, because we are interested in analysing the adaptability of plain LMs to temporal shifts.

3 Experimental Setting

In this section, we describe our experimental setting to investigate the effect of temporal shifts in LMs.

3.1 Evaluation Methodology

Let us define $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ as the training and test splits of a dataset $\mathcal{D}$ for a single downstream task (e.g. sentiment classification), where each dataset contains pairs of a text input and associated labels. Importantly, $\mathcal{D}_{\text{train}}$ is taken from the period prior to $\mathcal{D}_{\text{test}}$ without any temporal overlap. Given such dataset with temporal split, we consider the following two settings of out-of-time (OOT) and in-time (IT).

Out-of-Time (OOT). In the first setting, we simply train the models on $\mathcal{D}_{\text{train}}$ and evaluate them

[2] https://huggingface.co/datasets/tweettemposhift/tweet_temporal_shift

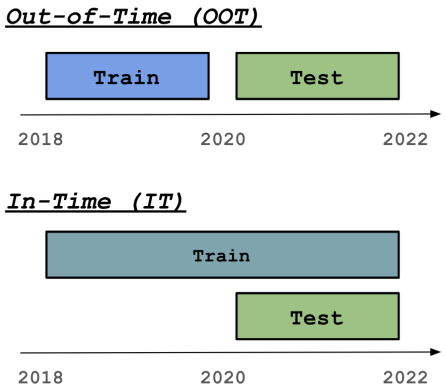


Figure 1: An illustrative example of the conceptual differences between the sampling time periods of the OOT and IT settings.

on $\mathcal{D}_{\text{test}}$. Noticeably, models have no access to the instances from the test period at the training phase in this setting, so we refer the setting as *out-of-time* (OOT) as an analogy to the out-of-domain (OOD).

In-Time (IT). As a comparison to OOT, we consider the second experimental setting, which is designed to assess the effect of training instances from the test period. The test set is randomly split into non-overlapped four chunks ($\mathcal{D}_{\text{test}} = \bigcup_{i=1}^{4} \mathcal{D}^i_{\text{test}}$) for cross validation, where models trained on $\mathcal{D}_{\text{train}} \bigcup \mathcal{D}_{\text{test}} \backslash \mathcal{D}^i_{\text{test}}$ are evaluated on $\mathcal{D}^i_{\text{test}}$. For each chunk of the test set, we downsample the IT training set to the same size as $\mathcal{D}_{\text{train}}$ with three random seeds and report the averaged metrics over the runs. To be precise, we consider a function $\mathcal{F}_s(\mathcal{D})$ that randomly samples $|\mathcal{D}_{\text{train}}|$ instances from $\mathcal{D}$, and we independently train models on $\mathcal{F}_s(\mathcal{D}_{\text{train}} \bigcup \mathcal{D}_{\text{test}} \backslash \mathcal{D}^i_{\text{test}})$ for $s = 0, 1, 2$. In contrast to OOT, we refer this setting as *in-time* (IT) setting.

Figure 1 presents an example overview of the differences between IT and OOT settings from the perspective of data sampling periods (data from 2018 to 2022 in the example).

3.2 Tasks & Datasets

We consider the following five diverse social media NLP tasks: hate speech detection, topic classification, sentiment classification, named entity disambiguation (NED), and named entity recognition (NER). For each task, we employ a public dataset for English and leverage its original temporal splits, unless there is overlap between the periods of training and test sets.

Hate Speech Detection. Hate speech detection in Twitter consists of identifying whether a tweet contains hateful content. We use the dataset proposed by Waseem and Hovy (2016) framed as binary classification as the dataset to create the training and test splits based on the timestamp. The first half is used for training and the rest for test split. The training split is further randomly split into 2:8 for validation:training. We use accuracy to evaluate the hate speech detection models.

Topic Classification. Topic classification is a task that consists of associating an input text with one or more labels from a fixed label set. For this evaluation, we rely on TweetTopic (Antypas et al., 2022), a multi-label topic classification dataset with 19 topics such as *sports* or *music*. As evaluation metric, we use micro-F1 score to measure the performance of topic classification models.

