The Temporal Persistence of Generative Language Models in Sentiment Analysis

Notebook for the LongEval Lab at CLEF 2023

Pablo Medina-Alias, Özgür Şimşek

University of Bath, Claverton Down, Bath, United Kingdom

Abstract

Pre-trained transformer-based language models (PLMs) have revolutionised text classification tasks, but their performance tends to deteriorate on data distant in time from the training dataset. Continual supervised re-training help address this issue but it is limited by the availability of newly labelled samples. This paper explores the longitudinal generalisation abilities of large generative PLMs, such as GPT-3 and T5, and smaller encoder-only alternatives for sentiment analysis in social media. We investigate the impact of time-related variations in data, model size, and fine-tuning on the classifiers' performance. Through competitive evaluation in the CLEF-2023 LongEval Task 2, we compare results from fine-tuning, few-shot learning, and zero-shot learning. Our analysis reveals the superior performance of large generative models over the benchmark RoBERTa and highlights the benefits of limited exposure to training data in achieving robust predictions on temporally distant test sets. The findings contribute to understanding how to build more temporally robust transformer-based text classifiers, reducing the need for continuous re-training with annotated data.

Keywords

text classification, temporal robustness, pre-trained language representations, text generation, social media analysis

1. Introduction

Pre-trained transformer-based language models (PLMs) such as BERT [1] have been extremely successful on text classification tasks. However, their performance tends to deteriorate when they are tested on data that is distant in time from the initial training dataset. This lack of "temporal persistence" is often addressed by continuously updating the models with newly annotated data [2, 3]. Simultaneously, recent progress in NLP has seen the rise of large-scale, generative PLMs such as GPT-3 [4] and T5 [5], which have redefined the state-of-the-art in a wide range of NLP tasks, even when only learning from a few examples. The enhanced performance of these models comes at the expense of larger model sizes, making it increasingly costly to update them with incoming samples. As such, understanding how to build more temporally robust transformer-based text classifiers is essential to help alleviate the need for continuous re-training with annotated data.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18âĂŞ21, 2023, Thessaloniki, Greece

pjm69@bath.ac.uk (P. Medina-Alias); o.simsek@bath.ac.uk (Ö. Şimşek)

0009-0001-4202-8664 (P. Medina-Alias); 0000-0001-5449-0437 (Ö. Şimşek)

With the aim of encouraging new research on temporally robust classifier systems, CLEF-2023 LongEval Task 2 [6] proposed to competitively evaluate the short and long-term persistence in performance of models for sentiment analysis in social media. In this paper, we narrow the research scope of the task to focus on the comparison of the longitudinal generalisation abilities between that of large pre-trained generative models and those of smaller encoder-only alternatives. Moreover, we strive to explore methods for training more temporally robust models. We hypothesise that extensive fine-tuning over a short-term optimisation objective might result in their decreased performance on new, evolving, datasets. Therefore, we address the following research questions:

1. How well do generative PLMs adapt to variations in the time-related characteristics of the test data?
2. How does the longitudinal drop in classification performance of encoder-only models compare to that of decoder-only and encoder-decoder models?
3. Does model size contribute to better longitudinal generalisation in PLMs?
4. How does the amount of fine-tuning affect the effectiveness of PLMs in preventing performance deterioration on temporally evolving datasets?

In an endeavour to address these questions while achieving competitive results in the task, we focused on GPT-3 and T5. We present the results obtained by these models in the evaluation phase of the competition with a comparative view of fine-tuning, few-shot, and zero-shot learning. Furthermore, we provide a post-evaluation analysis of the impact of model size and exposure to training data on the short and long-term classification scores. Our results show the superiority of large transformer-based generative models over the task benchmark, RoBERTa [7], and show that model performance in the temporally distant test set benefited from limited exposure to training data. This generalisation capability resulted in more robust predictions, outperforming all other competing models in this task (see Section 5).

The body of this paper is organised as follows: Section 2 provides an overview of previous related research. Section 3 summarises the details of the task. Our methodological approach is described in Section 4, followed by the analysis of the results in every phase of the competition in Section 5. Section 6 presents our conclusions.

