Evolving Domain Adaptation of Pre-trained Language Models for Text Classification

Yun-Shiuan Chuang
University of Wisconsin-Madison
yunshiuan.chuang@wisc.edu

Yi Wu
University of Wisconsin-Madison
ywu676@wisc.edu

Rheeya Uppaal
University of Wisconsin-Madison
uppaal@wisc.edu

Luhang Sun
University of Wisconsin-Madison
lsun232@wisc.edu

Makesh Narsimhan Sreedhar
University of Wisconsin-Madison
msreedhar@wisc.edu

Sijia Yang
University of Wisconsin-Madison
syang84@wisc.edu

Timothy T. Rogers
University of Wisconsin-Madison
ttrogers@wisc.edu

Junjie Hu
University of Wisconsin-Madison
junjie.hu@wisc.edu

Abstract

Pre-trained language models have shown impressive performance in various text classification tasks. However, the performance of these models is highly dependent on the quality and domain of the labeled examples. In dynamic real-world environments, text data content naturally evolves over time, leading to a natural *evolving domain shift*. Over time, this continuous temporal shift impairs the performance of static models, as their training becomes increasingly outdated. To address this issue, we propose two dynamic buffer-based adaptation strategies: one utilizes self-training with pseudo-labeling, and the other employs a tuning-free, in-context learning approach for large language models (LLMs). We validate our methods with extensive experiments on two longitudinal real-world social media datasets, demonstrating their superiority compared to unadapted baselines. Furthermore, we introduce a COVID-19 vaccination stance detection dataset, serving as a benchmark for evaluating pre-trained language models within evolving domain adaptation settings.

1 Introduction

Text classification using pre-trained language models (PLMs)(Devlin et al., 2018; Brown et al., 2020) is essential for tasks like sentiment analysis on platforms such as Twitter and Amazon. Given the ever-evolving content on these platforms, there's a constant need for timely annotation of large-scale time-series data. Consequently, adapting PLMs to current unlabeled data becomes invaluable (ALDayel and Magdy, 2021; Küçük and Can, 2020).

Existing approaches to text classification with PLMs fall into two main categories. Some methods fine-tune PLMs using a limited set of labeled data (Devlin et al., 2018), while others utilize few-shot in-context examples to guide large language models (LLMs) in their predictions (Min et al., 2022; Dong et al., 2023; Kim et al., 2022).

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: F_{avg} across tuning-based and prompting methods, along with baselines. Bold face highlights the best method within each setting.

Category	Model	Setting	COVID	WTWT
Tuning-Based	BERT	Src-Only	0.509	0.618
		AST	0.460	0.633
		BST	0.569	0.697
		CST	**0.611**	**0.739**
		Supervised	0.687	0.785
Prompting	ChatGPT	Zero	0.746	0.590
		Src-Only	0.780	0.632
		OIE	**0.791**	**0.645**
	FLAN-T5	Zero	0.779	0.560
		Src-Only	0.784	0.562
		OIE	**0.788**	**0.563**

However, these PLMs face a significant challenge: the data distribution often shifts between the training phase and the deployment phase. This is particularly evident in time-series text classifications, where an *evolving domain shift* (EDS) can occur (Alkhalifa et al., 2021; Alkhalifa and Zubiaga, 2022; Mu et al., 2023). For example, people's narratives about COVID-19 vaccines change over time with the emergence of new virus or variants brands. To address this, our study delves into *evolving domain adaptation* (EDA) techniques tailored for PLMs for text classification.

To this end, we propose two dynamic buffer-based EDA approaches: (1) *Online Buffered Self-Training* (OBS) for fine-tuning PLMs using self-training on fresh unlabeled data, inspired by Kumar et al. (2020); and (2) *online update of in-context examples* (OIE) for prompting large language models (LLMs). OBS employs a dynamic buffer for iterative fine-tuning, while OIE leverages it to choose pertinent in-context examples. Evaluations on health and finance stance detection datasets show both methods excel over unadapted models in handling evolving domain shifts.

