Towards Robust Sentiment Analysis of Temporally-Sensitive Policy-Related Online Text

Anonymous ACL submission

Abstract

Sentiment analysis in policy-related studies typically involves annotating a subset of data to fine-tune a pre-trained model, which is subsequently used to classify sentiments in the remaining unlabeled texts, enabling policy researchers to analyze sentiments in novel policy contexts under resource constraints. We argue that existing methods fail to adequately capture the temporal volatility inherent in policy-related sentiments, which are subject to external shocks and evolving discourse of opinions. We propose methods accounting for the temporal dynamics of policy-related texts. Specifically, we propose leveraging continuous time-series clustering to select data points for annotation based on temporal trends and subsequently apply model merging techniques – each fine-tuned separately on data from distinct time intervals. Our results indicate that continuous time-series clustering followed by fine-tuning a single unified model achieves superior performance, outperforming existing methods by an average F1-score of 2.71%. This suggests that language models can generalize to temporally sensitive texts when provided with temporally representative samples. Nevertheless, merging multiple time-specific models – particularly via greedy soup and TIES – achieves competitive performance, suggesting practical applications in dynamically evolving policy scenarios.

1 Introduction

Sentiment analysis in policy-related studies is often conducted using transfer learning on partially annotated datasets, where a subset of data is annotated and used to fine-tune a pre-trained model, subsequently employed to classify sentiments in the remaining unlabeled texts (An et al., 2023; Effrosynidis et al., 2022; Maceda et al., 2023; Melton et al., 2022). This allows policy researchers to systematically gauge public support (or opposition) toward policies from extensive online data, providing valuable insights to inform policy recommendations (Ceron and Negri, 2015). This approach enables researchers to leverage robust language models for sentiment classification even in novel policy contexts, where benchmarks datasets fail to adequately capture the evolving opinions or context-specific semantics associated with sentiments of emerging policies. For instance, terms like "Welfare Queen" may be associated with positivity among sentiments from benchmark datasets, but are considered derogatory in welfare policy contexts (Floyd-Thomas, 2016). Additionally, it helps overcome practical constraints such as limited resources, since annotating the entire dataset is often infeasible due to time and budgetary limitations.

We argue that these commonly employed methods fail to effectively capture the temporally-sensitive nature of sentiments associated with policy-related texts. Sentiments in such contexts are subject to volatile shifts, driven by factors such as external shocks which influence policy perception (Giuliano and Spilimbergo, 2024), the emergence of conflicting information over time (Dhingra et al., 2022) and the continuous introduction of new vocabulary or terminologies associated within evolving policy discourse (Alkhalifa et al., 2021; Azarbonyad et al., 2017). All these factors can alter the semantic context of underlying sentiments. Furthermore, temporal variations in online discourse often reflect shifts in public attention triggered by specific events or emerging issues, characterized by pronounced spikes or drops in online engagement (Yang and Leskovec, 2011).

These characteristics often lead to a non-uniform temporal distribution of trends surrounding online textual data. Pronounced fluctuations among sentiments from policy-related discourse could result in periods where texts are densely clustered around particular events or intervals. Consequently, random sampling for annotation is likely to disproportionately represent texts from these dense intervals, leaving other crucial periods sparsely annotated (Lazaridou et al., 2021). Such sampling bias impairs the generalizability of language models by limiting their exposure to representative texts and vocabulary, constraining their ability to adapt to evolving semantic contexts (Azarbonyad et al., 2017).

Hence, this study aims to leverage strategies in developing robust sentiment analysis models capable of generalizing across multiple time intervals, under realistic settings that mimic sentiment analysis in policy-related studies. We aim to integrate temporal aspects of policy-related online texts by (1) proposing continuous time-series clustering to segment the corpus timeline into variable-length clusters based on temporal trends, which yields a temporally representative training set for fine-tuning and (2) subsequently experimenting with advance merging methods to integrate multiple models – each fine-tuned separately on data from distinct time intervals – into a unified sentiment classifier.

We conduct extensive experiments on 3 benchmark datasets across 4 models, and demonstrate that continuous time-series clustering improves the average F1-score by 2.71% compared to random selection, benefitting from taking temporal shifts into account. Although certain merging techniques achieved competitive performance, it's overall performance deteriorated compared to the unified singular model finetuned across all time intervals. This suggests that language models can generalize to temporally volatile policy sentiments when fine-tuned on representative samples capturing meaningful semantic shifts in policy discourse.

Therefore, our contributions are as follows:

- We explicitly consider temporal trends of online texts by proposing continuous time-series clustering when sampling data for annotation and subsequent fine-tuning, thus accounting for fluctuations in online textual activity driven by external shocks and evolving discourse. Innovatively, our method incorporates aspects beyond purely textual considerations.

- We rigorously evaluate our methods on realistic policy-related datasets under settings closely resembling typical sentiment analysis tasks in policy studies. Our results hence provides practical insights for policy researchers regarding the expected effectiveness of our proposed approach.

- We rigorously explored advance model merging techniques to test their effectiveness in integrating models fine-tuned on distinct time intervals, despite observing an overall performance deterioration.

2 Related Works

2.1 Temporally-sensitive text classification

The limited ability of language models to generalize effectively across multiple time points has been extensively studied. Dhingra et al. (2022) attributes this limitation primarily to 'temporal staleness,' emphasizing that language models, typically trained on static data snapshots, fail to adapt adequately to temporal changes beyond their training snapshot, resulting in degraded performance. To address this, the authors propose prepending temporal information to the textual data.

Similarly, Lazaridou et al. (2021) observed that language models trained under static conditions consistently struggle to capture the dynamic and evolving nature of language. They further demonstrate that scaling models by using larger variants like TransformerXL, fails to remedy this limitation. However, their findings suggest that this limitation can be through sustained training across extensive time points.

Additionally, Röttger and Pierrehumbert (2021) demonstrated that fine-tuning an individual model for each month and testing it on the same month produced substantially better predictions than relying on a model fine-tuned with labeled data pooled across all time points when attempting to predict the political leaning of a given Reddit post. This demonstrates the pronounced temporal volatility of online texts with its associated downstream prediction and shortcomings of finetuned language models in generalizing across multiple time intervals.

