Temporal Adaptation of BERT and Performance on Downstream Document Classification: Insights from Social Media

Paul Röttger and **Janet B. Pierrehumbert**

University of Oxford

paul.rottger@oii.ox.ac.uk

Abstract

Language use differs between domains and even within a domain, language use changes over time. For pre-trained language models like BERT, domain adaptation through continued pre-training has been shown to improve performance on in-domain downstream tasks. In this article, we investigate whether temporal adaptation can bring additional benefits. For this purpose, we introduce a corpus of social media comments sampled over three years. It contains unlabelled data for adaptation and evaluation on an upstream masked language modelling task as well as labelled data for fine-tuning and evaluation on a downstream document classification task. We find that temporality matters for both tasks: temporal adaptation improves upstream and temporal fine-tuning downstream task performance. Time-specific models generally perform better on past than on future test sets, which matches evidence on the bursty usage of topical words. However, adapting BERT to time and domain does not improve performance on the downstream task over only adapting to domain. Token-level analysis shows that temporal adaptation captures event-driven changes in language use in the downstream task, but not those changes that are actually relevant to task performance. Based on our findings, we discuss when temporal adaptation may be more effective.

1 Introduction

Language use differs between domains and even within a domain, language use changes over time. In different domains, different communities share different social experiences as well as topical interests and thus produce different language (Church and Gale, 1995; Blei et al., 2003). At different times, some topics are discussed more actively while others fade into the background (Church, 2000; Altmann et al., 2009; Pierrehumbert, 2012). For NLP tasks, model performance therefore depends at least in part on how training and test data align in terms of domain and temporality. Sentiment analysis models trained on film reviews, for example, perform worse on restaurant reviews (Liu et al., 2019). Similarly, gender and age prediction models trained on one year's data perform increasingly worse on later years (Jaidka et al., 2018).

The widespread use of pre-trained language models like BERT (Devlin et al., 2019) motivates additional considerations about data selection. Such models are first trained *upstream* on large unlabelled corpora to learn general-purpose language representations (*pre-training*) before labelled task data is introduced *downstream* in a separate training phase (*fine-tuning*). In this setting, the choice of unlabelled pre-training data influences downstream model performance like the choice of labelled fine-tuning data does. In particular, we know that *domain* information, i.e. *where* pre-training data is sampled from, is highly relevant for downstream tasks. Domain *adaptation*, i.e. additional pre-training of an already-pre-trained model on domain data, has been shown to improve performance on a wide variety of in-domain downstream tasks (e.g. Gururangan et al., 2020). By contrast, there is little insight so far into the relevance of *temporality* in pre-training, i.e. *when* pre-training data is sampled from, as it relates to downstream tasks.

In this article, we work towards closing this research gap by investigating whether adapting BERT to time and domain can improve performance on a downstream document classification task relative to only adapting to domain. Our hypothesis is that temporal adaptation can capture changes in language use such as topical shifts that are relevant to the downstream task, which time-agnostic domain adaptation cannot account for.

To enable our analysis, we introduce a benchmark corpus of English-language text comments sampled from the social media site Reddit over three years. The corpus, which we call the Reddit Time Corpus (RTC), consists of a large set of un-

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2400–2412
November 7–11, 2021. ©2021 Association for Computational Linguistics

labelled comments for adaptation and evaluation on an upstream masked language modelling task (MLM), and a smaller set of labelled comments for fine-tuning and evaluation on a downstream five-way document classification task, which we call Political Subreddit Prediction (PSP).

We use RTC and a pre-trained BERT model to conduct a series of experiments on the upstream MLM and downstream PSP task (Figure 1). For MLM, we evaluate scale effects in domain adaptation (**DAda**) relative to no adaptation (**NAda**) as well as the effects of temporal adaptation (**TAda**). For PSP, we evaluate scale effects in **DAda** in relation to regular fine-tuning (**RFt**) as well as the effects of temporal fine-tuning (**TFt**). Lastly, we compare PSP performance across all six combinations of these adaptation and fine-tuning strategies (e.g. **TAda+TFt**). Overall, we find that temporal information matters for both tasks. **TAda** improves MLM performance and **TFt** improves PSP performance. **DAda** beats **NAda** on MLM and PSP. However, we do not find clear evidence that **TAda** outperforms **DAda** on PSP. More granular analysis suggests that this is because the event-driven changes in language use captured by **TAda** are not discriminative, i.e. relevant, for the PSP task.[1]

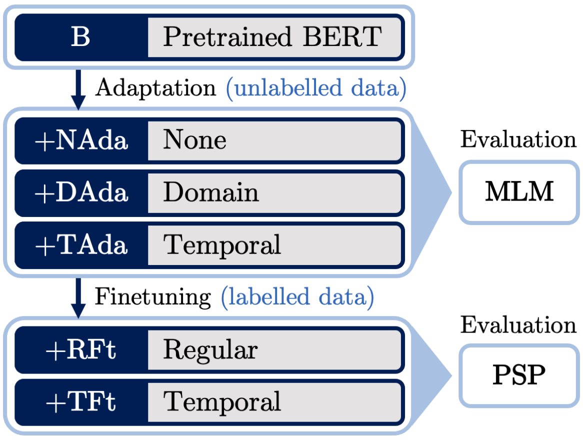

Figure 1: Schematic of our experimental setup. BERT is first adapted in one of three ways using unlabelled data and evaluated upstream on a masked language modelling task (MLM). Either adapted model is then fine-tuned in one of two ways using labelled data and evaluated downstream on Political Subreddit Prediction (PSP), a five-way document classification task.