2 Preliminaries

In **Evolving Domain Adaptation (EDA)**, we address a M-way text classification task over time. We have an initial labeled dataset $\mathcal{D}_0$ from distribution $P^0_{\mathcal{X}\mathcal{Y}}$ and a series of subsequent unlabeled datasets $(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T)$ each from a unique input distribution $P^t_{\mathcal{X}}$. There's an evolving shift in the input distribution across consecutive time steps, signified by $P^{t-1}_{\mathcal{X}} \neq P^t_{\mathcal{X}}$. Label distribution, $P^t_{\mathcal{Y}}$, also evolves over time, implying $P^{t-1}_{\mathcal{Y}} \neq P^t_{\mathcal{Y}}$. The dataset concatenation between two time steps is denoted as $\mathcal{D}_{t:t'}$. EDA's objective is leveraging all datasets $\mathcal{D}_{0:T}$ to optimize a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$ across all unlabeled domains from $t \in [1, T]$.

3 Methods

We outline three self-training methods for fine-tuning small language models using pseudo-labeled data. Subsequently, we present a tuning-free, prompt-based approach employing a dynamic buffer for prompting large language models.

Fine-tuning with Self-training In the context of EDA, we introduce different variants of training methods. Self-training (Scudder, 1965; Amini et al., 2022) is a common approach to semi-supervised learning, which leverages a base model to predict pseudo-labels for unlabeled data and further fine-tunes the model on pseudo-labeled data. **1. Buffered Self-Training (BST):** BST method uses a dynamic buffer, $\mathcal{B}_t$, of fixed size b. At the start ($t = 0$), it's initialized with the source labeled dataset, $\mathcal{B}_0 = \mathcal{D}_0$, with size $b = n_0$. For each subsequent time step $t \in [1, T]$, a model, $f_{\theta,t}$, is fine-tuned using the buffer $\mathcal{B}_{t-1}$ and then predicts pseudo-labels for unlabeled instances from $\mathcal{D}_t$. These pseudo-labeled examples, $\tilde{\mathcal{D}}_t$, are added to the buffer. The buffer size remains consistent using a First-In-First-Out (FIFO) strategy, meaning when full, adding a new instance results in the removal of the oldest one, represented as: $\mathcal{B}_t = \text{FIFO}(\mathcal{B}_{t-1} \cup \{(x_i, \tilde{y}_i) | x_i \in \mathcal{D}_t\}, b)$, where FIFO$(\cdot, b)$ is a function that maintains the most recent b instances in the buffer. [1] **2. Cumulative Self-Training (CST):** Unlike BST, which keeps a buffer of fixed size, CST's buffer, denoted as $\mathcal{B}_t$ at the t-th step, accumulates data and grows over time. The buffer update in line 5 of Algorithm 1 for CST becomes: $\mathcal{B}_t = \mathcal{B}_{t-1} \cup (x_i, \tilde{y}_i)|x_i \in \mathcal{D}_t$. This variant tests the importance of retaining historical data. **3. All Self-Training (AST):** In AST, an initial model, $f_{\theta,0}$, is first trained on source domain $\mathcal{D}_0$. This model pseudo-labels instances across target domains $\mathcal{D}_{1:T}$, producing labels like $\tilde{\mathcal{D}}_t$. The data merge into $\tilde{\mathcal{D}}'_{0:T} = \mathcal{D}_0 \cup \tilde{\mathcal{D}}_1, \ldots, \tilde{\mathcal{D}}_T$, on which the final model, $f_{\theta,T}$, is trained.

Algorithm 1: Buffered Self-Training (BST)

```
Input: Source domain $\mathcal{D}_0$, target domains $\mathcal{D}_{1:T}$
1  Initialize buffer $\mathcal{B}_0 = \mathcal{D}_0$, buffer size $b = |\mathcal{D}_0|$
2  for $t = 1$ to $T$ do
3      Train model $f_{\theta,t}$ on $\mathcal{B}_{t-1}$
4      Obtain pseudo-label $\tilde{y}_i = f_{\theta,t}(x_i), \forall x_i \in D_t$
5      Update $\mathcal{B}_t$ by FIFO
Output: $\{f_{\theta,t}\}_{t=1}^{T}$
```