Sentiment Analysis. Sentiment analysis is a standard social media task consisting of associating each post with its sentiment. In particular, we use the dataset from the task 2: LongEval-Classification from CLEF-2023 (Alkhalifa et al., 2023) in which the task is framed as binary classification with positive and negative labels. The original training split contains around 50k instances while 1k test split, which is highly imbalance and the effect of the IT sample can be very limited. To balance the training and test splits, we randomly sample 2.5k instances from each label, amounting 5k new training instance. We use accuracy to evaluate the sentiment classification models.

Named Entity Disambiguation (NED). NED is a a binary classification that consists of identifying if the meaning of a given target entity in context is the same as the one provided. We use the TweetNERD (Mishra et al., 2022) dataset and reformulated into NED following SuperTweetEval (Dimosthenis et al., 2023). Then, we create the train, validation, and test splits in the same way as the hate speech detection. We use accuracy to evaluate the NED models.

Named Entity Recognition (NER). NER is a sequence labelling task to predict a single named-entity type on each token on the input text. We rely on TweetNER7 (Ushio et al., 2022), a NER dataset on Twitter that contains seven named entity types. We use span F1 score to evaluate NER models.

	Split	Size	Date	Examples
Hate	Train	2,318	2013-09-23 / 2015-03-03	*Zebra undies #MKR chic in pink dress.* (Hate)
	Valid	579	2013-09-23 / 2015-03-03	*OMG fashion parade time #mkr.* (non-Hate)
	Test	1,475	2015-03-04 / 2015-03-14	*female football commentators just don't work.* (Hate)
Topic	Train	4,585	2019-09-08 / 2020-08-30	*So, when I can listen to watermelon sugar live in Jakarta Harry?*
	Valid	573	2019-09-08 / 2020-08-30	*@Harry_Styles* (celebrity, music)
	Test	1,679	2020-09-06 / 2021-08-29	*Glad to see the Chiefs crushed the Texans* (sports)
Sent.	Train	5,000	2014-02-06 / 2016-12-31	*I think I'm in love* (positive)
	Valid	1,344	2016-01-01 / 2016-12-31	*@user is making me very upset* (negative)
	Test	1,344	2018-01-01 / 2019-01-01	*Shoutout to @MENTION for donating to poor* (positive)
NED	Train	18,469	2020-02-26 / 2021-08-27	*Every concert I've seen announce lately, they are steering clear of Detroit* (Target: Detroit, Definition: Art museum, Label: False)
	Valid	4,617	2020-02-27 / 2021-08-27	*Me on stream: Happy Friday!, Australia: It's Saturday*
	Test	21,253	2021-08-28 / 2021-11-28	(Target: Australia, Definition: country, Label: True)
NER	Train	4,616	2019-09-08 / 2020-08-30	*UFC 245: Looking at the three title fights on tap at T-Mobile Arena*
	Valid	576	2019-09-08 / 2020-08-30	(UFC 245: corporation, T-Mobile Arena: location)
	Test	2,807	2020-09-05 / 2021-08-31	*Glad the Chiefs crushed the Texans* (Chiefs: group, Texans: group)

Table 1: The number of tweets and the period with examples of each dataset.

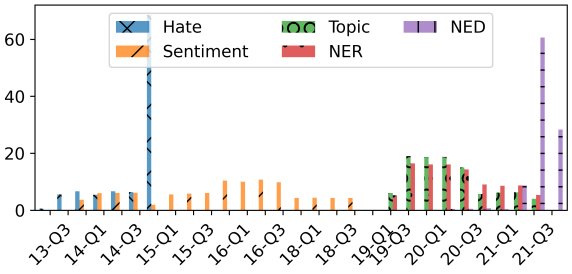


Figure 2: Quarterly breakdown of the number of tweets ratio (%) in each dataset. For example, a ratio of 5% in 13-Q3 for Dataset X would mean that 5% of all tweets in Dataset X belong to the third quarter (July-September) of 2013.

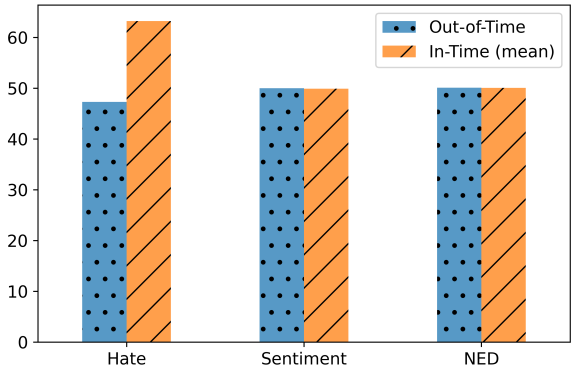


Figure 3: Comparisons of ratio (%) of positive labels in the training split of each task between OOT and IT.