2 Related Work

Previous work shows that models trained on texts from one time period perform increasingly worse on later time periods for a wide variety of tasks such as review and news article classification (Huang and Paul, 2018, 2019), gender and age prediction (Jaidka et al., 2018), sentiment analysis (Lukes and Søgaard, 2018) and hate speech detection (Nobata et al., 2016; Florio et al., 2020). However, such work has generally not used pre-trained models (e.g. Jaidka et al., 2018) and even if they are used, training and evaluation focuses on labelled task data alone (e.g. Florio et al., 2020). By contrast, our analysis aims to investigate the effects of unsupervised temporal adaptation in pre-training on downstream task performance.

Within the current paradigm of using pre-trained language models, research has focused more on the domain of pre-training data than its temporality. BERT and its variants have been pre-trained from scratch on in-domain data to improve performance on tasks such as hate speech detection (Tran et al., 2020), as well as tasks in scientific (Beltagy et al., 2019), clinical (Huang et al., 2019) and legal NLP (Zheng et al., 2021). Further, Gururangan et al. (2020) demonstrate that domain adaptation, a second phase of pre-training on in-domain data, similarly improves performance on in-domain downstream tasks (see also Alsentzer et al., 2019; Chakrabarty et al., 2019; Lee et al., 2020). We use their approach to domain adaptation as a baseline and extend it to temporality.

Incorporating temporal information in model pre-training has so far received little attention. Literature on diachronic embeddings for capturing temporal semantic change (e.g. Hamilton et al., 2016b; Rudolph and Blei, 2018; Tsakalidis and Liakata, 2020) is closely related, but mostly concerned with learning representations across a known time span and investigating their dynamics. Hofmann et al. (2021a) jointly model social and temporal information across time periods using a BERT model, showing that this improves performance on MLM and sentiment analysis. However, they do not evaluate task performance across time periods. By contrast, we adapt BERT to specific time periods with the aim of improving performance on a downstream task located in time. More directly related to our approach, Lazaridou et al. (2021) train autoregressive, left-to-right transformer models from scratch on unlabelled data sam-

[1] We make our code available on https://github.com/paul-rottger/temporal-adaptation.

pled up to a specific point in time and then evaluate them on a language modelling task using later data. They find that performance degrades over time and demonstrate that dynamic evaluation (Krause et al., 2019), a form of unsupervised online learning, mitigates this degradation. By contrast, the BERT models we use learn representations through MLM and are adapted to specific time periods, which is less computationally expensive than pre-training from scratch. Most importantly, we go beyond masked language modelling and evaluate the effects of temporal adaptation on a downstream document classification task, which is a more practically relevant use case of pre-trained language models.

3 Experiments

3.1 Data: Reddit Time Corpus

The Reddit Time Corpus (RTC) covers three years between March 2017 and February 2020 and is split into 36 evenly-sized monthly subsets based on comment timestamps. RTC is sampled from the Pushshift Reddit dataset published by Baumgartner et al. (2020). We provide a data statement (Bender and Friedman, 2018) for RTC in Appendix A.

Adaptation: Unlabelled News Comments We collect comments from *r/news* and *r/worldnews*, two of the most-subscribed and most active subreddits (i.e. discussion forums) on Reddit. *r/news* is primarily focused on US news content while *r/worldnews* describes itself as a "place for major news from around the world, excluding US-internal news". Both subreddits explicitly forbid overtly partisan posts in their community rules. For each of 36 months in our analysis, we sample one million comments, half from each of the two subreddits, for model adaptation. In total, we sample 36 million news comments.

Fine-Tuning: Labelled Politics Comments We collect comments from five subreddits for political discussion: *r/the_donald*, *r/libertarian*, *r/conservative*, *r/politics* and *r/chapotraphouse*. For each of 36 months in our analysis, we sample 25,000 comments at equal proportions across these subreddits and label them by the subreddit they were posted to, to create a balanced five-way classification task with equal class distribution across months, which we call Political Subreddit Prediction (PSP). 20,000 comments are used for model fine-tuning. 5,000 comments are used for evaluation, with labels for PSP and without for MLM. In total, we sample 0.9 million politics comments.

The subreddits we chose for PSP generally correspond to different political ideologies. *r/the_donald* was a subreddit for supporters of then-US President Donald Trump. *r/chapotraphouse* was one of the most active leftist subreddits, which grew out of a popular podcast. Both subreddits were shut down by Reddit in June 2020 for hosting content that promoted hate and violence. *r/conservative* and *r/libertarian* are subreddits for discussing conservative and libertarian politics. *r/politics* is not explicitly ideological but its subscribers tend to be liberal-leaning (Marchal, 2020). We thus expect distinctions between subreddits in PSP to be at least partially predictable based on comment text for two reasons: First, because language use differs between subreddits (Del Tredici and Fernández, 2017). Second, because distinguishing between political subreddits can be seen as a proxy for text-based ideology prediction, which is a well-established NLP task (e.g. Conover et al., 2011; Iyyer et al., 2014; Kannangara, 2018; Xiao et al., 2020).

Since both the labelled and the unlabelled comments in RTC are sampled from the same platform, we would expect some particular degree of similarity in language use between them. Based on Jaccard similarity of their vocabularies, comments from different politics subreddits are about as similar to each other as they are to comments from the news subreddits (Table 1). Comments from all subreddits are also more similar to each other than to paragraphs from the BooksCorpus (Zhu et al., 2015) that was used for pre-training BERT, along with English Wikipedia content. This motivates our use of news comments for domain adaptation. Further, we know that topical shifts, particularly those due to exogenous events, can drive changes in language use in both news and politics comments. For instance, Donald Trump's impeachment in December 2019 was immediately and actively discussed in news as well as politics subreddits. This motivates our use of monthly subsets of news comments for adapting models to both domain and time.