Tuning-free Method by Prompting LLMs In addition to fine-tuning, we explore a tuning-free, dynamic prompting technique using in-context examples to guide large language models (LLMs) for classification. [2] **Online Update of In-Context Examples (OIE):** Mirroring the dynamic approach of BST (§3), we maintain a dynamic buffer, $\mathcal{B}_t$, which evolves with each time step $t \in [1, T]$. From this buffer, we derive in-context examples, $\mathcal{C}_t \subseteq \mathcal{B}_t$. At each step t, for each class $y \in \mathcal{Y} = \{1, 2, ..., M\}$, we choose the most *representative* example from each class within $\mathcal{B}_{t-1}$. The representativeness is determined by first calculating the mean embedding for class y examples: $\mathbf{u}^y_{t-1} = \frac{1}{\mathcal{B}^y_{t-1}} \sum_{(x,y)\in\mathcal{B}^y_{t-1}} g(x)$, with g being a PLM encoder and $\mathcal{B}^y_{t-1}$ the set of pseudo-labeled examples for class y. We then pick the example closest to the mean embedding $\mathbf{u}^y_{t-1}$ using cosine similarity (formally defined in Appendix A). The most representative examples for all classes form the in-context set: $\mathcal{C}_{t-1} = \{(x_{\text{rep}}, v(y))|y \in \mathcal{Y}\}$. Using $\mathcal{C}_{t-1}$, we construct prompts for instances in $\mathcal{D}_t$, obtain the model's predictions, and add these pseudo-labeled examples to $\mathcal{B}_t$. This buffer also updates using the FIFO strategy, consistent with BST.

Algorithm 2: Online Update of In-context Examples (OIE)

```
Input: Source domain $\mathcal{D}_0$, target domains $\mathcal{D}_{1:T}$, #examples per class as $k$, label space $\mathcal{Y}$
Initialize buffer $\mathcal{B}_0 = \mathcal{D}_0$, buffer size $b = |\mathcal{D}_0|$
for $t = 1$ to $T$ do
    Initialize the in-context example set $\mathcal{C}_{t-1} = \varnothing$
    for $y$ in $\mathcal{Y}$ do
        Compute mean embedding $\mathbf{u}^y_{t-1}$
        Find the most-representative example $(x_{\text{rep}}, y)$
        Collect the in-context example $\mathcal{C}_{t-1} = \mathcal{C}_{t-1} \cup \{(x_{\text{rep}}, y)\}$;
    Create a few-shot prompt using $x'_i = \mathcal{T}(x_i, \mathcal{C}_{t-1}), \forall x_i \in \mathcal{D}_t$
    Obtain pseudo-label $\tilde{y}_i = v^{-1}(f_{\text{LLM}}(x'_i)), \forall v^{-1}(x'_i) \in \mathcal{D}_t$
    Update $\mathcal{B}_t$ by FIFO
Output: $\{C_t\}_{t=0}^{T-1}$, the set of all in-context examples
```

[1] Notably, when fine-tuning $f_{\theta,t}$, the training set in $\mathcal{B}_{t-1}$ is first upsampled to mitigate label imbalance. Therefore, regardless of label shift, there are always equal number of instances per class to fine-tune the model.

[2] Classification via prompting involves a verbalizer, $v : \mathcal{Y} \to \mathcal{Z}$, translating each label into a text phrase. An instance x_i is transformed into a prompt x'_i using a template function $\mathcal{T}$, integrating x_i with n_c in-context examples ($\mathcal{C}$). $x'_i = \mathcal{T}(x_i, \mathcal{C}) = [I_1]\, \underbrace{[x_1]\, [z_1] \ldots}_{\text{in-context } \mathcal{C}}\, [x_i]\, [I_2]$, where I_1 and I_2 are text instructions. The LLM is instructed using this prompt to produce text, $\tilde{z}_i = f_{\text{LLM}}(x'_i)$. This output is then parsed back to the original label using an inverse-verbalizer, resulting in $\tilde{y}_i = v^{-1}(\tilde{z}_i)$.