2. Related Work

Performance deterioration over time. Recent work has demonstrated the lack of temporal persistence of machine learning models on NLP tasks. Lukes and Søgaard [8] studied how shifts in lexical polarity negatively impacted the performance of sentiment classifiers that use logistic regression. Florio et al. [9] dynamically tested BERT for hate speech detection, showing that shifts in the topical composition of the test data resulted in a decrease in classification scores. Lazaridou et al. [2] showed that the performance degradation of transformer-based language models over time transfer into downstream tasks such as part-of-speech (POS) tagging or question answering (QA), which cannot be prevented with a larger model size. This last result is in line with the findings of Agarwal and Nenkova [10], who also found evidence against the performance deterioration of pre-trained language representations in tasks with more stable label definitions such as sentiment and domain classification. More recently, Alkhalifa et al. [3] extended the analysis of longitudinal persistence to a larger set of models for text classification, including autorregressive architectures such as Hierarchical Attention Network (HAN) [11] and Generative Pre-trained Transformer 2 (GPT-2) [12]. They observed a generalised performance drop across models, with transformer-based classifiers consistently yielding better results. Our work brings new evidence on the longitudinal performance of two generative models that, to the best of our knowledge, have not been tested for this task: GPT-3 and T5.

Improving temporal persistence. In relation to building classification models that dynamically adapt to non-stationary NLP tasks, Lukes and Søgaard [8] proposed a feature selection approach based on estimating the lexical polarity rank for a given period of time to induce temporal robustness in sentiment classification. He et al. [13] introduced two evolutionary learning methods for neural networks based on temporal parameter smoothing and diachronic propagation when training on temporarily split data. Other approaches focus on the continual adaptation of language representations to prevent performance loss in new streams of data [14, 15]. Alkhalifa et al. [16] mitigated performance drop in neural stance detection by incrementally updating n-gram based embeddings. Alternatively, integrating temporal information into contextual representations; as a result of masked pre-training [17, 18, 19], a combination with temporal embeddings [20], or by its inclusion into the model's attention mechanism [21], effectively reduced temporal deterioration in downstream tasks. In this regard, Röttger and Pierrehumbert [19] found that although temporal adaptation did contribute to improved language modelling and text classification, it was less effective than domain adaptation.

3. Task Description

The CLEF-2023 LongEval Task 2 [6] provided participants with a binary sentiment classification challenge, wherein the test data was progressively more distant in time from the training data. The training data comprised a corpus of approximately 50,000 distantly-annotated English Tweets from 2016, extracted from the TM-Senti Dataset [22]. Model performance was assessed using the Macro-F1 score and the relative performance drop (RPD)[1]. The evaluation datasets used by [6] consisted of human-annotated development and test sets associated with three temporal benchmarks: 2016 ("within time" as this overlapped the training data), 2018 ("short-term"), and 2021 ("long-term"). To facilitate the utilisation of unsupervised methods, an additional set of around 1 million unlabelled data samples, along with their corresponding creation year spanning from 2013 to 2021, was provided.

The task encompassed three distinct phases. During the **development phase**, participants were granted access to the training and development data, thereby enabling them to develop and pre-evaluate their classifiers. In the **evaluation phase**, candidate models underwent testing against one test set per temporal benchmark. The number of submission attempts was limited to a maximum of five. Finally, the **post-evaluation phase** facilitated further analysis of model performance by disclosing the ground truth labels of the test sets.

[1] The RPD is calculated as the relative change in the F1-score between two testing periods.

4. Methods

We focused on three well established transformer-based architectures for natural language tasks: (1) RoBERTa, (2) T5, and (3) GPT-3. Table 1 provides an overview of the models used in the present work.

Table 1
Summary of models

Model	Type	Param. Size	Vocab Size (EN)
RoBERTa_base	Encoder	125M	50K
RoBERTa_large	Encoder	355M	50K
T5_base	Encoder-Decoder	222M	32K
T5_large	Encoder-Decoder	737M	32K
GPT-3 (babbage)	Decoder	1.3B	52K
GPT-3 (curie)	Decoder	6.7B	52K
GPT-3 (davinci)	Decoder	175B	52K

RoBERTa, serving as our baseline and representative encoder-only model, has consistently exhibited robust performance in natural language understanding tasks. We fine-tuned[2] two sequence classifiers with RoBERTa_base and RoBERTa_large encoders, respectively. Given the nature of the competition, we selected the best model check-point and hyper-parameters attending to loss minimisation in the "within-time" development set (2016). All RoBERTa classifiers were trained by using AdamW with linear learning rate decay (decay=1e-3). We found 2e-05 and 1e-05 to be the best learning rates for RoBERTa_base and RoBERTa_large, respectively.