3.2.1 Data Statistics

Table 1 shows the size and the period of the training and the test sets for each dataset, and Figure 2 displays the number of tweets per quarter for each task. Topic classification and NER use the same tweets, which are sampled uniformly from each month, while NED and hate speech detection have the majority of the tweets in the latest quarter. Sentiment analysis covers the longest period in the dataset that spans over four years. Figures 3 and 4 show the comparisons of the label distribution of the binary (i.e., hate speech, sentiment classification, and NED) and multi-classification tasks (i.e., NER and topic classification), respectively. As can be observed, hate speech detection has fewer positive labels in OOT than in IT, while the other two tasks have the same ratio of the positive labels between OOT and IT. The same pattern can be observed for topic classification and NED, for which the label distribution does not substantially change.

3.3 Models

We investigate an established general-purpose LM, RoBERTa (Liu et al., 2019) as well as other LMs pre-trained on tweets including BERTweet (Nguyen et al., 2020), TimeLM (Loureiro et al., 2022a), and BERNICE (DeLucia et al., 2022). For RoBERTa and BERTweet, we consider the base and the large models, referred as RoBERTa (B), RoBERTa (L), BERTweet (B), and BERTweet (L). For TimeLM, we consider the base models trained on the tweets up to 2019, 2020, 2021 and 2022, referred as TimeLM2019 (B), TimeLM2020 (B), TimeLM2021 (B) and TimeLM2022 (B), and the large model trained upto 2022, referred as

Model	Parameters	HF Name	Citation
RoBERTa$_{BASE}$	123M	`roberta-base`	(Liu et al., 2019)
RoBERTa$_{LARGE}$	354M	`roberta-large`	
BERTweet$_{BASE}$	123M	`vinai/bertweet-base`	(Nguyen et al., 2020)
BERTweet$_{LARGE}$	354M	`vinai/bertweet-large`	
TimeLM2019$_{BASE}$	123M	`cardiffnlp/twitter-roberta-base-2019-90m`	(Loureiro et al., 2022a)
TimeLM2020$_{BASE}$	123M	`cardiffnlp/twitter-roberta-base-dec2020`	
TimeLM2021$_{BASE}$	123M	`cardiffnlp/twitter-roberta-base-2021-124m`	
TimeLM2022$_{BASE}$	354M	`cardiffnlp/twitter-roberta-base-2022-154m`	
TimeLM2022$_{LARGE}$	354M	`cardiffnlp/twitter-roberta-large-2022-154m`	
BERNICE	278M	`jhu-clsp/bernice`	(DeLucia et al., 2022)

Table 2: Language models used in the paper with the number of parameters and model aliases on Hugging Face.

	Hate	Topic	Sentiment	NED	NER
RoBERTa	✓		✓		
BERTweet	✓		✓		
BERNICE	✓	✓	✓	✓	✓
TimeLM2019	✓		✓		
TimeLM2020	✓		✓		
TimeLM2021	✓	✓	✓	✓	✓
TimeLM2022	✓	✓	✓	✓	✓

Table 3: The overlap between the test period and the pre-trained corpus of each LM (✓indicates that the LM is pre-trained on the corpus including the test period of the task).

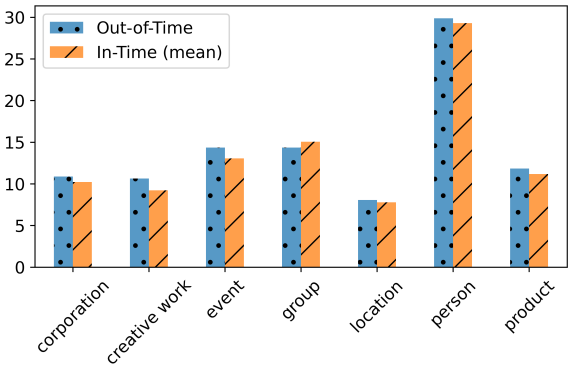
TimeLM2022 (L). The end date of the pre-trained corpus for each model is 2019-02 (RoBERTa), 2019-08 (BERTweet), 2019-12 (TimeLM2019), 2020-12 (TimeLM2020), 2021-12 (TimeLM2021 and BERNICE), and 2022-12 (TimeLM2022). All the model weights are taken from the transformers model hub (Wolf et al., 2020) and Table 2 shows the details of models we used in the paper.[3] Table 3 shows the overlap between the period of the pre-trained corpus and the test set for each task, which will be relevant for the analysis on the effect of pre-training in Section 5.1. These models are then fine-tuned in the datasets presented in the previous section, in both OOT and IT settings. For model fine tuning, we run hyperparameter search with Optuna (Akiba et al., 2019) with the default search space.