4 Experimental Settings

4.1 Datasets and Evaluation Metric

Datasets **1. COVID-19 Vaccination Dataset (COVID)** is constructed by the authors (See Appendix C). It contains 5002 tweets about COVID-19 vaccination, sampled daily from December 1, 2020, to June 30, 2022. Tweets are categorized as '*Against*' (anti-vaccine views) or '*Not-against*' (pro-vaccine, neutral, or ambiguous views). The source domain combines the first six months, with the subsequent year divided monthly to form ten target domains (See Appendix D). **2. Will-They-Won't-They Dataset (WTWT)** features tweets about Mergers and Acquisitions (M&A). They are classified into '*Support*', '*Refute*', '*Comment*', and '*Unrelated*'. We use the subset of 44,717 tweets from June 2015 to December 2018 (Conforti et al., 2020). The first year is the source domain, with subsequent tweets bi-monthly grouped into 14 target domains (See Appendix D).

Performance Evaluation Metric Model performance is assessed on target domain test sets $\mathcal{D}_{1:T}$. For each $t \in [1, T]$, the macro-averaged F1 Score, $F_{\text{macro},t}$, is computed over the label space $\mathcal{Y}$. The global performance metric, F_{avg}, is the average of all $F_{\text{macro},t}$, defined as $F_{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T} F_{\text{macro},t}$.

4.2 Fine-tuning Approach

Self-Training Methods and Baselines **1,2,3. BST, CST, AST:** Detailed in §3. **4. Source-only Baseline (Src-Only):** Trained only on source domain $\mathcal{D}_0$ and tested on target domains, $\mathcal{D}_{1:T}$. It acts as an unadapted benchmark. **5. Fully-supervised Baseline (Supervised):** Unlike other EDA methods, Supervised uses true labels from target domains, training on all labeled data and setting a performance upper bound for EDA. For all of the aforementioned methods, we use the BERT-large model (Devlin et al., 2018) as the backbone model for all subsequent methods. Hyperparameters for fine-tuning are described in Appendix G.

4.3 Tuning-free Approach by Prompting LLMs

Selection of Large Language Models We select two high-performing LLMs for text classification: FLAN-T5 and ChatGPT.[3] **1. FLAN-T5:** particularly the FLAN-T5-XXL variant (Chung et al., 2022; Ziems et al., 2023). **2. ChatGPT:** version gpt-3.5-turbo-0301, an OpenAI decoder model fine-tuned via RLHF. (Ziems et al., 2023; Gilardi et al., 2023).[4].

In-context Example Selection Methods and Baselines: **1. OIE:** Detailed in §3 with the few-shot prompt template in Table 9. **2. Source-only Baseline (Src-Only):** Uses representative examples solely from the source domain $\mathcal{C} \subseteq \mathcal{D}_0$, showcasing limitations of using in-context examples only from the source domain. **3. Zero-shot Baseline (Zero):** Only contains task instruction I and query text x_i, excluding in-context examples (see Table 8). Note that for OIE and Src-Only, to retrieve the text embeddings, we employ the E5-Large-V2 model as our encoder $g : \mathcal{X} \to \mathbb{R}^d$ to transform texts into 1024-sized embeddings. The E5-Large-V2 embeddings excel in tasks like semantic similarity, retrieval and classification (Wang et al., 2022) .[5]