2.2 Merging multiple time-specific models

To address temporal sensitivity in text classification, recent methods propose merging models fine-tuned on discrete intervals (e.g., months or years). Model merging essentially blends weights across multiple models to capture complementary knowledge without additional retraining or ensembling.

For instance, Nylund et al. (2024) proposed merging multiple fine-tuned models, each trained on distinct fixed intervals (e.g., individual months or years), through "model souping". However,

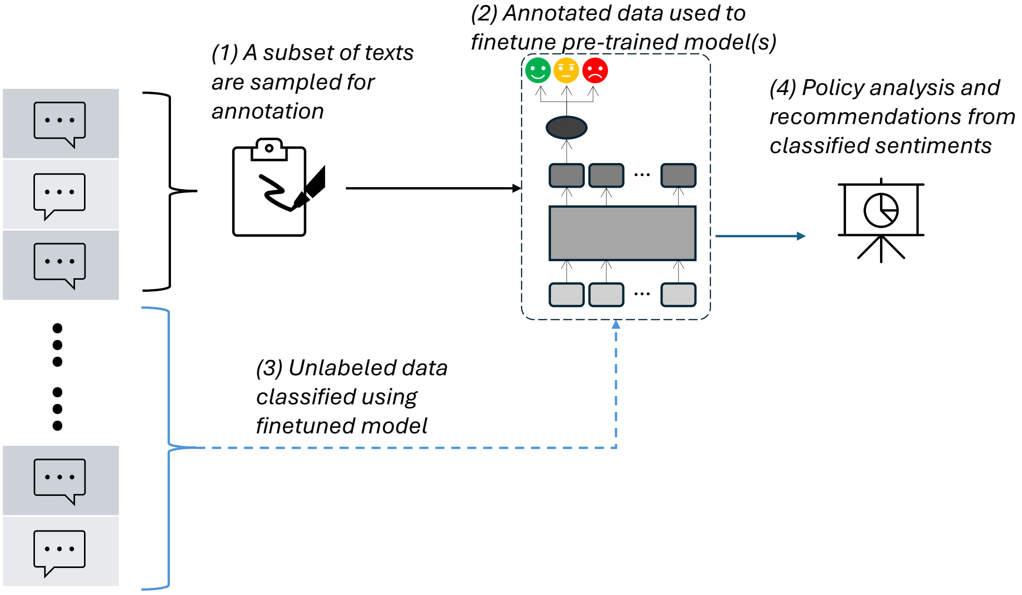

(1) A subset of texts are sampled for annotation
(2) Annotated data used to finetune pre-trained model(s)
(3) Unlabeled data classified using finetuned model
(4) Policy analysis and recommendations from classified sentiments

Figure 1: Typical sentiment analysis in policy-related studies, where sampled data is annotated and used to fine-tune a model, subsequently classifying unlabeled data. This approach is beneficial in novel policies, where benchmarks fail to capture the context-specific discourse associated with sentiments of emerging policies, and annotating the entire dataset is resource-prohibitive.

results showed that these merged models generally performs worse in generalizing across multiple time periods compared to a single model fine-tuned on labeled data from all intervals. Although interpolation between two time vectors successfully improved predictions for unknown intervals such as future or intervening periods, merging multiple fine-tuned models simultaneously via souping did not yield similar benefits, underscoring the challenge of improving generalization with unseen data spanning multiple temporal intervals.

Dziadzio et al. (2025) similarly addressed this issue in a streaming context using the Temporal Integration of Model Expertise (TIME) framework. At each interval, TIME initializes training from an exponential moving average (EMA) of prior checkpoints, fine-tunes on the current interval, then merges the newly trained expert back into the EMA. Although TIME outperformed standard continual fine-tuning and other merging methods, its sequential training assumption limits direct applicability to scenarios involving generalization across multiple intervals simultaneously. Nevertheless, TIME motivates us to explore intermediate processing steps rather than directly merging fixed-interval models (Nylund et al., 2024).

3 Methods

3.1 Selecting data points for annotation

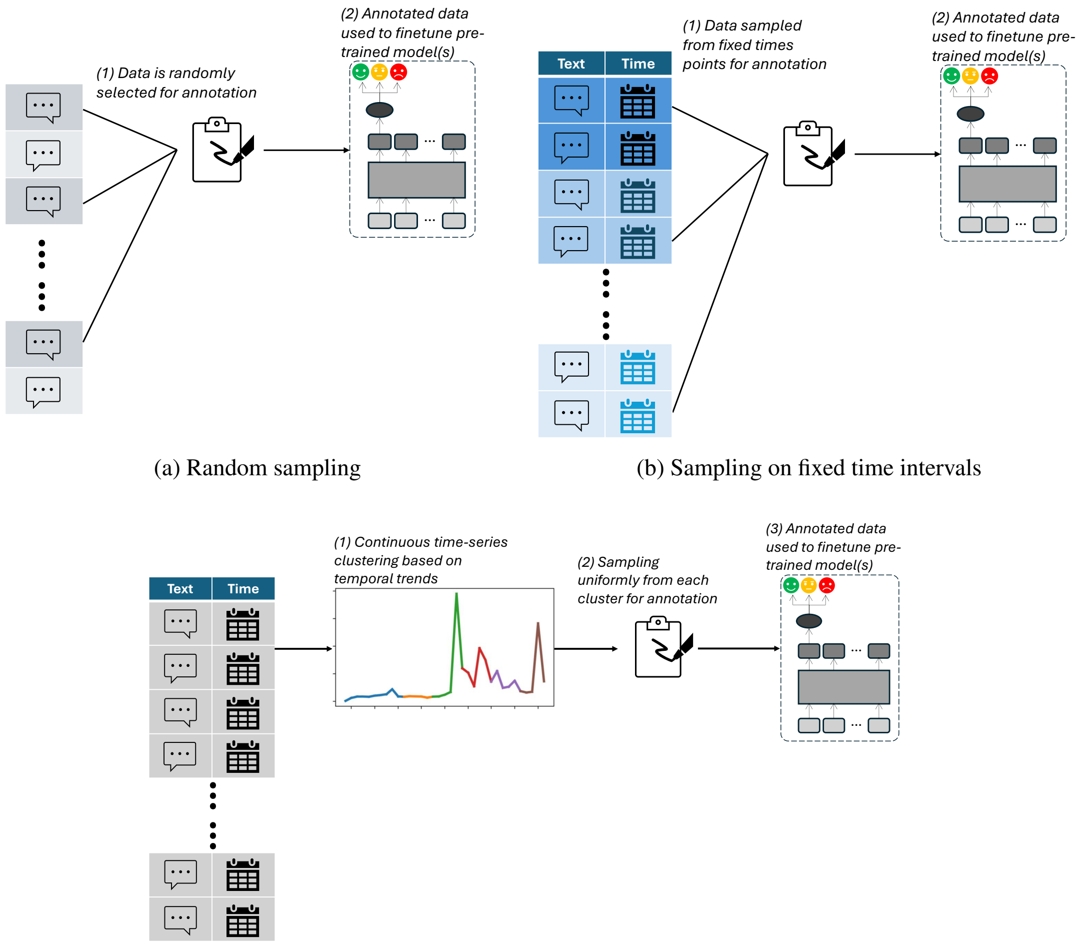
As illustrated in Figure 1, sentiment analysis in policy-related studies typically begins by sampling a subset of data points for professional annotation. These labeled data are subsequently used to fine-tune sentiment classification model(s).