Pre-Processing During sampling, we restrict RTC to English-language comments using the `langdetect` Python library. We replace URLs and emojis with [URL] and [EMOJI] tokens, remove line breaks and collapse white space. We remove comments posted by bots, which we identified heuristically. We also remove comments that

	LIB	CTH	CON	POL	T_D	NWN	BC
LIB	1.00	0.42	0.48	0.47	0.44	0.46	0.33
CTH	0.42	1.00	0.43	0.43	0.42	0.42	0.33
CON	0.48	0.43	1.00	0.48	0.46	0.46	0.34
POL	0.47	0.43	0.48	1.00	0.45	0.46	0.34
T_D	0.44	0.42	0.46	0.45	1.00	0.44	0.34
NWN	0.46	0.42	0.46	0.46	0.44	1.00	0.34
BC	0.33	0.33	0.34	0.34	0.34	0.34	1.00

Table 1: Jaccard similarity between vocabularies for random sets of comments (n = 50k) from the five political subreddits (LIB, CTH, CON, POL, T_D) as well as the union of the two news subreddits (NWN) in RTC and a random sample of paragraphs (n = 50k) from the BooksCorpus (BC) used in BERT's pre-training.

users have deleted from Reddit and drop duplicates within each monthly subset of the corpus.

3.2 Model Architecture: BERT

We use uncased BERT-base (Devlin et al., 2019) for all experiments. For adapting to unlabelled news comments, we initialise BERT with default pre-trained weights and then continue pre-training on the MLM objective for one epoch, i.e. one pass over all additional data. For fine-tuning on labelled politics comments, we add a linear layer with softmax output and train for three epochs. Further details on model training and parameters as well as implementation can be found in Appendix B.

3.3 Upstream Task: MLM

Scale Effects in Domain Adaptation First, we evaluate the relative advantage of adapting BERT to domain using unlabelled news comments (**B+DAda**) and the extent to which this advantage scales with the amount of adaptation data. To eliminate temporal effects, news comments for adaptation and evaluation are sampled in equal proportions across all 36 months in RTC. As an evaluation metric, we report pseudo-perplexity, which we calculate as the exponential of the average cross-entropy loss across masked tokens (Table 2).

MLM performance clearly benefits from adapting to domain. Pseudo-perplexity on the test set decreases by 57.22%, from 19.54 to 8.36, for **B+DAda** with one million news comments compared to **B+NAda**. Performance further improves with the amount of adaptation data, although incremental improvements are diminishing.

Adaptation Data	Pseudo-Perplexity
0 (= **B+NAda**)	19.54
1 million	8.36
2 million	7.77
5 million	7.10
10 million	6.62

Table 2: Pseudo-perplexity of **B+DAda** on overall MLM test set (n = 5k unlabelled politics comments) for different amounts of adaptation data.

Temporal Adaptation Second, we introduce temporality by adapting to and evaluating on comments sampled from specific months. We adapt pre-trained BERT to one million news comments from each month in RTC, which yields 36 models (**B+TAda**). We then evaluate each month-adapted model on each monthly test set of 5,000 politics comments, so that in total we perform 1,296 evaluations. Pseudo-perplexity is comparable between models on the same test set but not between different test sets. Thus, we report percentage differences in pseudo-perplexity relative to the pseudo-perplexity of a domain-adapted control model (**B+DAda** with one million news comments) on a given test set. For readability, Table 3 shows results for every fourth month.

Adapt.	17-04	17-08	17-12	18-04	18-08	18-12	19-04	19-08	19-12
17-04	-0.56	0.53	2.10	3.24	4.29	5.37	4.34	5.19	5.74
17-08	0.52	-1.62	1.96	2.89	1.98	4.58	4.05	3.20	4.87
17-12	1.48	0.42	-0.87	1.05	1.50	2.61	2.26	2.65	2.24
18-04	1.14	0.69	1.82	-0.95	0.98	2.47	2.14	2.67	2.54
18-08	1.19	-0.20	-0.06	0.34	-1.12	0.38	1.38	0.81	1.38
18-12	0.79	0.78	0.90	0.69	0.08	-1.03	0.98	1.14	1.34
19-04	2.01	1.09	1.64	1.28	0.18	0.62	-0.66	1.10	-0.19
19-08	3.05	2.06	2.21	1.54	1.85	2.08	1.26	-0.12	1.22
19-12	2.04	1.00	1.47	1.82	1.26	1.23	-0.18	0.70	-2.36

Table 3: % difference in pseudo-perplexity of month-adapted models (**B+TAda**) relative to the control model (**B+DAda**). Rows correspond to adaptation sets (n = 1m unlabelled news comments), columns to test sets (n = 5k unlabelled politics comments).

For each monthly test set of politics comments, the best-performing model is the one adapted to news comments from the same month. When adaptation month matches test month, **TAda** outperforms **DAda** by 1.03% on average. For other months, **TAda** generally performs worse than **DAda**. As the temporal distance between adaptation month and test month increases, **TAda**'s performance decreases. Lastly, relative to the month they were adapted to, models generally perform better on past than on future test data. A one-sided Wilcoxon signed-rank test of all pairs of matching off-diagonal results (e.g. for a model adapted to 17-12, tested on 18-12 vs. a model adapted to 18-12, tested on 17-12) confirms that this finding is highly significant ($p < 0.001$).