4 Results

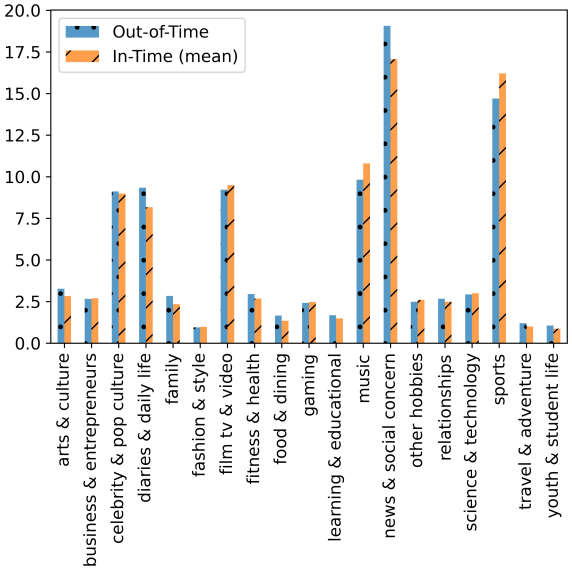
Figures 5 to 9 show the comparisons of IT and OOT in hate speech detection, NED, NER, topic classification and sentiment analysis. As can be observed, hate speech detection, NED and NER present inconsistencies in both settings, decreasing the performance from IT to OOT. In contrast, this cannot be observed for both sentiment analysis, and especially topic classification. The average decrease of OOT performance for each of the tasks is 4.5, 2.4, 1.7, 0.8 and -0.1 for hate speech detection, NER, NED topic classification and sentiment analysis.

One of the main differences of those two groups of tasks (i.e. hate/NED/NER v.s. topic/sentiment) entity-centric or event-driven nature of the former. NER and NED are clearly related to named entities. Hate speech detection does not relate to named entities explicitly, but since the tweets for hate speech detection are collected by querying specific events, they are often about events or celebrities which peak around the sampled timestamp (Gómez et al., 2023). On the other hand, events or named entities are not as important in sentiment analysis, as the sentiment can be estimated from the context in most cases. Topic classification depends on the topic, with some of them related to entities (e.g. those related to celebrities or TV) and others not (e.g., daily life, family or food), but in the main clearly identifiable by the context. Through the lens of entity relevancy, this result may suggest that the temporal shift can be caused by named entities, which includes meaning drift of existing named

[3]Note that for this analysis we are not interested in the performance of zero-shot LLMs such as GPT-4, but rather on the effect of fine-tuned LMs.



(a) Ratio of entities in NER.



(b) Ratio of labels in topic classification.

Figure 4: Comparisons of label distributions between OOT and IT settings.

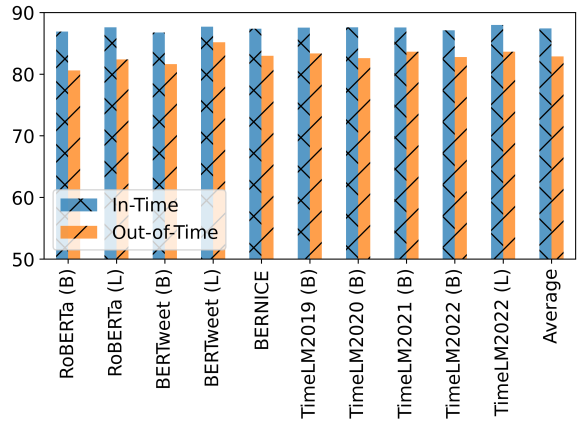


Figure 5: Comparisons of IT and OOT performance (accuracy) for hate speech detection.