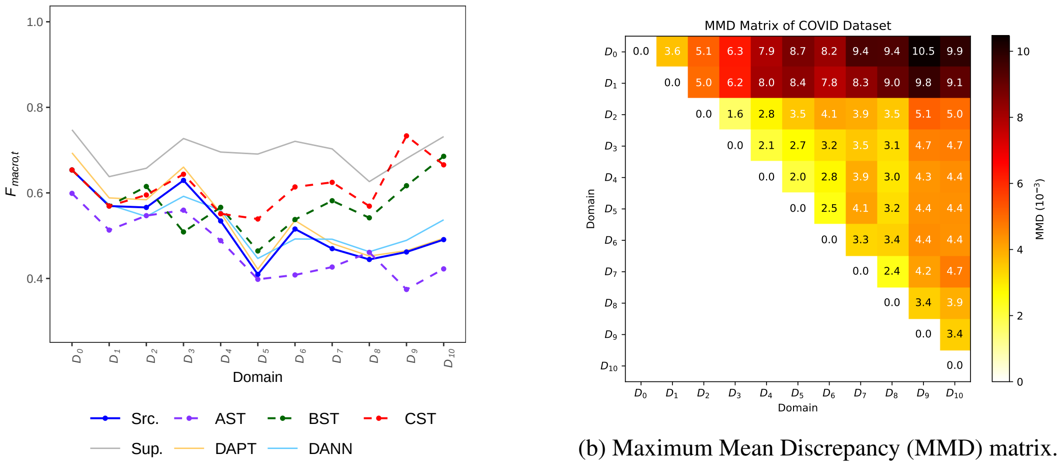
5 Results

Dynamic Buffer Helps Mitigate EDA **1. Fine-tuning with Self-training:** Table 1 reveals that BST and CST significantly outperform Src-Only. Notably, BST's gains come without increasing the training data size, highlighting the efficacy of a dynamic pseudolabeled data buffer. Between BST and CST, CST is superior, indicating the benefits of retaining all prior examples when storage is available. On the other hand, AST's performance is poor presumably because it ignores the temporal nature

[3] We use greedy decoding (with temperature = 0 for temperature sampling) for reproducibility in both FLAN-T5-XXL and ChatGPT, retaining default generation settings.

[4] In addition, as a closed-source model, ChatGPT's current API does not provide the logit for each generated text, presenting a practical limitation for certain strategies. This context underscores the need to devise a strategy that is compatible with closed-source models as well.

[5] As of June 21, 2023, E5-Large-V2 tops the MTEB Leaderboard (Muennighoff et al., 2022): https://huggingface.co/spaces/mteb/leaderboard



(a) $F_{\text{macro},t}$ of fine-tuning methods over domains. Five different fine-tuning methods are represented: Src-Only (blue), BST (green), CST (red), AST (purple), and Supervised (grey).

(b) Maximum Mean Discrepancy (MMD) matrix. Each cell represents the MMD between a pair of domains, calculated based on the marginal distribution of text embeddings $P_{g(\mathcal{X})}$ projected by the E5-Large-V2 model. The numbers in the heatmap are in the units of 10^{-3}.

Figure 1: Results of the COVID Dataset.

of the data. **2. Prompting with Evolving In-context Examples:** As shown in Table 1, it is clear that leveraging in-context examples is better than using a zero-shot prompt, in line with the literature. Secondly, OIE outperforms the use of static in-context examples from the source domain. When both models OIE, they demonstrate improved performance over Src-Only method, underscoring the benefits of continuously updating in-context examples in response to evolving data.

EDA is robust against evolving domain shift Model performance across domains for the COVID dataset is depicted in Figure 1a. While the Src-Only model performs well initially, its effectiveness diminishes over time. In contrast, BST, CST, and AST remain consistent. Domain divergence, calculated using MMD Gretton et al. (2012), provides insight into these trends (Figure 1b). Results on the WTWT dataset follows a similar trend, detailed in Appendix B and H.

6 Conclusion

We introduce evolving domain adaptation (EDA) methods that use dynamic buffering to mitigate the challenges posed by evolving domain shifts in text classification using PLMs. Our methods use fine-tuning of SLMs and prompting of LLMs. Our results highlight the importance of using up-to-date data for EDA, the significant role of intermediate domains, and the critical reliance of our strategies on accurate pseudo-labeling. Together, these insights offer an innovative perspective for addressing text classification in time-series data with pre-trained language models.

References

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

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.

Rabab Alkhalifa and Arkaitz Zubiaga. 2022. Capturing stance dynamics in social media: open challenges and research directions. *International Journal of Digital Humanities*, 3(1-3):115–135.

Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Emilie Devijver, and Yury Maximov. 2022. Self-training: A survey. *arXiv preprint arXiv:2202.12040*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv preprint arXiv:2005.00388*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Ananya Kumar, Tengyu Ma, and Percy Liang. 2020. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Yida Mu, Mali Jin, Kalina Bontcheva, and Xingyi Song. 2023. Examining temporalities on stance detection towards covid-19 vaccination. *arXiv preprint arXiv:2304.04806*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

H. Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

Supplementary Material

A Detailed Formula of Finding the Most Representative Example in OIE

For clarification, we formally define some seemingly obvious formulas about OIE (§5).

When constructing the in-context examples, we use the most representative example for class y as the example whose embedding is closest to the mean embedding $\mathbf{u}_{t-1}^{y}$. Formally, the representative example for class y at time $t-1$, denoted as (x_{rep}, y), is selected as follows:

$$(x_{\text{rep}}, y) = \operatorname*{argmax}_{(x_i,y_i)\in\mathcal{B}_{t-1}, y_i=y} \cos(g(x_i), \mathbf{u}_{t-1}^{y}), \tag{1}$$

where $\cos(g(x_i), \mathbf{u}_{t-1}^{y})$ is the cosine similarity between the embedding of example x_i and the mean embedding $\mathbf{u}_{t-1}^{y}$ for class y. This would select the example whose embedding $g(x)$ is the most similar to the mean embedding of class y.

B Result Figures on the WTWT dataset

Figure 2 visualizes the model performance over domains for the WTWT dataset when using fine-tuning with self-training. It highlights a noteworthy phenomenon in the performance of online learning models. When data from domain $\mathcal{D}_6$ is introduced, both BST and CST undergo a sharp performance drop, mirroring the decline in Src-Only. However, unlike the baseline model, BST and CST recover quickly in the subsequent domain domains. The key to this recovery lies in their ability to update the buffer with pseudo-labeled examples from $\mathcal{D}_6$. As a result, the model is able to adapt to the domain shift.

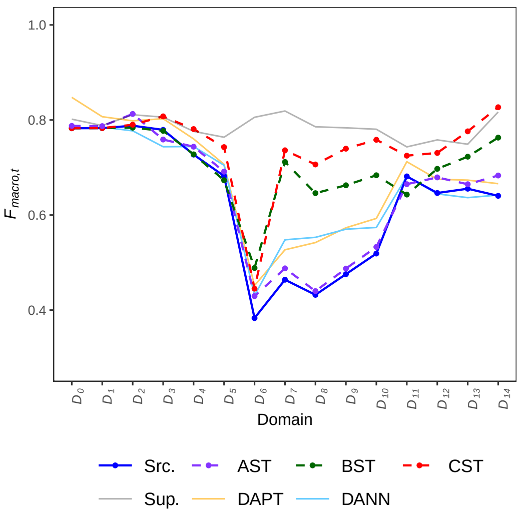


Figure 2: Fine-tuning with self-training over domains on the WTWT dataset. The x-axis denotes the domains, beginning with the source domain and all subsequent target domains. The Macro-F1 score for each domain are plotted on the y-axis. Five different fine-tuning methods are represented in this plot: Src-Only (blue), BST (green), CST (red), AST (purple), and Supervised (grey). The actual time stamp of the start of each domain is appended to the domain name in the format of (YY-MM).

C COVID-19 Vaccination (COVID) Dataset

C.1 Data Retrieval

We collected Twitter data via Twitter Academic API 2.0 endpoint using a list of keywords related to COVID-19 vaccination in English (see Table 6 for details). The time frame of the dataset is from

December 1, 2020 to June 30, 2022. We took a random sample on daily basis for human annotation in terms of valence classification (N = 9,373) into two levels, “against” or “not-against” COVID-19 vaccination.

C.2 Task Definition and Annotation Guidelines

The main goal of the human annotation is to identify the valence toward COVID-19 vaccination of each tweet. We classified the valence into two categories, which included “against” and “non-against” labels in terms of COVID-19 vaccination:

1. The “Against” label can be a) the author of the tweet is personally against COVID-19 vaccines (anti-vaccine) or vaccination policies; b) the tweet message indicates negative consequences of COVID-19 vaccination, such as severe side effects or health misinformation; etc.