Token-Level Analysis To further investigate why **TAda** outperforms **DAda** on MLM when adaptation month matches test month, we analyse changes in cross-entropy loss on individual masked tokens and how they contribute to the overall performance improvement.[2] Since we would expect dynamics of language use to vary between word classes, we use part-of-speech (POS) tags to structure our analysis. Specifically, we apply the `spaCy` POS tagger to all 36 test sets, link the tags to the WordPiece tokens generated by BERT and then compare performance improvements on masked WordPiece tokens by POS tag (Figure 2).

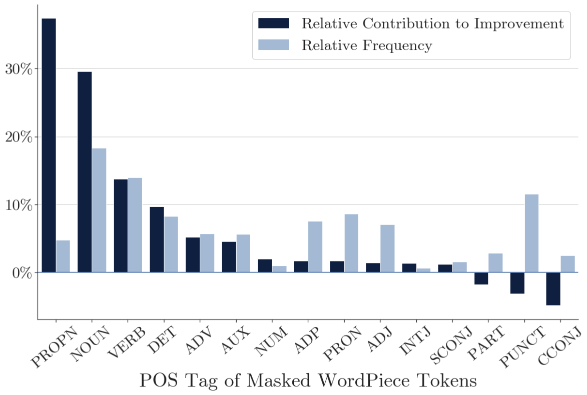

Figure 2: Relative contribution to reduction in cross-entropy loss, **B+TAda** over **B+DAda** on MLM, and relative frequency of masked tokens ($n \approx$ 1m) by POS.

Masked tokens in proper and common nouns drive 67.09% of the overall performance improvement. The former in particular contribute disproportionately much (37.46%) despite making up just 4.76% of all masked tokens. The contribution of tokens in other open-class words like verbs and adverbs roughly matches their frequency. By contrast, tokens in closed-class words such as conjugations contribute disproportionately little.

Qualitative analysis of those tokens in proper nouns for which cross-entropy loss was reduced the most from **TAda** over **DAda** suggests that **TAda** was most effective in capturing event-driven changes in topical language use (Table 4). These changes are generally bursty. The WordPiece "##ugh" as in "Kavana**ugh**", for example, was not used as part of a proper noun in any 2017 test set. Its use peaked when Kavanaugh was proposed for the US Supreme Court in September 2018 (107/5,000 test comments) and confirmed the month after (67/5,000). After December 2018, it was used at most nine times per test month.

Proper Noun	Time	Event
2019-**nCov**	20-02	WHO Covid press conference
Rex Till**erson**	18-03	Fired by Trump
Aziz Ansari	18-01	Abuse allegations
Kim Foxx	19-04	Prosecuting Jussie Smollet case
Liz **Warren**	19-11	Presidential run
Moscow	19-08	Trump: "Moscow Mitch"
Tide pods	18-02	Meme about eating them
Cville	17-08	"Unite the Right" rally
Ciaram**ella**	19-11	Revealed as CIA whistleblower
Kavana**ugh**	18-10	Supreme Court confirmation

Table 4: Top ten most-improved masked tokens (**bold**) in proper nouns from **TAda** over **DAda**, the test month the tokens are from and the event they correspond to.

3.4 Downstream Task: PSP

Scale Effects in Adaptation and Fine-tuning First, we evaluate relative scale effects in domain adaptation (**DAda**) and regular fine-tuning (**RFt**). To eliminate temporal effects, we sample news comments for adaptation as well as politics comments for fine-tuning and evaluation in equal proportions across all 36 months in RTC. Table 5 reports macro F1 on a scale from 0 to 100. Since there are five balanced classes in PSP, random choice would yield an expected macro F1 of 20.

Adapt.	1k	2k	5k	10k	20k	40k	80k	160k	320k
NAda	34.19	35.70	39.22	41.65	43.22	44.65	46.65	47.92	49.44
1m	37.76	39.26	41.31	42.91	44.03	45.18	46.87	47.94	49.51
2m	38.45	39.57	42.05	42.69	43.43	45.39	46.93	48.38	48.98
5m	38.78	39.84	42.16	43.42	44.46	45.86	47.37	47.93	49.82
10m	38.70	40.37	42.40	43.47	44.53	45.84	47.23	48.44	50.05

Table 5: Macro F1 of **B+DAda+RFt** on overall PSP test set (n = 10k labelled politics comments). Rows correspond to different amounts of adaptation data (unlabelled news comments), columns to different amounts of fine-tuning data (labelled politics comments).

Performance monotonically increases with the amount of politics comments used for **RFt**. Even

[2] BERT uses a WordPiece vocabulary. Each token is an instance of a WordPiece, which may be a word or sub-word.

for large amounts of fine-tuning data, there is no clear sign of a plateau. **DAda** using news comments is relatively more effective when there is less fine-tuning data. Its effectiveness moderately scales with the amount of adaptation data, but the biggest difference in performance is between the non-adapted model (**NAda**) and the model adapted using one million news comments, particularly for smaller amounts of fine-tuning data.

Temporal Fine-Tuning Second, we introduce temporality to PSP by fine-tuning and evaluating models on labelled politics comments from specific months (**TFt**). We fine-tune a pre-trained, non-adapted BERT model using 20,000 politics comments from each month in RTC, which yields 36 models (**B+NAda+TFt**). We then evaluate each month-tuned model on each monthly test set of 5,000 politics comments. Just like pseudo-perplexity in MLM, macro F1 on PSP is comparable between models on the same test set but not between different test sets. Therefore, we report percentage differences in macro F1 relative to the macro F1 of a control model with regular fine-tuning (**B+NAda+RFt** with 20,000 politics comments) on a given test set. For readability, we report results only for every fourth month in Table 6.