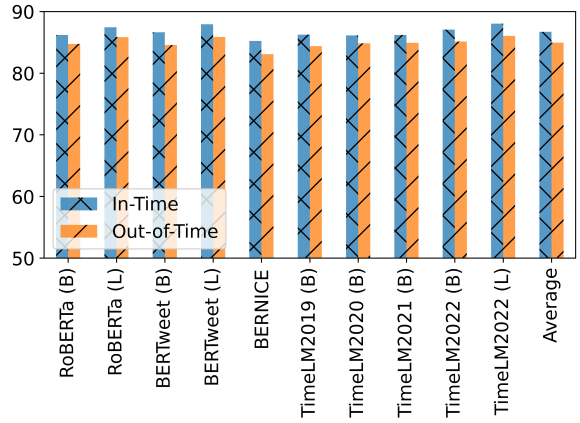


Figure 6: Comparisons of IT and OOT performance (accuracy) for NED.

entities or emerging new named entities. Topic classification can be seen as a mixture of entity-related instances and not, which results in not fully consistent gain from OOT, but still significant in the average.

5 Analysis

This section focuses on the second research question (RQ2) and analyses the main causes behind temporal shift performance degradation of LMs.

5.1 Effect of Pre-Training

A possible direction to mitigate the temporal shift is to pre-train the LMs on the text from the test period, which does not require any labeling. Figure 10 visualizes the performance and relative IT improvement of LMs with/without pre-training corpus covering the test period of each task for topic classification/NED/NER[4]. At a glance, we cannot observe see any relationship between the pre-training corpus and the performance. The averaged relative gains of the metrics from OOT within the LMs pre-trained on the test period and the others are 2.0 and 0.6 (topic classification), 3.5 and 3.8 (NER), and 2.1 and 1.9 (NED) respectively. Therefore, all models are affected by the temporal shift irrespective of the pre-training corpus date. This implies that the temporal shift cannot be robustly resolved by only adding data from the test period to the pre-training corpus, a conclusion that was also reached by Luu et al. (2022).

5.2 Effect of Label Distribution

In supervised machine learning label distribution, the distribution of the binary label over the test instances, shifts can affect a model's performance. In

[4]The test periods of hate speech detection and sentiment classification are covered by all the LMs we considered in the experiment.

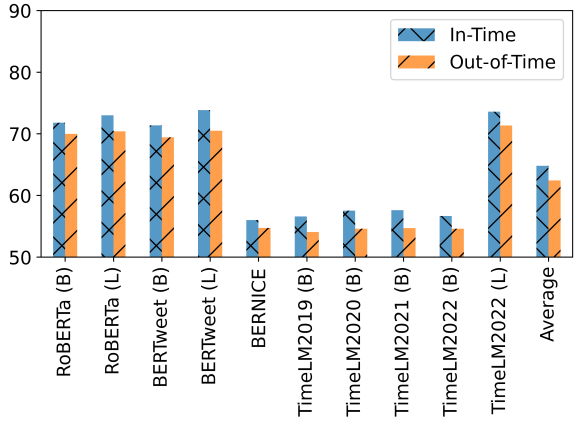


Figure 7: Comparisons of IT and OOT performance (F1 score) for NER.

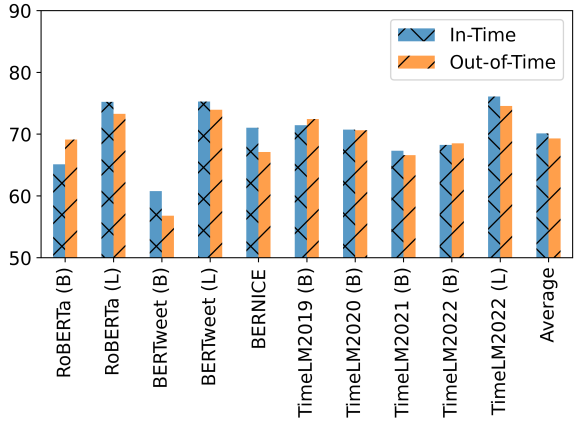


Figure 8: Comparisons of IT and OOT performance (F1 score) for topic classification.