For the evaluation of encoder-decoder models, we chose **T5**. Specifically, we fine-tuned T5_base and T5_large for binary text-to-text classification. T5 models were optimised with Adafactor, and a learning rate of 1e-04. The maximum sequence length was fixed to 128 during RoBERTa and T5 training.

Finally, we used **GPT-3** to test the performance of decoder-only models at scale. In addition, so as to evaluate the in-context and fine-tuned effectiveness of LLMs, we evaluated GPT-3 in three distinct learning paradigms: zero-shot (0-S) and few-shot (F-S) learning with instruction prompting, as well as fine-tuning. We selected the largest GPT-3 model, *'davinci'* [3], for prompted inference; and the smaller *'babbage'* and *'curie'* versions for supervised re-training. GPT-3 models were re-trained using the OpenAI API, with learning rates of 4e-5 for GPT-3 babbage and 2.2e-5 for GPT-3 curie. For the few-shot scenario, we retrieved fifteen random samples from the annotated training data. The prompts used for zero and few-shot inference can be found in section A of the appendix.

[2] Only the labelled sets were used during training in this work. Therefore, future mentions to training data exclude the additional unlabelled samples provided in the competition.

[3] https://platform.openai.com/docs/models

5. Results and Analysis

We present the results of the 2023 CLEC-LongEval [4] classification in Table 2.

Table 2
Final leader board for LongEval 2023 Task 2.

Position	Team	F1 Within	F1 Short	F1 Long	Overall Drop	Overall Score
1	**This work**	0.7377	0.6739	**0.6971**	-0.0708	***0.7029***
2	CLEF-LE (baseline)	**0.7459**	**0.6839**	0.6549	-0.1025	0.6945
3	Cordyceps	0.7246	0.6771	0.6549	-0.0669	0.6923
4	saroyehun	0.7203	0.6674	0.6874	-0.0596	0.6917
5	pakapro	0.533	0.4648	0.4910	**-0.0504**	0.4863

Our winning model is T5, which, surprisingly was not the best performing on the development set. The results in the development and evaluation splits are presented in Tables 3 and 4. During the development phase, we observed that the fine-tuned versions of GPT-3 and T5 achieved higher F1 scores compared to the baseline RoBERTa model. However, due to the typical formulation of sentiment analysis as a three-class problem, the zero-shot predictions of GPT-3 often included the 'neutral' label despite our efforts in prompt design, which negatively impacted its score. Although GPT-3 achieved the highest overall F1 score, T5 demonstrated better short-term stability.

Table 3
Results on the development set, both within (2016) and short-term (2018).

	F1 Within	F1 Short	RP Drop	Overall Score
RoBERTa_base (baseline)	0.788	0.761	-0.034	0.775
T5_base	0.791	0.771	**-0.025**	0.781
GPT3_davinci (0-S)	0.718	0.693	-0.035	0.706
GPT3_davinci (F-S)	0.746	0.697	-0.066	0.722
GPT3_babbage (F-T)	**0.823**	**0.798**	-0.030	**0.811**

In the evaluation phase, T5 emerged as the best-performing model overall, achieving an **average F1 score of 0.703** across the three temporal benchmarks. However, it was outperformed by GPT-3 in the short-term evaluation. Fine-tuned GPT-3's overall performance was inferior to the benchmark, which we attribute to overfitting towards the 2016 score during the development phase. Notably, GPT-3Âťs largest version outcompeted the baseline when prompted with a few examples in the *within time* period, but this performance was not sustained further in time. Regarding longitudinal performance drop, the fine-tuned version of GPT-3 exhibited more robust predictions. Surprisingly, the zero and few-shot versions of GPT-3 displayed the highest performance drop among all models.

To explore the impact of model size on performance, we also investigated the results obtained by the large versions of these models in the post-evaluation phase. For that, we selected the

[4] https://codalab.lisn.upsaclay.fr/competitions/12762#results

Table 4
Results on the Test set, within, short, and long term.

	F1 Within	F1 Short	F1 Long	Overall Drop	Overall Score
**in evaluation phase*					
RoBERTa (baseline)	0.727	0.670	0.687	-0.067	0.695
T5_base	**0.738**	0.674	**0.697**	-0.071	**0.703**
GPT3_davinci (0-S)	0.706	0.632	0.647	-0.094	0.662
GPT3_davinci (F-S)	0.728	0.664	0.656	-0.093	0.683
GPT3_babbage (F-T)	0.720	**0.682**	0.675	**-0.058**	0.692
**in post-evaluation phase*					
RoBERTa_large	0.742	0.679	0.681	-0.084	0.701
T5_large	**0.747**	**0.682**	**0.725**	-0.058	**0.718**
GPT3_curie (F-T)	0.725	0.689	0.687	**-0.051**	0.700

6.7B-parameter version of GPT-3, ‘curie’; T5_large; and RoBERTa_large. All models yielded higher F1 scores than their smaller versions in the three temporal benchmarks. However, a larger number of parameters did not prevent longitudinal performance drop. Again, the encoder-decoder option provided the best overall classification results, showcasing its superiority among the considered models.