Finetune	17-04	17-08	17-12	18-04	18-08	18-12	19-04	19-08	19-12
17-04	6.81	0.48	-0.23	-1.13	-5.83	-1.32	-4.47	-8.58	-8.32
17-08	-1.20	6.00	-0.32	0.67	-3.15	0.28	-2.77	-5.02	-5.23
17-12	-4.84	1.16	4.95	-2.76	-0.81	-1.03	-5.16	-5.80	-2.63
18-04	-2.76	-1.72	-0.14	3.60	-0.04	0.37	-2.93	-5.14	-6.28
18-08	-2.71	-0.54	-0.66	-0.99	1.35	1.70	-2.89	-3.07	-5.70
18-12	-3.91	-1.64	-3.53	-0.89	-0.70	6.01	-0.79	-2.85	-5.22
19-04	-5.89	-4.82	-4.72	-2.96	-1.51	-0.49	6.12	-0.58	1.57
19-08	-10.82	-6.43	-7.52	-3.62	-4.62	-2.13	-1.10	3.88	-1.78
19-12	-10.31	-7.02	-5.82	-5.35	-4.88	-1.93	-4.75	0.92	3.03

Table 6: % difference in macro F1 of month-tuned models (**B+NAda+TFt**) relative to the control model (**B+NAda+RFt**). Rows correspond to fine-tuning sets (n = 20k labelled politics comments), columns to test sets (n = 5k labelled politics comments).

Overall, the results for month-tuned **TFt** models on PSP resemble those for month-adapted **TAda** models on MLM (Table 3). The best-performing model on a given test month is the one fine-tuned on politics comments from that month. When fine-tuning month matches test month, **TFt** outperforms **RFt** by 5.09% on average. For other months, **TFt** generally performs worse than **RFt**, although there are some exceptions when fine-tuning and test month are not far apart. For instance, **TFt** models on average perform 1.11% better than the **RFt** model on the test month directly after their fine-tuning month. As temporal distance between fine-tuning and test month grows, the performance of **TFt** models generally worsens. Models generally perform better on past than on future test data relative to the month they were fine-tuned on. A one-sided Wilcoxon signed-rank test of all pairs of matching off-diagonal results confirms that this finding is highly significant ($p < 0.001$).

Adaptation and Downstream Effects Third, we compare PSP performance across all six combinations of adaptation (**NAda**, **DAda**, **TAda**) and fine-tuning strategies (**RFt**, **TFt**). Our main interest is in evaluating whether **TAda** provides additional performance benefits on PSP compared to **DAda**. As an evaluation metric, we report average macro F1 across all 36 monthly PSP test sets. For models that incorporate temporality in adaptation (**TAda**) and/or fine-tuning (**TFt**), we consider those where adaptation and/or fine-tuning month matches the test month. Given that we found adaptation to be more effective for smaller amounts of fine-tuning data (Table 5), we report results for fine-tuning sizes of 2,000 and 20,000 (Table 7).

	2k	**20k**
B+NAda+RFt	35.95	43.21
B+DAda+RFt	**39.11**	**43.84**
B+TAda+RFt	39.01	43.81
B+NAda+TFt	37.59	45.41
B+DAda+TFt	40.19	46.02
B+TAda+TFt	**40.38**	**46.12**

Table 7: Average macro F1 across all 36 monthly PSP test sets (n = 180k labelled politics comments) for the six main model configurations, split between models using **RFt** and **TFt**. Best performance is **bold**. Columns correspond to different amounts of labelled politics comments used for fine-tuning.

Our central finding is that models adapted to time and domain (**TAda**) show no clear performance improvement over models adapted to just domain (**DAda**). Macro F1 is marginally higher for **B+TAda+TFt** than **B+DAda+TFt**, but marginally lower for **B+TAda+RFt** than **B+DAda+RFt**. Further, we find that **DAda** outperforms **NAda** and that **DAda** is more beneficial for models fine-tuned on less data, which matches results from Table 5.

MLM Improvements and PSP Performance Finally, we investigate why the benefits of **TAda** over **DAda** on MLM (Table 3) did not manifest in better performance on the downstream PSP task. For this purpose, we focus on masked tokens in proper nouns, which we identified as the main driver of **TAda**'s MLM improvements (Figure 2).

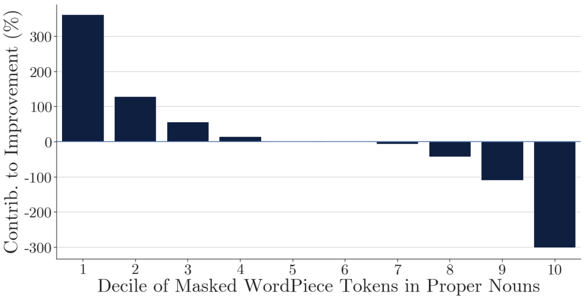

Figure 3: Relative contribution to reduction in cross-entropy loss, **B+TAda** over **B+DAda** on MLM, for masked tokens in proper nouns (n = 48,107) by decile.