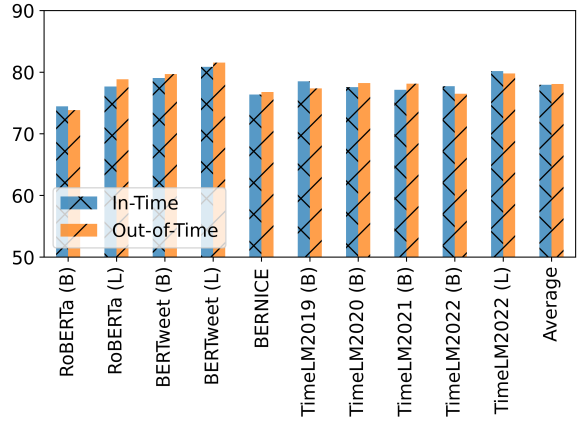


Figure 9: Comparisons of IT and OOT performance (accuracy) for sentiment classification.

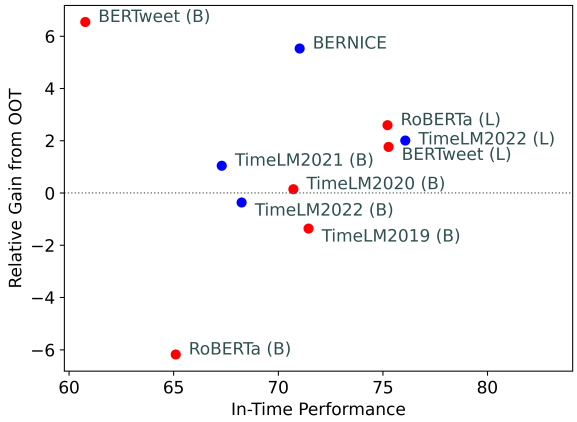
	Original	Balanced
RoBERTa (B)	7.25	5.19
RoBERTa (L)	5.96	-0.95
BERTweet (B)	5.91	4.84
BERTweet (L)	2.88	-0.30
BERNICE	5.04	4.72
TimeLM2019 (B)	4.80	4.16
TimeLM2020 (B)	5.71	5.39
TimeLM2021 (B)	4.51	5.52
TimeLM2022 (B)	4.97	0.15
TimeLM2022 (L)	4.94	1.89
Average	5.20	3.06

Table 4: Comparisons of relative accuracy gain from OOT to IT between original (unbalanced) and balanced label distributions for hate speech detection.

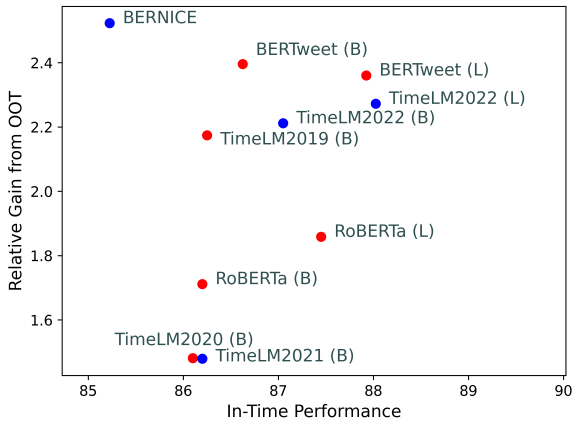
this section, we analyse this potential effect when it comes to temporal shifts. For this, we rely on hate speech detection, which presents the largest decrease in performance from IT to OOT, with a different label distribution between training and test (see Figure 3). For the other tasks, the label distribution appears to be largely similar. To separate the effect of label distributional shift between IT and OOT from the temporal shift, we conduct a controlled experiment by balancing the label distribution of each IT training split to be the same as OOT training split. This is achieved by undersampling the size of the training set. Table 4 shows the results, where the average relative gain is still positive, although it becomes less dominant in balanced experiment. This highlights how label distribution may change over time and this itself have an effect in model performance. A similar finding was already discussed by Luu et al. (2022).

5.3 Qualitative Analysis

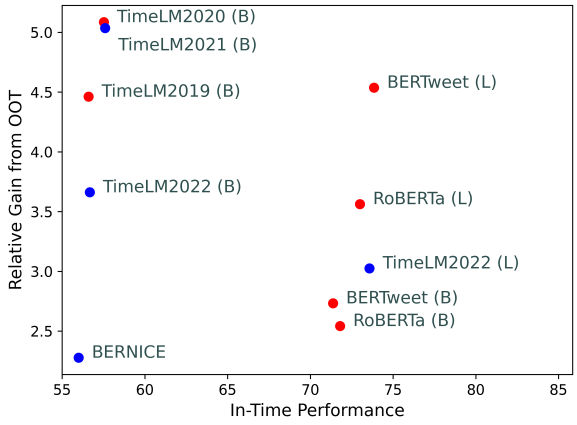
In this analysis, we have a closer look on the test instances that are incorrect in OOT, turning to be correct in IT. To be precise, we sort the test instance in a single task based on the number of models where the error in OOT setting has been corrected in IT setting over all the random seeds. In other words, given a test instance, we check whether a model prediction is incorrect in OOT, but correct in the IT setting. This particular instance is counted as a correction. In total, we have 10 models with 3 independent runs with different random seed to construct the training data, so 30 would be the maximum number of corrections. For sentiment classification, hate speech detection, topic classification and NED, we simply count instance-level corrections. Given the complex nature of NER evaluation, we decided to only focus on the entity type