5.1. Longitudinal Overfitting

Given the significant difference in the model performance between the development and the longitudinally-distant test sets, we investigate the possible causes for performance drop using three tests. The first is a statistical analysis of the similarity of each of the data splits, inspired by Tayyar Madabushi et al. [23]. The second consists of comparing the topical distribution in each temporal split, inspired by Florio et al. [9], who show that such topical differences result in a drop in performance across temporal splits. Finally, we explore the impact of the number of training samples on T5, our best performing model on the test set, given the fact that the training data is, by virtue of the task, similar only to one of the evaluation sets (“within”).

For our statistical exploration of splits, we use the Wilcoxon signed-rank test [24]. The process involves randomly sampling observations from different data splits and comparing their frequency counts to assess whether the datasets exhibit a statistically similar distribution of words. Table 5 presents the range of minimum and maximum p-values obtained from ten runs of each pair, along with their corresponding interpretations. It can be observed that the statistical similarity between the “within dev” (what is optimised for) and the “long test” is the least with a p-value of less than 0.05, which is also what is reflected in terms of the performance drop of our model.

For our analysis of the topical distribution of the different evaluation splits, we use a BERT-based pre-trained topic classifier to extract topical information [25]. This result, presented in Table 6, shows that the distribution of most common topics does not exhibit significant changes along the temporal data splits, except for the “News" category, likely associated to the COVID-19 pandemic. Thus, we conclude that while the broader topics were similar, the nature of these

Table 5
Results from the Wilcoxon Signed-rank test for Corpus Similarity and corresponding p-values.

Corpus Pair	Min p value	Max p value	%Same
Train vs Train	0.2302	0.9314	100
Train vs Within_eval	0.2244	0.9347	100
Train vs Short_eval	0.1673	0.2969	100
Train vs Long_eval	**0.0123**	0.6077	90
Within_dev vs Within_eval	0.2629	0.9987	100
Within_dev vs Short_eval	**0.0284**	0.6752	90
Within_dev vs Long_eval	**0.0108**	0.0761	40

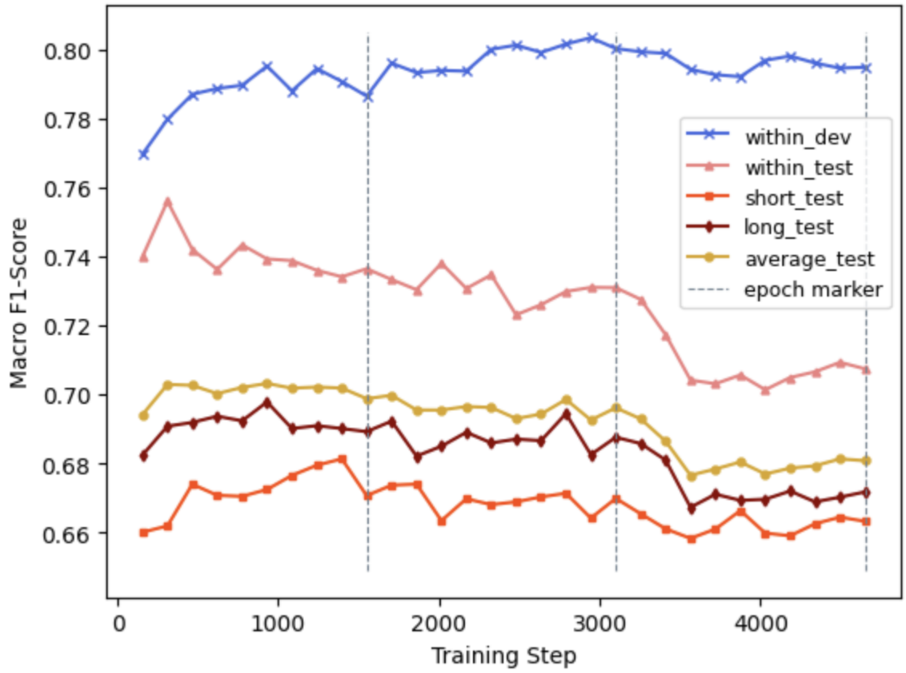
Table 6
Frequence (%) of top 10 most common of topics in all data splits.