Figure 3 shows that **TAda** improvements in MLM performance on masked tokens in proper nouns are overwhelmingly driven by the top 10% most-improved tokens (e.g. "2019-**n**Cov"), which we found to closely map onto exogenous news events (Table 4). We relate this set of most-improved tokens to PSP as follows: For each token, we take the WordPiece it is an instance of and the PSP test set the comment with the token is from. In that test set, we then count how many subreddits the WordPiece was used in and how many comments in each subreddit the WordPiece was used in, filtering only uses in proper nouns. For example, one of the most-improved tokens is "Kavana**ugh**" in October 2018. That month, the WordPiece "##ugh" was used in a proper noun in 67 out of 5,000 test comments across all five politics subreddits.

Table 8 suggests that the tokens in proper nouns that drive MLM improvements of **TAda** are not relevant to the PSP task. First, most WordPieces corresponding to these tokens are not distinctive of individual subreddits, with 2,717 WordPieces (65.95%) used in more than one subreddit in a given test set. The more subreddits the WordPieces are used in, the more frequent they are overall and within each subreddit they are used in. Second, more distinctive WordPieces are much rarer. WordPieces that are used in fewer subreddits are used much less frequently overall and used less frequently in the subreddits that they do appear in. The 1,403 WordPieces (34.05%) that are used in just one subreddit in a given test set are used on average in just 1.28 comments. 1,156 WordPieces are used in just one comment.

Subs.	WordPs	Comments	Avg. Freq.
1	1,403	1,789	1.28
2	769	2,316	1.51
3	559	3,336	1.99
4	570	6,950	3.05
5	819	33,090	8.08

Table 8: Frequency measures for WordPieces corresponding to the top 10% most-improved tokens in proper nouns by **TAda** over **DAda** for MLM (n = 4,120 after deduplication). Grouping is by the number of subreddits (n = 5) that a given WordPiece was used in in a given PSP test set (n = 5k). Average frequency is calculated for the subreddits the WordPieces were used in (Avg. Freq. = Comments / WordP's / Subs.).

4 Discussion

4.1 Results

We find that **DAda** yields large performance improvements on upstream MLM (Table 2) and the downstream PSP document classification task (Table 7) when compared to **NAda**, which matches previous findings (Alsentzer et al., 2019; Chakrabarty et al., 2019; Lee et al., 2020; Gururangan et al., 2020). Further, **DAda** is more effective when there is little fine-tuning data (Table 5).

We also find that temporality matters for both MLM and PSP. For upstream MLM, **TAda** outperforms **DAda** when adaptation month matches test month (Table 3). For downstream PSP, **TFt** outperforms **RFt** when fine-tuning month matches test month (Table 6). For both tasks, model performance decreases as the temporal distance between (pre-)training and test set grows. These findings are consistent with previous evidence for MLM (Lazaridou et al., 2021) and other document classification tasks (e.g. Huang and Paul, 2018; Florio et al., 2020). The results also illustrate a trade-off between temporal specificity and generalisability across time periods, which mirrors an equivalent trade-off in domain adaptation (Gururangan et al., 2020). Further, relative to the month they were adapted to (Table 3) or fine-tuned on (Table 6), models perform significantly better on past than on future test sets for MLM and PSP. This matches evidence on the usage of topical words, which tend to occur in bursts, often triggered by an exogenous event, followed by a slower decay (Church, 2000; Altmann et al., 2009; Pierrehumbert, 2012).

Despite these positive results for the individual tasks, we cannot confirm that the benefits of **TAda** over **DAda**, which are evident in MLM, transfer downstream to PSP. **DAda** and **TAda** perform about equally well on PSP (Table 7). This holds for different fine-tuning strategies (**RFt** and **TFt**) and different amounts of fine-tuning data.

Several trivial explanations for this negative finding can be eliminated due to our systematic experimental approach. First, we know that the language used in news comments is informative for PSP, since **DAda** consistently outperforms **NAda** on PSP (Tables 5 and 7). Second, we know that discriminatory language cues for PSP change over time, since **TFt** consistently outperforms **RFt** when fine-tuning month matches test month, and since **TFt** performs increasingly worse as temporal distance between fine-tuning and test month increases (Table 6). Lastly, we know that **TAda**, which uses news comments, allows models to capture some changes in language use in politics comments, since for each monthly test set of politics comments in MLM, the best-performing model is the one adapted to news comments from that same month (Table 3). Therefore, we can conclude that the changes in language use in politics comments that are captured by **TAda** using MLM on news comments are by and large not the changes in discriminatory language cues that are relevant to PSP.

In our token-level analysis, we find that most of **TAda**'s improvements over **DAda** on MLM (Figure 2) are for masked tokens in nouns. Predictions on masked tokens in proper nouns improve disproportionately much, especially for tokens that directly correspond to bursty changes in topical language use driven by exogenous news events (Table 4). However, in relation to the PSP task, the WordPieces corresponding to these tokens generally appear non-discriminative, since most of them are used in several politics subreddits rather than just one (Table 8). Intuitively, many news events are not just relevant to one political ideology, although they may differ in the way they are framed (Card et al., 2015; Demszky et al., 2019; Hofmann et al., 2021b). In March 2018, for example, when Donald Trump fired his secretary of state Rex Tillerson, an *r/politics* user in the corresponding test set said they were "sympathetic" to him, while an *r/the_donald* user called him a "globalist cuck". Since **TAda** uses comments from news subreddits, it cannot easily capture such distinctive frames.

4.2 Promising Uses of Temporal Adaptation

Based on our findings for this particular application of **TAda**, we can formulate positive expectations about the circumstances in which **TAda** would likely be more effective.