(a) Topic classification.



(b) NED.



(c) NER.

Figure 10: Relative improvement (%) from OOT to IT for each task (topic, NED and NER). LMs with pre-training corpus including the test period are in blue, and those without temporal overlap in red.

predictions for this analysis.

Table 5 shows the top instances in terms of IT corrections for each of the task. We can observed the marked differences across tasks, with NED and hate speech detection including instances which were corrected 100% in the OOT setting. In fact,

Task	Top corrected	Avg Top 10
NED	30/30 (100%)	30.0
Hate	30/30 (100%)	30.0
NER	28/30 (93.3%)	24.0
Sentiment	19/30 (63.3%)	13.6
Topic	16/30 (53.3%)	12.5

Table 5: Top instances in terms of number of predictions corrected with an IT split. The second column indicates the top 10 average.

Task	Instance	Gold	Times corrected
NED	so cute how <Aoki> describes Ida. "thinks about things seriously" *(Japanese manga series)*	False	30/30 (100%)
	Will Ram & <Priya> go on a honeymoon it'll be a nice break for them (...) #BadeAchhe-LagteHain2 *(Indian actress)*	False	30/30 (100%)
Hate	#MKR God Kat you are awful awful person. Oh you are humiliated? GOOD.	False	30/30 (100%)
	#katandandre gaaaaah I just want to slap her back to WA #MKR	False	30/30 (100%)

Table 6: Two examples from the NED and hate speech detection datasets in which the prediction was corrected 100% of the times with an IT split. For NED, the definition is provided in parenthesis and target word indicated between < and >.

there are respectively 44 and 15 instances for which this is the case in these two tasks. Similarly for NER, the number of corrections is high. This is correlated with the main results of the paper (see Section 4) which showed clear improvements for these tasks in the IT setting, but not for sentiment and and topic classification.

Finally, Table 6 shows some of these instances for NED and hate speech detection. In the case of NED, the tweets relate to two new TV series that were on air at test time (Japanese *Kieta Hatsukoi* in the first example and Indian B*ade Achhe Lagte Hain* in the second, both from 2021). This is similar to the hate speech detection in which the examples

belong to the *My Kitchen Rules TV* show. This highlights the event-driven nature of social media, and the importance of acquiring the background context for the specific task.

6 Conclusion

We proposed an evaluation method to assess the adaptability of LMs for temporal shifts on social media with five diverse downstream tasks including sentiment classification, NER, NED, hate speech detection, and topic classification. We have tested diverse LMs trained on Twitter under different temporal settings. The experimental results indicate that the adaptability gets consistently worse on entity or event-driven tasks (NED, NER, and hate speech detection) while the effect is limited in the other tasks. This conclusion was similar to previous work in more general domains, which observed a variation across different types of task when it comes to temporal degradation (Luu et al., 2022; Agarwal and Nenkova, 2022). Finally, our analysis shows that pre-training on a corpus from the test period is not enough to solve the temporal shift issue, with performance still being degraded in comparison to models fine-tuned on the labeled dataset from the test period.

Limitations

Regardless of some similarities between Twitter and other streaming data such as news and other social media platforms being real-time and trend-driven, they can have different characteristics, and the results of our study may apply to Twitter exclusively. For our evaluation we rely on a single dataset for each of the tasks. Of course, these datasets are not a faithfully representation of the task and may contain their own biases. Therefore, even for the same task, the findings in this paper may differ if using a different dataset.

Ethical Statement

The datasets we used in the experiments are all from Twitter. Data has been anonymized (only information about legacy-verified users is kept) so that they do not contain any personal identifiable information (PII). We do not gather information from individual accounts but rely on aggregated information and metrics only. Please note that the text may contain sensitive content due to the nature of social media and the task, in particular hate speech detection.