	Daily Life	Relationships	Celebrities	Music	Film and TV
Train	57.34	11.48	8.24	7.40	5.33
Within_dev	57.89	11.77	6.03	6.70	4.02
Short_dev	54.46	8.33	6.47	7.07	5.73
Within_eval	62.54	10.57	4.74	5.84	5.84
Short_eval	54.96	7.71	5.84	6.61	6.50
Long_eval	48.79	6.17	7.16	6.28	6.72
	Food	Sports	Education	News	Family
Train	4.72	4.39	3.72	3.24	3.23
Within_dev	5.06	4.84	4.76	3.72	3.20
Short_dev	5.51	6.18	2.68	3.27	3.35
Within_eval	5.18	4.41	4.08	4.52	3.19
Short_eval	5.29	6.61	2.53	3.63	2.53
Long_eval	4.19	4.18	3.30	8.37	3.30

discussions, subtopic distributions, or their associated sentiment might have evolved, resulting in a drop in our classifier performance.

Finally, we investigate the impact of additional training steps on the longitudinal robustness of T5, our best performing model. Figure 1 illustrates the F1-Score obtained in the “within" development set and the three test splits as the model is fine-tuned with the training data during three epochs. We averaged the results of three training instances initiated with different random seeds and the hyperparameters described in section 4. We note that T5 pre-trainning is enough to offer competitive results in this task. However, this capability includes predicting the label “neutral”, which can be adapted to the binary specification within a few hundred steps. All subsequent training results in over-fitting, and consequently, in a drop in longitudinal performance. We believe this to be a result of catastrophic forgetting of the information embedded during pre-training in favour of the new samples and conclude that, for pre-trained models, additional training data might not always be helpful when temporal robustness is at play.

Figure 1: F1 Score on the evaluation sets along T5 fine-tuning.



6. Conclusions and Future work

In this work, we evaluate the robustness of several well-known PLMs to temporal variations in text classification. We find that generative PLMs outperform encoder-only alternatives such as RoBERTa (the task baseline) across most temporal splits. Additional analysis confirms our hypothesis that increased training on the downstream short-term objective can hinder the generalisation capabilities of the pre-trained model, despite the similarity of topical distributions across the different temporal splits. Our findings on the role of model size in this task are in line with previous results in the field, showing that a larger number of parameters does not prevent performance drop.

In the future, we intend to more thoroughly test and improve the temporal robustness of generative PLMs under a larger set of carefully controlled temporal splits. In particular, exploring methods to improve their zero and few-shot capabilities; such as prompt fine-tuning, chain-of-thought prompting, or informative sample selection, might effectively reduce the re-training requirements of these models. Alternatively, we regard unsupervised methods for identifying polarity shifts towards topics and named entities as a promising research direction towards building more temporally-aware sentiment classifiers.

Acknowledgments

This work was supported by UKRI, under grant number EP/S023437/1.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[2] A. Lazaridou, A. Kuncoro, E. Gribovskaya, D. Agrawal, A. Liska, T. Terzi, M. Gimenez, C. de Masson d'Autume, T. Kocisky, S. Ruder, et al., Mind the gap: Assessing temporal generalization in neural language models, Advances in Neural Information Processing Systems 34 (2021) 29348–29363.

[3] R. Alkhalifa, E. Kochkina, A. Zubiaga, Building for tomorrow: Assessing the temporal persistence of text classifiers, Information Processing & Management 60 (2023) 103200.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, CoRR abs/2005.14165 (2020). URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[6] R. Alkhalifa, I. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, G. Gonzalez-Saez, P. Galuščáková, L. Goeuriot, E. Kochkina, M. Liakata, D. Loureiro, H. Tayyar Madabushi, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Longeval: Longitudinal evaluation of model performance at clef 2023, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[8] J. Lukes, A. Søgaard, Sentiment analysis under temporal shift, in: Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis, 2018, pp. 65–71.

[9] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, Time of your hate: The challenge of time in hate speech detection on social media, Applied Sciences 10 (2020) 4180.

[10] O. Agarwal, A. Nenkova, Temporal effects on pre-trained models for language processing tasks, Transactions of the Association for Computational Linguistics 10 (2022) 904–921.