Random Sampling The selection of data points for annotation is often randomly sampled, where a fixed number (n) of data points – determined based on factors such as the researcher's annotation budget or desired annotation volume – is drawn uniformly at random (without replacement) from the entire dataset (An et al., 2023; Hayawi et al., 2022; Hossain et al., 2020). This can be illustrated in Figure 2a.

Sampling Based on Fixed Time Intervals To account for the temporality inherent in online data, some studies propose uniformly sampling data points from each predefined fixed time interval t (e.g., monthly or yearly), where $n_t \approx \frac{n}{|\mathcal{T}|}$ for $t \in \mathcal{T}$ (Nylund et al., 2024; Röttger and Pierrehumbert, 2021; Dhingra et al., 2022), as illustrated in Figure 2b.

Sampling based on continuous time series clustering We propose employing continuous time-series clustering to sample data points from each

Codes accessible at the anonymous GitHub repository: https://github.com/anonAclSrw/tempsentpolicy


(1) Data is randomly selected for annotation
(2) Annotated data used to finetune pre-trained model(s)
Text
Time
(1) Data sampled from fixed times points for annotation
(2) Annotated data used to finetune pre-trained model(s)
Text
Time
(1) Continuous time-series clustering based on temporal trends
(2) Sampling uniformly from each cluster for annotation
(3) Annotated data used to finetune pre-trained model(s)

(a) Random sampling

(b) Sampling on fixed time intervals

(c) Sampling based on continuous time series clustering

Figure 2: The distinct strategies when selecting data points for annotation, which will subsequently be used to finetune a model to classify the sentiments of the remaining corpus.

identified cluster, as illustrated in Figure 2c. We utilize Ruptures (Truong et al., 2020), as it effectively detects structural shifts or change points in discrete time-series data, serving our overarching purpose of modeling temporal trends across online texts.

We begin by aggregating the entire corpus into a univariate count series $\mathbf{N} = (N_1, \ldots, N_T)$, where $N_t \in \mathbb{N}$ is the total number of policy-related texts (e.g., Tweets) observed in time bin t (e.g., day, month, or year). Ruptures then segments this series into contiguous clusters by locating change-points that minimize the penalized within-segment cost

$$\hat{\boldsymbol{\tau}} = \arg \min_{\boldsymbol{\tau} \subset \{1,\ldots,T-1\}} \Bigg\{ \underbrace{\sum_{k=0}^{|\boldsymbol{\tau}|} \mathcal{L}\big(N_{t_k+1:t_{k+1}}\big)}_{\text{segment-cost}} + \underbrace{\beta\,|\boldsymbol{\tau}|}_{\text{penalty}} \Bigg\}$$

where the segment-cost

$$\mathcal{L}(N_{a:b}) = \min_{\alpha,\gamma} \sum_{t=a}^{b} \big(N_t - (\alpha + \gamma t)\big)^2$$

fits a local linear trend $N_t \approx \alpha + \gamma t$ to each subsequence $[a{:}b]$, and the ℓ_0 penalty $\beta|\boldsymbol{\tau}|$ to discourage over-segmentation (Truong et al., 2020).

The optimal set $\hat{\boldsymbol{\tau}}$ partitions the timeline into $M = |\hat{\boldsymbol{\tau}}|+1$ trend-homogeneous segments $\mathcal{C} = \{C_1, \ldots, C_M\}$, which we treat as continuous time-series clusters. From each cluster C_m ($m = 1, \ldots, M$) we then uniformly draw $n_{C_m} \approx \frac{n}{M}$ texts at random, yielding an annotation pool that is temporally representative of all detected discourse regimes.

In this approach, time intervals are dynamically defined by temporal trends in policy-related discourse, capturing sentiment shifts triggered by external shocks and evolving opinions that unfold over variable-length periods.

3.2 Building a model

3.2.1 Finetuning a single model

Upon annotating the sampled data, the most straightforward and commonly employed approach is to finetune a single unified model using all the annotated data-points.

3.2.2 Merging multiple models across time intervals

To account for temporal dynamics across data points, some propose fine-tuning separate models – each trained exclusively on data from a specific time interval – and subsequently merging them into a unified models (Aghapour and Rahili, 2024; Wortsman et al., 2022; Nylund et al., 2024). This approach aims to embed time into the model's weights by integrating multiple specialized models, each of which is fine-tuned to a specific time interval. We hence experimented the following merging techniques:

Souping Souping, which involves averaging the weights of multiple models, remains a commonly employed merging technique across distinct time intervals (Wortsman et al., 2022; Nylund et al., 2024). Two variants are commonly used: uniform souping, which equally averages the weights of all models from each time interval, and greedy souping, an iterative approach that sequentially adds models into the averaged ensemble, retaining each new model only if it improves performance on a held-out validation set.