E.g., “*Pfizer’s Covid Jabs Shown to Decrease Male Fertility for Months After Vaccination*” and “*My son-in-law committed suicide today. He was vaxxed, boosted x 2. Started losing weight and lost control of his bladder, had to be catheterized. He weighed 149 lbs at his death. Tests were pending to see what was wrong. He left a 11 yr old son and 5 yr old daughter.*”

2. The “Not-against” label can be a) the author of the tweet personally supports or promotes COVID-19 vaccines (pro-vaccine); b) the tweet message reports positive news of COVID-19 vaccines; or c) the tweet is ambiguous to identify its valence.

E.g., “*Good morning. Please get vaccinated*” and “*By the way, vaccination is not a ‘deeply personal decision.’ It is a routine public health requirement in a civilized society.*”

C.3 Data Annotation

Eight volunteers pursuing undergraduate studies were recruited to annotate the Twitter data, with each tweet being annotated by three different annotators. Prior to the annotation task, the annotators underwent a comprehensive training process. The annotation process took place over nine months, from December 2021 to August 2022.

The annotation task was divided into two main steps: 1) Relevancy: Annotators first determined whether each tweet was relevant to the subject of COVID-19 vaccination. This acted as a screening question to filter out unrelated content. 2) Stance: If a tweet was deemed relevant, the annotators were required to assign it an "against" or "non-against" label. Out of the original 9,373 tweets, 5,002 were considered relevant to COVID-19 vaccination. In instances where the three annotators disagreed on the coding of a tweet, a majority vote rule was applied to reach a final decision. This rule was chosen due to its efficacy in resolving disagreements in human annotations.

C.4 Quality Assessment

Given that each tweet was annotated by three annotators from a team of eight, we calculated the inter-coder reliabilities for each three-person sub-team (see Table 3 for details). The resulting weighted-average Krippendorff’s alpha, weighted by the number of samples annotated by each sub-team, was 0.64. This Krippendorff’s alpha is deemed acceptable as it exceeds 0.6 **?**Landis and Koch (1977).

To evaluate the quality of the majority vote rule, one of the authors, an expert in this field, randomly selected 300 tweets and provided expert annotations as gold labels. The comparison between the annotators’ labels and the gold labels resulted in an accuracy (percentage agreement) of 89.7% indicating a high level of concordance and thereby affirming the reliability of our annotation process (Table 4 shows the agreement matrix).

D Data Preprocessing

D.1 Partitioning the Data Chronologically

To study evolving domain adaptation (EDA), we partitioned the dataset in a chronological manner. This approach ensured the oldest instances served as the labeled source domain, while subsequent

Table 2: The keyword list for COVID-19 vaccine Twitter data collection

vaccine, vaccines, vaccination, vaccinations, vaccinate, vaccinated, vax, vaxx, vaxxx, vaxxed, covax, shot, shots, dose, doses, covidvaccine, covid19vaccine, coronavaccine, coronavirusvaccine, covaxin, mrna, nvic, booster, boosters, pfizer, moderna, gamaleya, "oxford-astrazeneca", astrazeneca, cansino, "johnson & johnson", "j&j", "j & j", "vector institute", novavax, sinopharm, sinovac, "bharat biotech", janssen, cepi, biontech, sputnikv, bektop, zfsw, nvic, pfizerbiontech, "biontechvaccine", "warp speed", "delta variant", oxfordvaccine, pfizervaccine, pfizercovidvaccine, modernavaccine, modernacovidvaccine, biotechvaccine, biotechcovidvaccine, biontechvaccine, biontechcovidvaccine, bektopvaccine, simopharmvaccine, johnsonvaccine, janssenvaccine, azvaccine, astrazenecacovidvaccine, astrazenecavaccine, thisisourshot, vaxhole, notocoronavirusvaccines, getvaccinated

Table 3: Krippendorff's alphas (α) for annotator teams

Sub-team Index	Percentage of Tweets (%)	α
1	3%	0.56
2	14%	0.56
3	11%	0.64
4	33%	0.64
5	15%	0.70
6	5%	0.61
7	4%	0.55
8	14%	0.69
Weighted α	100%	0.64