First, we expect **TAda** to be more effective if it captured changes in language use that were more specific to individual classes in the downstream task, i.e. more discriminative. Such changes in language use would occur when an event is relevant to just one class or when the same event is relevant to different classes at different times. For instance, learning about a regional news event in adaptation would likely help a classifier distinguish between comments from regional news sites.

Second, we expect **TAda** to be more effective over longer time scales than the 36 months covered by RTC. News and politics are suitable domains for our analysis because topical shifts are visible on short time scales (Figure 2), but over decades and centuries rather than months and years, cultural shifts and linguistic drift add to shorter-term event-driven changes in language use (Hamilton et al., 2016a,b). For tasks based on long-term corpora, such as the Corpus of Historical American English (Davies, 2012) temporal adaptation would thus likely improve model performance.

Lastly, we may also expect **TAda** to be more effective if it used pre-training objectives that were more aligned with downstream tasks. Clark et al. (2020a) argue that there is an inherent mismatch between task-agnostic pre-training that uses masked tokens and fine-tuning that does not, which recent work on discriminative pre-training tries to resolve (Clark et al., 2020b). Future work could explore the use of such techniques for model adaptation.

Even in circumstances in which **TAda** is effective, researchers and practitioners will need to consider the performance trade-off between temporal specificity and generalisability across time periods. For example, when deploying a hate speech detection model for content moderation, performance on newly posted content is most important, and tailoring the model to the current month is desirable even at a cost to reduced performance on past months. However, for applications where temporality is less relevant, more heterogeneous training data sampled across months is preferable.

5 Conclusion

In this article, we investigated whether adapting a pre-trained BERT model to time and domain can increase its performance on a downstream document classification task compared to only adapting it to domain. Overall, we found no clear evidence for this. By devising a systematic experimental approach based on the novel RTC benchmark corpus, we showed that temporality is relevant for both upstream MLM and the downstream PSP document classification task. Temporal adaptation improved MLM performance and temporal fine-tuning improved PSP performance. Further, domain adaptation improved performance on both tasks. Time-specific models generally performed better on past than on future test sets for both tasks, which matches evidence on the bursty usage of topical words. However, the upstream benefits of temporal adaptation for MLM did not translate into better downstream performance on PSP compared to domain adaptation alone. Token-level analysis showed that temporal adaptation captured event-driven changes in language use in downstream task data, but not those changes that are relevant to performance on it. This suggests that temporal adaptation may well be effective for other tasks under circumstances we outlined, which future work could investigate.

Acknowledgments

Paul Röttger was funded by the German Academic Scholarship Foundation. Janet B. Pierrehumbert was supported by EPSRC Grant EP/T023333/1. We thank Hannah Rose Kirk for helpful comments and all reviewers for their constructive feedback.

Ethics Statement

Data Collection All data in RTC is sampled from the Pushshift Reddit dataset made publicly available by Baumgartner et al. (2020). This dataset, in turn, was collected via Reddit's own public API in line with the site's terms of service. Our use of this Reddit dataset was also approved by the University of Oxford's Central University Research Ethics Committee. Labels for politics comments in RTC were created from comment metadata, so that no manual annotation was necessary.

Data Characteristics We describe the characteristics of RTC in the main body of this paper and provide additional detail in a data statement (Bender and Friedman, 2018) in Appendix A. In particular, we highlight RTC's limited scope in terms of data source (Reddit) and language (English), which limits the generalisability of models trained on it.

Intended Use The intended use of temporal adaptation is as an alternative to existing strategies for continued pre-training, particularly domain adaptation. Our article explores a specific application of temporal adaptation using monthly sets of news comments for a downstream classification task of politics comments. Temporal adaptation could be applied to most other NLP tasks as long as pre-training and task data can be located in time, although the effectiveness of temporal adaptation may differ. Effective temporal adaptation stands to improve diachronic model performance and thus reduce error rates in real-world applications.

Potential Misuse As with domain adaptation, temporal adaptation creates a trade-off between specificity and generalisability. Models adapted to a particular time period and domain should not be used for other time periods and domains without careful consideration of resulting biases.

Environmental Impact Temporal adaptation is more computationally expensive than just fine-tuning using a (smaller) set of labelled task data but much less computationally expensive than pre-training from scratch on even larger unlabelled datasets. Relative to the concerns raised around the environmental costs of the latter (Strubell et al., 2019; Henderson et al., 2020; Bender et al., 2021), we consider the environmental costs of temporal adaptation to be relatively minor. In practical applications, researchers could consider cumulative approaches to temporal adaptation, rather than adapting separate models for each time period, to avoid redundant computations.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Eduardo G Altmann, Janet B Pierrehumbert, and Adilson E Motter. 2009. Beyond word frequency:

Bursts, lulls, and scaling in the temporal distributions of words. *PLOS one*, 4(11):e7678.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO fine-tuning improves claim detection. 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)*, pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenneth W. Church. 2000. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than $p2$. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Kenneth Ward Church and William A Gale. 1995. Poisson mixtures. *Nat. Lang. Eng.*, 1(2):163–190.

Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher D. Manning. 2020a. Pre-training transformers as energy-based cloze models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 285–294, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020b. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer. 2011. Predicting the political alignment of Twitter users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 192–199.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora*, 7(2):121–157.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. 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)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. 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.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. 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, Berlin, Germany. Association for Computational Linguistics.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021a. 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)*, pages 6970–6984, Online. Association for Computational Linguistics.

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2021b. Modeling ideological agenda setting and framing in polarized online groups with graph neural networks and structured sparsity. *CoRR*, abs/2104.08829.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. 2018. Diachronic degradation of language models: Insights from social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200. Association for Computational Linguistics.