Task Arithmetic Task Arithmetic uses "task vectors" that capture the parameter-space direction of a task (Ilharco et al., 2022). Task vectors τ can be defined as the element-wise difference between a model fine-tuned on time interval T and the pre-trained weights θ_{pre}. Hence, we learn a task vector for each interval T and add them to the base parameters $\left(\theta_{\text{pre}} + \lambda \sum_{T \in \mathcal{T}} \tau_T\right)$ to obtain a merged model.

TIES Merging TrIm, Elect Sign, and Merge (TIES Merging) trims each task vector to the top $k\%$ largest-magnitude values, then elects the sign with the greatest total magnitude across the trimmed vectors before merging (Yadav et al., 2023). In doing so, it aims to remove redundant parameters and resolve sign conflicts during merging.

DARE Drop And REscale (DARE) proposes randomly dropping $p\%$ of *delta* parameters and rescaling the remaining ones $\left(\text{by } \frac{1}{1-p}\right)$ before merging the models (Yu et al., 2024), aiming to eliminate small and redundant changes witnessed in fine-tuned models from their pre-trained variants.

Fisher Merging Across multiple fine-tuned models derived from the same pretrained model, Fisher Merging first estimates the diagonal Fisher information for each model using a small batch of task-specific data (Matena and Raffel, 2022). Subsequently, for each parameter, it computes a weighted average across the models, with weights determined by the Fisher scores. Parameters considered more informative thus have greater influence, enabling the merged model to retain essential updates and minimize interference.

RegMean Merging Regression Mean (RegMean) merging treats model merging as a regression problem by computing an optimal weighted average of parameters across fine-tuned models (Matena and Raffel, 2022). Specifically, it uses the inner product matrices of layer inputs from each model to find parameters minimizing the squared difference between merged and individual model outputs. This hence reweighs and linearly combines parameter rows based on their importance.

4 Experimental Setup

4.1 Datasets

We perform our above-mentioned methods on 3 datasets that meet the following criteria: (1) a sentiment classification task, (2) data is policy-relevant, (3) all texts are professionally annotated, (4) dataset details, particularly the time-stamps, are available, and (5) is sufficiently large. Details of each dataset are elaborated in Appendix A.

Climate Change Twitter Dataset The Climate Change Twitter Dataset (Effrosynidis et al., 2022; Bauch and Qian, 2018) contains 43,943 annotated tweets surrounding climate change sentiments spanning Apr 27, 2015 and Feb 21, 2018. Tweets are labeled as `Pro-`, `Anti-`, `Neutral-` and `News-` stance towards climate change.

AI Perceptions The "Long-Term Trends of Public Perception of Artificial Intelligence (AI)", which we will call the AI Perceptions dataset, is a dataset that captures nearly 30 years of public perceptions regarding AI. Annotators labeled perceptions based on 5,685 paragraphs extracted from

New York Times (NYT) articles related to AI, spanning 1986 to 2016 (Fast and Horvitz, 2017; Shahane et al., 2018). Perceptions are categorized as either Positive, Negative, or Neutral/Mixed.

COVID Vaccine Twitter Dataset The COVID Vaccine Twitter Dataset contains 6,000 tweets annotated with sentiment labels (positive, negative, or neutral) toward COVID-19 vaccines. The tweets were collected during the initial months following the vaccine's release, spanning December 2020 through April 2021 (Preda, 2021b,a).

4.2 Model fine-tuning and evaluation

To mimic the typical sentiment analysis process employed in policy-related studies – where large datasets are classified using models fine-tuned on partially annotated subsets (An et al., 2023; Efrosynidis et al., 2022; Maceda et al., 2023; Melton et al., 2022) – we sample 10,000, 2,000, and 3,000 annotated data points from the Climate Change Twitter, AI Perceptions, and COVID-19 Vaccine Twitter datasets, respectively, using the strategies detailed in Section 3.1. These sampled data points are used to fine-tune pretrained models. The remaining data points are reserved for evaluation, mimicking the practical scenario in which models trained on a subset of annotated data are subsequently used to classify sentiments of remaining unlabeled corpora.

We performed our experiments on four pretrained models commonly employed in text classification: DeBERTa$_{\text{large}}$ (He et al., 2021), RoBERTa$_{\text{large}}$ (Liu et al., 2019), BERT$_{\text{large}}$ (Devlin et al., 2019), and a domain-specific model selected based on the dataset – BERTweet$_{\text{large}}$ (Nguyen et al., 2020) for Twitter data and NewsBERT (Wu et al., 2022) for news data. The training hyperparameters are detailed in Appendix B.

5 Results

5.1 Selecting data points for labeling

We begin by evaluating the sampling approaches described in Section 3.1 in selecting annotated data points to fine-tune a unified sentiment classification model. When sampling through fixed time intervals, we set the temporal granularity to monthly for the Climate Change Twitter and COVID-19 Vaccine Twitter datasets, and annually for the AI Perceptions dataset. Similarly, when sampling through continuous time series clustering, we cluster base on the daily, monthly and annual trends for the COVID-19 Vaccine Twitter, Climate Change Twitter, and AI Perceptions datasets, respectively. The clusters identified through continuous time-series clustering for each dataset are shown in Figure 3.

Our overall results demonstrate competitive or superior performances relative to prior studies (Efrosynidis et al., 2022; Almars et al., 2022; Thenmozhi et al., 2024; Akpatsa et al., 2022), even though those studies employed traditional train-test splits, whereas we used smaller annotated subsets to mimic realistic annotation constraints in policy-related research.