Acknowledgements

Jose Camacho-Collados is supported by a UKRI Future Leaders Fellowship.

References

Chowdhery Aakanksha, Narang Sharan, Devlin Jacob, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Ramesh Aditya, Pavlov Mikhail, Goh Gabriel, et al. 2021. Zero-shot text-to-image generation. *Preprint*, arXiv:2102.12092.

Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the ACL*, 10:904–921.

Takuya Akiba, Shotaro Sano, Yanase, et al. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Rabab Alkhalifa, Iman Bilal, Hsuvas Borkakoty, et al. 2023. Overview of the clef-2023 longeval lab on longitudinal evaluation of model performance. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 440–458. Springer.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, et al. 2023. SuperTweetEval: A challenging, unified and heterogeneous benchmark for social media NLP research. In *Findings of EMNLP 2023*, pages 12590–12607, Singapore.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, et al. 2022. Twitter topic classification. In *Proceedings of COLING*, pages 3386–3400, Gyeongju, Republic of Korea.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the ACL: EMNLP 2020*, pages 1644–1650, Online. ACL.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of LREC*, pages 258–266, Marseille, France.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on EMNLP*, pages 6191–6205, Abu Dhabi, United Arab Emirates. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, Minneapolis, Minnesota.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the ACL*, 10:257–273.

Antypas Dimosthenis, Ushio Asahi, Barbieri Francesco, et al. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In *Findings of EMNLP 2023*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the ACL*, 8:34–48.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Jesús Gómez, Alberto Matilla-Molina, Ma Pilar Amado, Dimosthenis Antypas, Jose Camacho-Collados, Carlos J Máñez, Tomás Fernández-Villazala, Alicia Méndez-Sanchís, and Javier López. 2023. The interaction between offensive and hate speech on twitter and relevant social events in spain. In *News Media and Hate Speech Promotion in Mediterranean Countries*, pages 81–109. IGI Global.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. ACL.

Touvron Hugo, Martin Louis, Stone Kevin, et al. 2023a. Llama 2: Open foundation and fine-tuned chat models.

Touvron Hugo, Lavril Thibaut, Izacard Gautier, et al. 2023b. Llama: Open and efficient foundation language models.

Wei Jason, Tay Yi, Bommasani Rishi, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Kasai Jungo, Sakaguchi Keisuke, Takahashi Yoichi, et al. 2022. Realtime qa: What's the answer right now? *Preprint*, arXiv:2207.13332.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363.

Mike Lewis, Yinhan Liu, Naman Goyal, et al. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*, pages 7871–7880, Online.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, et al. 2022a. TimeLMs: Diachronic language models from Twitter. In *Proceedings of ACL: System Demonstrations*, pages 251–260, Dublin, Ireland.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, et al. 2022b. TempoWiC: An evaluation benchmark for detecting meaning shift in social media. In *Proceedings of COLING*, pages 3353–3359, Gyeongju, Republic of Korea.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States.

Bonan Min, Hayley Ross, Elior Sulem, et al. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Shubhanshu Mishra, Aman Saini, Raheleh Makki, et al. 2022. Tweetnerd–end to end entity linking benchmark for tweets. *arXiv preprint arXiv:2210.08129*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*, pages 9–14, Online. ACL.

Rubenstein Paul K., Asawaroengchai Chulayuth, Dung Nguyen Duc, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, et al. 2019. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*, pages 2463–2473, Hong Kong, China.

Alec Radford, Jeffrey Wu, Rewon Child, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Anil Rohan, Dai Andrew M., Firat Orhan, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 833–841, New York, NY, USA.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the ACL: NAACL 2022*, pages 1498–1508, Seattle, United States. ACL.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 4593–4601, Florence, Italy. ACL.

Brown Tom B., Mann Benjamin, Ryder Nick, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Asahi Ushio, Francesco Barbieri, Vitor Sousa, et al. 2022. Named entity recognition in Twitter: A dataset and analysis on short-term temporal shifts. In *Proceedings of the 2nd Conference of AACL*, pages 309–319, Online only.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, et al. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of ACL*, pages 3609–3624, Online.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. ACL.

Thomas Wolf, Lysandre Debut, Victor Sanh, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*, pages 38–45, Online.