Sandeepa Kannangara. 2018. Mining Twitter for fine-grained political opinion polarity classification, ideology detection and sarcasm detection. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 751–752.

Ben Krause, Emmanuel Kahembwe, Iain Murray, and Steve Renals. 2019. Dynamic evaluation of transformer language models. *CoRR*, abs/1904.08378.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Sebastian Ruder, Dani Yogatama, Kris Cao, Tomás Kociský, Susannah Young, and Phil Blunsom. 2021. Pitfalls of static language modelling. *CoRR*, abs/2102.01951.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ruijun Liu, Yuqian Shi, Changjiang Ji, and Ming Jia. 2019. A survey of sentiment analysis based on transfer learning. *IEEE Access*, 7:85401–85412.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Jan Lukes and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–71, Brussels, Belgium. Association for Computational Linguistics.

Nahema Marchal. 2020. The polarizing potential of intergroup affect in online political discussions: Evidence from Reddit r/Politics. *SSRN*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Janet B Pierrehumbert. 2012. Burstiness of verbs and derived nouns. In Diana Santos, Krister Linden, and Wanjiku Ng'ang'a, editors, *Shall We Play the Festschrift Game?*, pages 99–115. Springer.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1003–1011. International World Wide Web Conferences Steering Committee.

Statista. 2019. Reddit: Traffic by country.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Se Rim Park. 2020. HABERTOR: An efficient and effective deep hatespeech detector. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7486–7502, Online. Association for Computational Linguistics.

Adam Tsakalidis and Maria Liakata. 2020. Sequential modelling of the evolution of word representations for semantic change detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8485–8497, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. 2020. Timme: Twitter ideology-detection via multi-task multi-relational embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2258–2268.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Data Statement

Following Bender and Friedman (2018), we provide a data statement, which documents the generation and provenance of labelled and unlabelled documents in the Reddit Time Corpus (RTC).

A. CURATION RATIONALE The purpose of RTC is to enable our analysis of temporal adaptation of pre-trained language models and downstream task performance. RTC comprises text comments that were posted to the social media site Reddit between March 2017 and February 2020. The unlabelled portion of RTC consists of 36 million comments sampled from *r/news* and *r/worldnews*, two of the most active subreddits (i.e. sub-forums) on the site, which are dedicated to discussion of current news events. The labelled portion of RTC consists of 0.9 million comments sampled in equal proportions from five subreddits for political discussion (*r/the_donald*, *r/libertarian*, *r/conservative*, *r/politics* and *r/chapotraphouse*). Comments are labelled based on which subreddit they are from. All data is split into 36 evenly-sized subsets based on comment timestamps.

B. LANGUAGE VARIETY RTC only contains English-language text documents, as determined by the `langdetect` Python library. We opted for English language due to data availability. Further, all data in RTC is sourced from Reddit. We consider this a limitation of our analysis and suggest expansion to other languages and data sources as a priority for future research.

C. SPEAKER DEMOGRAPHICS The speakers in RTC are a sample of all Reddit users who posted a comment to one of the seven subreddits covered by RTC between March 2017 and February 2020. In February 2020, *r/worldnews* had around 23.1m subscribers, *r/news* 19.9m, *r/politics* 5.76m, *r/the_donald* 0.79m, *r/libertarian* 0.36m, *r/conservative* 0.30m and *r/chapotraphouse* 0.15m. Reddit does not make information on user demographics available but a February 2019 survey of US users indicated that roughly two-thirds were male, and that user age was skewed towards 18 to 29 years (Statista, 2019).

D. ANNOTATOR DEMOGRAPHICS We did not employ any annotators. All labels in RTC are based on comment metadata, specifically which subreddit a given comment is from.

E. SPEECH SITUATION All comments in RTC were posted to Reddit between March 1st 2017 and February 29th 2020. The intended audience is other subreddit users and site visitors.

F. TEXT CHARACTERISTICS All documents are individual text comments. Preprocessing steps are described in §3.1. For the labelled portion of RTC, we provide a label based on which of the five political subreddits in RTC they were posted to. The class distribution is balanced in RTC overall and in each monthly subset.

B Model Training & Parameters

Model Architecture We implemented uncased BERT-base models (Devlin et al., 2019) using the `transformers` Python library (Wolf et al., 2020). Uncased BERT-base, which is trained on lower-cased English text, has 12 layers, a hidden layer size of 768, 12 attention heads and a total of 110 million parameters. For PSP, we added a linear layer with softmax output.

Training Parameters For both MLM and PSP, we used cross-entropy loss. As an optimiser, we used AdamW (Loshchilov and Hutter, 2019) with a 5e-5 learning rate and a 0.01 weight decay. For regularisation, we set a 10% dropout probability. Maximum input sequence length is 128 tokens. For adapting to unlabelled data, we trained for one epoch, i.e. one pass over all additional data, which matches Gururangan et al. (2020). Training batch size was 128. For fine-tuning on labelled data, we trained for three epochs with a batch size of 32, which corresponds to default settings recommended by Devlin et al. (2019). For comparability, we used these same untuned hyperparameters across all experiments.

Computation All experiments were run between March and May 2021 using Nvidia Tesla K80 and V100 GPUs accessed through the University of Oxford's Advanced Research Computing service. Runtime varied from experiment to experiment. Adapting BERT to one million comments for one epoch took around three hours on a V100. Fine-tuning BERT on 20,000 comments for three epochs took around 15 minutes.

Source Code We make all our code available at https://github.com/paul-rottger/temporal-adaptation.