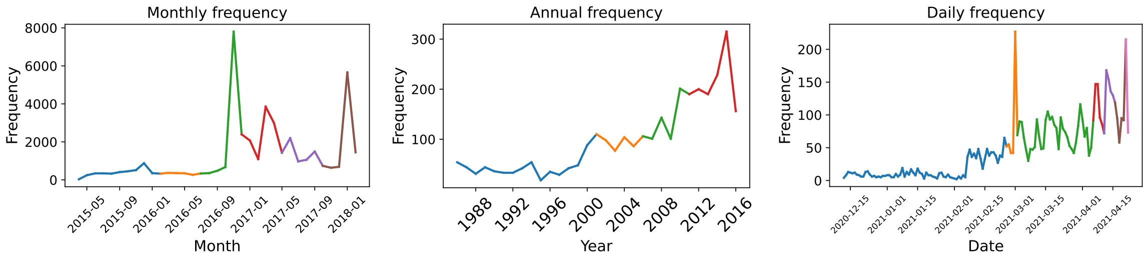
As shown in Table 1, our proposed method of using continuous time-series clustering to select data points for annotation and model fine-tuning consistently outperforms random selection – improving upon average F1-score and accuracy by 2.71% and 1.18%, respectively. Similarly, our method of selecting through continuous time-series sampling improves upon fixed time-interval sampling by an average F1-score and accuracy score of 4.03% and 1.92%, respectively. Surprisingly, fixed-interval sampling results in a slight performance deterioration relative to random selection, with an average decrease in F1-score of 0.99%.

5.2 Building a robust model across time intervals

Having determine the best strategy when selecting the data for annotation towards model fine-tuning, we proceed to assess the effectiveness of the merging methods outlined in Section 3.2.2, wherein models fine-tuned separately on data from distinct time intervals are merged. We then compare the performance of these merged models against the single unified model fine-tuned across all intervals in Section 5.1.

As shown in Figure 4, our results show that fine-tuning a single unified model using data from all time intervals consistently outperforms merging individually fine-tuned models from separate intervals. The sole exception arises from the DeBERTa$_{\text{large}}$ variant from the AI perceptions dataset, in which greedy souping outperforms a single unified model by 0.89%.

Nonetheless, in many cases, certain merging techniques – particularly greedy souping and TIES merge – yields very competitive performances, often coming a few percentage points off a single unified model. This suggests that merging separately fine-tuned models may still be advantageous in scenarios involving incremental or online learning,

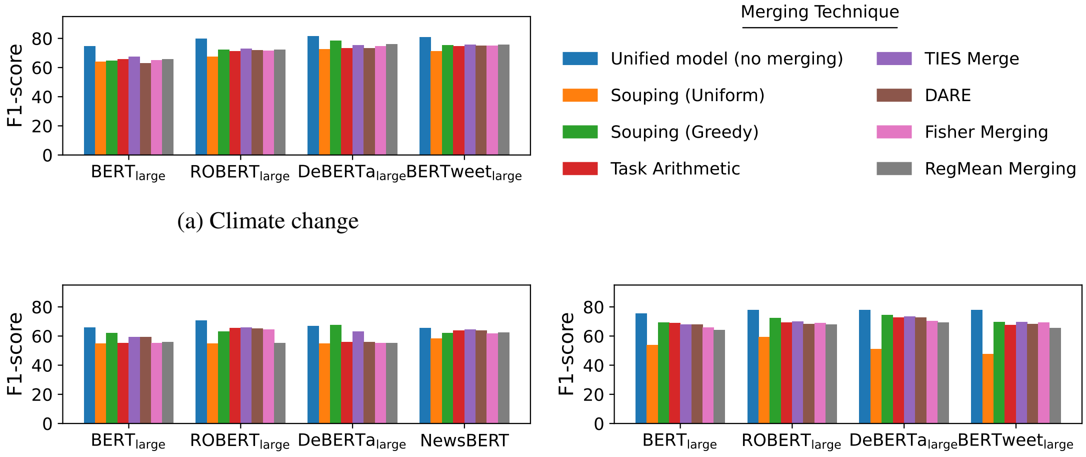


(a) Climate change (b) AI Perceptions (c) COVID-19 Vaccine

Figure 3: Clusters obtained from continuous time-series clustering based on temporal trends within each dataset. Distinct colors correspond to individual clusters.

Type	Model	Climate Change			AI Perceptions			COVID vaccine		
		Accuracy	F1	AUROC	Accuracy	F1	AUROC	Accuracy	F1	AUROC
Random Sample	RoBERTa$_{\text{large}}$	79.93%	79.26%	93.48%	68.58%	58.09%	76.42%	77.37%	76.88%	87.46%
Fixed intervals		79.65%	79.26%	93.00%	69.03%	58.58%	75.17%	77.37%	77.02%	87.00%
Continous time series clusters		**80.34%**	**79.81%**	**93.63%**	**72.75%**	**70.49%**	**77.38%**	**77.58%**	**77.68%**	**87.64%**
Random Sample	BERT$_{\text{large}}$	74.79%	74.28%	90.12%	68.77%	58.75%	72.00%	74.23%	71.90%	85.03%
Fixed intervals		74.54%	74.06%	89.66%	67.75%	54.72%	69.07%	73.91%	71.12%	83.96%
Continous time series clusters		**75.40%**	**74.78%**	**90.14%**	**71.35%**	**65.75%**	**73.27%**	**76.05%**	**75.49%**	**85.69%**
Random Sample	DeBERTa$_{\text{large}}$	81.67%	81.37%	93.90%	69.06%	62.51%	73.69%	77.60%	76.81%	**86.83%**
Fixed intervals		80.75%	80.65%	93.66%	71.34%	66.24%	73.95%	77.98%	77.62%	86.26%
Continous time series clusters		**81.79%**	**81.49%**	**94.05%**	**71.90%**	**66.69%**	**74.90%**	**78.27%**	**77.92%**	86.58%
Random Sample	BERTweet$_{\text{large}}$ / NewsBERT	80.99%	80.41%	93.93%	70.64%	64.23%	**75.24%**	77.77%	77.56%	87.96%
Fixed intervals		80.01%	79.55%	93.48%	69.63%	60.49%	73.37%	70.53%	66.87%	74.54%
Continous time series clusters		**81.38%**	**80.87%**	**94.09%**	**70.89%**	**65.63%**	75.10%	**77.87%**	**77.94%**	**88.18%**

Table 1: Results spanning the distinct sampling approaches in selecting data points for annotation and model fine-tuning. Among each dataset, the best performing results across each model are **bolded** and the best results across all models are underlined.



(a) Climate change

(b) AI Perceptions (c) COVID-19 Vaccine

Figure 4: Results spanning the distinct merging techniques.

where new data continually streams in as policies and associated events evolve over time.