[11] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 1480–1489.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are

unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[13] Y. He, J. Li, Y. Song, M. He, H. Peng, Time-evolving text classification with deep neural networks, in: IJCAI International Joint Conference on Artificial Intelligence, 2018, p. 2241.

[14] X. Jin, D. Zhang, H. Zhu, W. Xiao, S.-W. Li, X. Wei, A. Arnold, X. Ren, Lifelong pretraining: Continually adapting language models to emerging corpora, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 4764–4780.

[15] F.-K. Sun, C.-H. Ho, H.-Y. Lee, Lamol: Language modeling for lifelong language learning, in: International Conference on Learning Representations, 2019.

[16] R. Alkhalifa, E. Kochkina, A. Zubiaga, Opinions are made to be changed: Temporally adaptive stance classification, in: Proceedings of the 2021 workshop on open challenges in online social networks, 2021, pp. 27–32.

[17] B. Dhingra, J. R. Cole, J. M. Eisenschlos, D. Gillick, J. Eisenstein, W. W. Cohen, Time-aware language models as temporal knowledge bases, Transactions of the Association for Computational Linguistics 10 (2022) 257–273.

[18] G. D. Rosin, I. Guy, K. Radinsky, Time masking for temporal language models, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 833–841.

[19] P. Röttger, J. Pierrehumbert, Temporal adaptation of bert and performance on downstream document classification: Insights from social media, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2400–2412.

[20] V. Hofmann, J. Pierrehumbert, H. Schütze, Dynamic contextualized word embeddings, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 6970–6984.

[21] G. Rosin, K. Radinsky, Temporal attention for language models, in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 1498–1508.

[22] W. Yin, R. Alkhalifa, A. Zubiaga, The emojification of sentiment on social media: Collection and analysis of a longitudinal twitter sentiment dataset, CoRR abs/2108.13898 (2021). URL: https://arxiv.org/abs/2108.13898. arXiv:2108.13898.

[23] H. Tayyar Madabushi, E. Kochkina, M. Castelle, Cost-sensitive BERT for generalisable sentence classification on imbalanced data, in: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 125–134. URL: https://aclanthology.org/D19-5018. doi:10.18653/v1/D19-5018.

[24] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (1945) 80–83. URL: http://www.jstor.org/stable/3001968.

[25] D. Antypas, A. Ushio, J. Camacho-Collados, V. Silva, L. Neves, F. Barbieri, Twitter topic classification, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3386–3400. URL: https://aclanthology.org/2022.coling-1.299.

A. Appendix

GPT-3 Prompts

Prompt for zero-shot classification.

```
Please label the following tweet as either positive or negative.
-Tweet: < Text_sample>
-Label:
```

Prompt for few-shot (15 examples) classification.

```
Please label the following tweet as either Positive or Negative.
-Tweet: Where else can I get a pumpkin coffee in the morning because DD does not have their shit together.
-Label: negative
-Tweet: my mentions though  I couldn't be more grateful for what happened tonight I love Shawn so much
-Label: negative
-Tweet: I sneezed on the beat and the beat got sicker
-Label: positive
-Tweet: @MENTION Hi Niall, I hope you're fine, I love you so much! Thanks for everything.
   You make me smile  can you follow me please? 1450
-Label: positive
-Tweet: @MENTION Are You bored by listening pop and radio music ?
    Join us, This Channel let's You discover new emotions in each track
-Label: positive
-Tweet: So it smells of weed in the car.
Pretty sure it's coming from a certain someone's reading rucksack  @MENTION
-Label: positive
-Tweet: hate to see a good guy get fucked over  like wyd girl
-Label: negative
-Tweet: When Nike leaves the security tag on your shoes and you back to get it off,
    and the alarm goes off, but didn't go off when you left.
-Label: negative
-Tweet: Coz' I was born for you.. @MENTION  @MENTION  HAH joke..
-Label: positive
-Tweet: Can't fault her the last nigga spoiled her.
-Label: positive
-Tweet: Trying to have the #bun #hairstyle. Gotta wait a few months but the wait gonna be worth it
-Label: positive
-Tweet: Fan was soooo AMAZING Loved it You MUST watch it @MENTION was again the BEST actor in the world
-Label: positive
-Tweet: Bout to take a nap n then wake up n do some with my life!!
-Label: positive
-Tweet: Send emojis for a tbh because I'm bored asf
-Label: negative
-Tweet: Pahabol essays and hw are the worst
-Label: negative
-Tweet: < Text_sample>
-Label:
```