We further examined whether merging models fine-tuned on fixed intervals, as opposed to continuous time series clusters, might improve performance. Additional experiments, detailed in Appendix C, shows that merging models base on fixed intervals performed even worse than merging cluster-based models, reinforcing the advantage of continuous clustering for both unified and merged-model strategies.

6 Discussion

Despite advancements in LLMs enhancing sentiment classification among complex, nuanced policy texts, existing methods often neglect the temporally volatile nature of its associated sentiments, which continuously evolves due to external shocks and evolving discourse of opinions. To this end, we propose methods to account for the temporally-sensitive nature of policy-related texts (Alkhalifa et al., 2021; Giuliano and Spilimbergo, 2024) and experimentally evaluate them in realistic settings that mimic sentiment analysis as conducted in policy-related studies. Specifically, we propose leveraging continuous time-series clustering to select data points for annotation based on temporal trends before subsequently applying advance merging techniques to merge multiple models, each fine-tuned separately on data from distinct time intervals.

Our results demonstrate that sampling data points for annotation through continuous time-series clustering, and subsequently fine-tuning a single unified model using all annotated data, yields the best performance. These findings are unsurprising given that they echo the results of Nylund et al. (2024), who found that fine-tuning a single model across all time intervals outperformed merging individually fine-tuned models trained separately on each time interval in all but one instance, despite the merged models collectively receiving five times more training data – albeit in a different downstream task from ours.

Our results suggests that language models can generalize across temporally volatile sentiments associated with policy-related texts across multiple time points, provided they are fine-tuned on representative samples that capture meaningful semantic variations within evolving policy discourse (Azarbonyad et al., 2017). Hence, leveraging machine learning methods to identify distinct temporal patterns allows us to select more representative samples for annotation and model fine-tuning, effectively capturing varying trends associated with sentiment shifts driven by external shocks or evolving opinions across variable-length periods (Alkhalifa et al., 2021). These patterns align with previous studies, which have demonstrated that accounting for temporality when applying language models to downstream tasks – especially in domains subject to temporal volatility – can improve performances (Röttger and Pierrehumbert, 2021; Lazaridou et al., 2021; Dhingra et al., 2022).

Nonetheless, the attainment of competitive performances when merging multiple models – each trained on intervals determined through continuous time-series clustering – using techniques such as greedy souping and TIES merging could be beneficial in certain practical scenarios. For instance, when significant events or shifts – such as political transitions – lead to external shocks that substantially alter public sentiment (e.g., sudden changes in online immigration-policy rhetoric following President Trump's emergence and subsequent election (Quinonez, 2018)) that may necessitate the collection and annotation additional data to update already-tuned language models in order to facilitate an up-to-date policy analysis of sentiments (Azarbonyad et al., 2017; Alkhalifa et al., 2021). Under such conditions, merging newly fine-tuned models with previously trained models offers an efficient and flexible alternative to retraining a single classifier from scratch.

7 Conclusions

Sentiments in policy-related texts exhibit high volatility due to external shocks and evolving discourse. We posit that these temporal dynamics are typically overlooked by existing methods. To address this, we propose leveraging continuous time-series clustering to select temporally representative data points for annotation, followed by advance merging techniques to combine models fine-tuned on distinct time intervals.

Our results show that continuous time-series clustering combined with fine-tuning a single unified model outperforms conventional random sampling by an average F1-score of 2.71%. Although merging multiple models typically reduces performance compared to a unified model, certain merging methods – particularly greedy souping and TIES merging – yield competitive results. These findings suggest language models effectively generalize to temporally sensitive policy texts when trained on representative samples. Furthermore, the competitive performance of merged time-specific models indicates practical advantages in dynamically evolving policy contexts.

Limitations

Our analyses – from the experimental setup and selected datasets to the choice of models – were explicitly designed to mimic sentiment analysis tasks in policy-related contexts. While our results are consistent with similar studies (Nylund et al., 2024; Lazaridou et al., 2021), as discussed in Section 6, further research is needed to explore whether these findings generalize effectively to other downstream tasks across distinct domains.

Additionally, our experiments employed transfer learning on partially annotated datasets to mimic practical constraints – such as limited annotation resources – which represent the most common and straightforward method for leveraging robust language models for policy-related sentiment analysis (An et al., 2023; Effrosynidis et al., 2022; Maceda et al., 2023; Melton et al., 2022). Nonetheless, further research could explore incorporating unannotated examples and their temporal contexts, potentially enhancing the generalizability of predictions across multiple time intervals through weak supervision (Tong et al., 2024) and semi-supervised learning techniques (Shi et al., 2023).

Furthermore, fine-tuning on limited subsets may directly influence the predictive performance of our models. While our chosen subset sizes were guided by prior studies in policy-related contexts (An et al., 2023; Effrosynidis et al., 2022; Maceda et al., 2023; Melton et al., 2022), the precise relationship between relative training sample size and predictive performance remains unclear, as does the optimal subset size within commonly employed setups for policy-related sentiment analysis. We therefore highlight these as important considerations for future work.

Moreover, as open-source LLMs with impressive reasoning capabilities (Grattafiori et al., 2024; Guo et al., 2025) continue to emerge, their performance in classifying sentiments within temporally volatile policy contexts under few-shot settings remains unclear. If such models excel under these conditions, the practical advantages of our approach may be diminished. Thus, comparing the effectiveness of few-shot learning with larger, reasoning-focused LLMs against our proposed methods represents an important avenue for future research.

Finally, our work was evaluated on benchmark datasets covering global policy topics—climate change, artificial intelligence perceptions, and COVID-19 vaccine attitudes—primarily due to the extensive availability of fully annotated datasets in these domains. However, sentiment analysis is also commonly applied to national and local policies (Maceda et al., 2023; Haqbeen et al., 2021; Chen and Wei, 2023; An et al., 2023), where typically only a subset of data is annotated, similar to our experimental setup. Since national and local policies often exhibit greater temporal volatility (Henisz, 2004), it remains unclear if our findings would generalize to these contexts.

Ethical Considerations

Given that sentiments expressed in policy-related opinions in online spaces are often intertwined with racial, gender, age, and socio-economic stereotypes, there is an inherent risk that fine-tuned language models may similarly associate stereotype-embedded terminologies with particular sentiments (Lee et al., 2024). Furthermore, policy-related sentiments can be highly subjective; thus, annotators may inadvertently introduce their own biases or stereotypical associations into the manual annotation process, potentially embedding these biases into models during fine-tuning (Sap et al., 2022; Davani et al., 2023).

References

Elahe Aghapour and Salar Rahili. 2024. Beyond fine-tuning: Merging specialized llms without the data burden. Medium.

Samuel Kofi Akpatsa, Xiaoyu Li, Hang Lei, and Victor-Hillary Kofi Setornyo Obeng. 2022. Evaluating public sentiment of covid-19 vaccine tweets using machine learning techniques. *Informatica*, 46(1).

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2021. Opinions are made to be changed: Temporally adaptive stance classification. In *Proceedings of the 2021 workshop on open challenges in online social networks*, pages 27–32.

Abdulqader M Almars, El-Sayed Atlam, Talal H Noor, Ghada ELmarhomy, Rasha Alagamy, and Ibrahim Gad. 2022. Users opinion and emotion understanding in social media regarding covid-19 vaccine. *Computing*, 104(6):1481–1496.

Ruopeng An, Yuyi Yang, Quinlan Batcheller, and Qianzi Zhou. 2023. Sentiment analysis of tweets on soda taxes. *Journal of Public Health Management and Practice*, 29(5):633–639.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518.

Chris Bauch and Edward Qian. 2018. Twitter climate change sentiment dataset.

Andrea Ceron and Fedra Negri. 2015. Public policy and social media: How sentiment analysis can support policy-makers across the policy cycle. *Rivista Italiana di Politiche Pubbliche*, 10(3):309–338.

Kehao Chen and Guiyu Wei. 2023. Public sentiment analysis on urban regeneration: A massive data study based on sentiment knowledge enhanced pretraining and latent dirichlet allocation. *Plos one*, 18(4):e0285175.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

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.

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 Association for Computational Linguistics*, 10:257–273.

Sebastian Dziadzio, Vishaal Udandarao, Karsten Roth, Ameya Prabhu, Zeynep Akata, Samuel Albanie, and Matthias Bethge. 2025. How to merge multimodal models over time? In *ICLR 2025 Workshop on Modularity for Collaborative, Decentralized, and Continual Deep Learning*.

Dimitrios Effrosynidis, Alexandros I Karasakalidis, Georgios Sylaios, and Avi Arampatzis. 2022. The climate change twitter dataset. *Expert Systems with Applications*, 204:117541.

Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Juan M Floyd-Thomas. 2016. Welfare reform and the ghost of the" welfare queen". *New Politics*, 16(1):29.

Paola Giuliano and Antonio Spilimbergo. 2024. Aggregate shocks and the formation of preferences and beliefs. *IMF Working Papers*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jawad Haqbeen, Sofia Sahab, Takayuki Ito, and Paola Rizzi. 2021. Using decision support system to enable crowd identify neighborhood issues and its solutions for policy makers: An online experiment at kabul municipal level. *Sustainability*, 13(10):5453.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Witold Jerzy Henisz. 2004. Political institutions and policy volatility. *Economics & politics*, 16(1):1–27.

Tamanna Hossain, Robert L Logan Iv, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, and 1 others. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Messi HJ Lee, Jacob M Montgomery, and Calvin K Lai. 2024. America's racial framework of superiority and americanness embedded in natural language. *PNAS nexus*, 3(1):pgad485.

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

Lany L Maceda, Arlene A Satuito, and Mideth B Abisado. 2023. Sentiment analysis of code-mixed social media data on philippine uaqte using fine-tuned mbert model. *International Journal of Advanced Computer Science and Applications*, 14(7).

Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.

Chad A Melton, Brianna M White, Robert L Davis, Robert A Bednarczyk, and Arash Shaban-Nejad. 2022. Fine-tuned sentiment analysis of covid-19 vaccine–related social media data: Comparative study. *Journal of Medical Internet Research*, 24(10):e40408.

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 Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Kai Nylund, Suchin Gururangan, and Noah Smith. 2024. Time is encoded in the weights of finetuned language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2571–2587, Bangkok, Thailand. Association for Computational Linguistics.

Gabriel Preda. 2021a. Covid-19 all vaccines tweets.

Gabriel Preda. 2021b. Covid-19 vaccine tweets with sentiment annotation.

Erika Sabrina Quinonez. 2018. welcome to america: a critical discourse analysis of anti-immigrant rhetoric in trump's speeches and conservative mainstream media. Master's thesis, California State University - San Bernardino.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of bert and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Saurabh Shahane, Ethan Fast, and Eric Horvitz. 2018. Public perception of ai.

Zhengxiang Shi, Francesco Tonolini, Nikolaos Aletras, Emine Yilmaz, Gabriella Kazai, and Yunlong Jiao. 2023. Rethinking semi-supervised learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5614–5634.

M Thenmozhi, G Shubigsha, G Sindhuja, and V Dhinakar. 2024. Sentiment analysis on climate change using twitter data. In *2024 2nd International Conference on Networking and Communications (ICNWC)*, pages 1–6. IEEE.

Yongqi Tong, Sizhe Wang, Dawei Li, Yifan Wang, Simeng Han, Zi Lin, Chengsong Huang, Jiaxin Huang, and Jingbo Shang. 2024. Optimizing language model's reasoning abilities with weak supervision. *arXiv preprint arXiv:2405.04086*.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and 1 others. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Di Wu, Wasi Uddin Ahmad, and Kai-Wei Chang. 2022. Pre-trained language models for keyphrase generation: A thorough empirical study. *arXiv preprint arXiv:2212.10233*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: resolving interference when merging models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 7093–7115.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, pages 57755–57775. PMLR.

A Dataset details

Climate Change Twitter Dataset Tweets were annotated, by Bauch and Qian, as `Pro` if it supports the concept of man-made climate change, `Anti` if the tweet denies man-made climate change, `News` if it contains factual news information regarding climate change, and `neutral` if it neither beliefs nor denies the role of man-made climate change. In total, there were 22962 (52.25%) `Pro`, 9276 (21.11%) `news`, 420 (17.56%) `neutral`, and 3990 (9.08%) `Anti` sentiments. Missing timestamps were imputed based on the nearest-neighbor tweet ID, as tweet IDs are generated incrementally and correspond directly to the chronological posting order.

AI Perceptions The dataset was annotated, by Fast and Horvitz, as either "positive" or "negative" based on several key indicators. Positive indicators include its beneficial impact on (1) education, (2) transportation, (3) entertainment, (4) healthcare, (5) decision-making, (6) work, (7) positive singularity, (8) merging of Ai and human applications, otherwise known as cyborg (e.g., robotic limbs for the disabled) and (9) others. Negative indicators included (1) loss of control, (2) negative impact on work, (2) military applications, (3) ethics, (4) military applications, (5) lack of progress, (6) negative singularity, (7) negative cyborg applications (e.g., cyborg soldiers), and (8) others. Among each annotator, we consider their sentiment to be negative if majority of the selected indicators were negative, and vice-versa. We consider the sentiments to be "neutral or mixed" if none of the indicators were selected or an equal amount of negative and positive indicators were selected. In total, there were 4065 (71.47%) `neutral / mixed`, 1220 (21.45%) `positive`, and 402 (7.07%) `negative` sentiments. The final sentiment label was determined based on a majority vote among the annotators. In lieu of some text having missing timestamps, we sampled the annotated data-points (and plotted Figure 3) from texts with corresponding time-stamps.

COVID-19 Twitter Dataset Tweets were annotated, by Preda, based on their sentiments towards the COVID-19 vaccine during the initial months following the vaccine's roll-out and approval, on December 11 2020, spanning December 2020 through April 2021 (Preda, 2021b,a). The vaccines that were covered in the dataset included Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford / Astra Zeneca, Covaxin, and the Sputnik V vaccines. In total, there were 3680 (61.33%) `neutral`, 1900 (31.66%) `positive`, and 420 (7%) `negative` sentiments. Missing timestamps were imputed based on the nearest-neighbor tweet ID, as tweet IDs are generated incrementally and correspond directly to the chronological posting order.

B Hyper-parameters

B.1 Finetuning Parameters

We fine-tune all models using learning rates of $\{1 \times 10^{-5},\ 2 \times 10^{-5}\}$, batch sizes of 6 for RoBERTa$_{\text{large}}$; 8 for RoBERTa$_{\text{large}}$, BERT$_{\text{large}}$, and BERTweet$_{\text{large}}$; and 12 for NewsBERT. Additionally, we use a warmup ratio of 5% and weight decay of $\{0.01,\ 0.1\}$. Models fine-tuned across all time intervals are trained for up to 3 epochs with an early stopping patience of 2, while models fine-tuned within each time interval are trained for up to 8 epochs, also with an early stopping patience of 2 – though early stopping criteria are mostly met before reaching the maximum number of epochs. These hyper-parameters are adapted from previous studies employing the same datasets (Effrosynidis et al., 2022; Almars et al., 2022; Thenmozhi et al., 2024; Akpatsa et al., 2022). All models were fine-tuned on a Nvidia GeForce RTX 4090.

B.2 Parameters for Continuous Time-Series Clustering

When sampling data using continuous time-series clustering, we set the temporal granularity t to daily, monthly, and yearly trends for the COVID-19 Vaccine Twitter, Climate Change Twitter, and AI Perceptions datasets, respectively. The penalty parameter $\beta|\boldsymbol{\tau}|$ for clustering was set to 0.5 for the COVID-19 Vaccine Twitter dataset and 0.1 for both the Climate Change Twitter and AI Perceptions datasets.

B.3 Model merging parameters

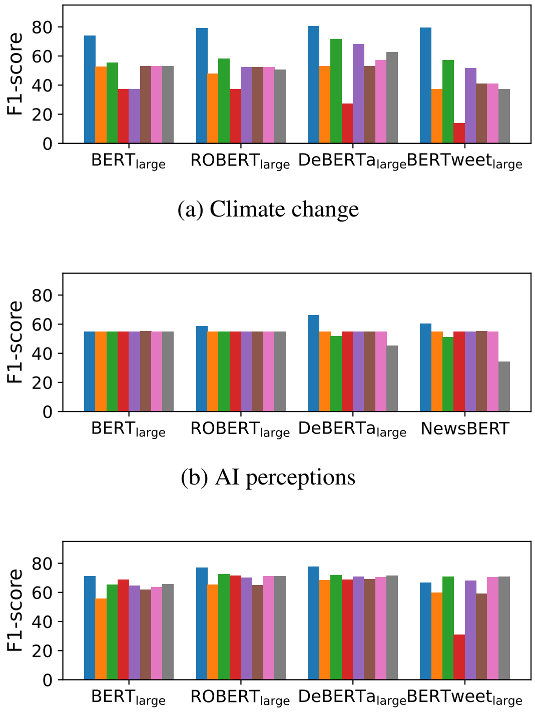
Table 2 summarizes the range of hyperparameters explored across the different model merging techniques. For each merging technique, hyperparameter configurations were evaluated on a held-out validation set, and the optimal parameters were selected. We adopted these range of hyper-parameters from Yu et al., Yadav et al., and Ilharco et al..

Merging method	Range of hyper-parameters
Task Arithmetic	λ: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
TIES Merging	λ: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] $k\%$: [10, 20, 30]
DARE Merging	λ: [0.1, 0.3, 0.5, 0.7, 0.9, 1.0] p: [0.5, 0.6, 0.7, 0.8, 0.9]

Table 2: Searched ranges of hyper-parameters of model merging methods

C Additional Results

Results when merging merging models fine-tuned on fixed intervals, as opposed to continuous time series clusters are shown in Figure 5. Note that unlike the aforementioned section, the λ parameters were fixed here but the remaining parameters were selected via a held-out validation set (similar to Section B.3). Overall, results of models merged on fixed intervals performed even worse than models merged on time series clusters. The observations are similar to the results in Section 5.2: fine-tuning a single unified model using data from all time intervals consistently outperforms merging individually fine-tuned models from separate intervals.



(a) Climate change

(b) AI perceptions

(c) COVID Vaccine

Figure 5: Results when merging models fine-tuned on fixed intervals, as opposed to continuous time series clusters.