Table 4: Agreement Matrix: Annotations' Label vs Expert Gold Label

		Expert Gold Label	
		Not Against	Against
Annotators' Label	Not Against	193	8
	Against	23	76

instances were arranged into a series of target domains based on their timestamps. This alignment, by natural time units like months instead of fixed-size partitions, is designed to emulate real-world scenarios. For the COVID dataset, we used one-month units for partitioning, and for the WTWT dataset, we used two-month units, with a few exceptions, e.g., period from February to March 2022 was mergd into a single domain to ensure an adequate number of instances for reliable evaluation.

D.2 Source and Target Domains of the COVID-19 Vaccination Dataset (COVID)

For the COVID dataset, instances ranging from December 2020 to May 2021 were merged to form the source domain. The rest of the instances from June 2021 to June 2022 were used to create 10 target domains using a one-month interval. For the detailed correspondence between times and domain, please refer to Table 5.

D.3 Source and Target Domains of the Will-They-Won't-They Dataset (WTWT)

For the WTWT dataset, instances from June 2015 to June 2016 were combined to create the source domain. The remaining instances from July 2016 to December 2018 were used to create 14 target domains using a two-month interval. For the detailed correspondence between times and domain, please refer to Table 6.

D.4 Training, Validation, Testing Partitions

Each domain was further divided into training, validation, and test sets in a 5:1:4 ratio. The testing set was not used during the training process, and is only used to evaluate the model performance at each domain. Only instances from the training and validation sets were pseudolabeled if the method required pseudolabeling. The validation set was used for fine-tuning, allowing us to select the best epoch checkpoint (see Appendix G).

E Label and Topic Distribution over Domains

Table 5, 6, 7 , Figure 3 presents the label and topic distribution shifts across different domains.

For COVID Dataset (Table 5, Figure 3a), the proportion of labels related to COVID-19 vaccines increased over domains.

The proportions of labels in the WTWT Dataset (Table 6, Figure 3b) exhibits strong shifts in label proportions. The portions of each labels all fluctuated significantly.

The proportions of topics in the WTWT dataset (Table 7, Figure 3c) also demonstrate significant shifts. AFT v.s. HUM and ANTM v.s. CI are the majority of the topics in the early domains, then the proportion FDXA v.s. DIS become the majority of the topics until $\mathcal{D}_{11}$. After that, the portion of CI v.s. ESRX and CSV v.s. AET increased and CVS v.s. AET dominates the topic at last.

Table 5: Label Distribution of the COVID Dataset across Domains

Domain Type	Domain	Against	Not-against	Total
		Count (%)	Count (%)	
Source Domain	$\mathcal{D}_0$: 2020-12 to 2021-05	224 (14.34%)	1338 (85.66%)	1562
Target Domains	$\mathcal{D}_1$: 2021-06	51 (22.67%)	174 (77.33%)	225
	$\mathcal{D}_2$: 2021-07	109 (25.77%)	314 (74.23%)	423
	$\mathcal{D}_3$: 2021-08	132 (26.4%)	368 (73.6%)	500
	$\mathcal{D}_4$: 2021-09	160 (37.04%)	272 (62.96%)	432
	$\mathcal{D}_5$: 2021-10	157 (46.73%)	179 (53.27%)	336
	$\mathcal{D}_6$: 2021-11	117 (37.86%)	192 (62.14%)	309
	$\mathcal{D}_7$: 2021-12	142 (36.41%)	248 (63.59%)	390
	$\mathcal{D}_8$: 2022-01	144 (40.22%)	214 (59.78%)	358
	$\mathcal{D}_9$: 2022-02 to 2022-03	132 (56.65%)	101 (43.35%)	233
	$\mathcal{D}_{10}$: 2022-04 to 2022-06	133 (56.84%)	101 (43.16%)	234
Total	-	1501 (30.01%)	3501 (69.99%)	5002

Table 6: Label Distribution of the WTWT Dataset across Domains

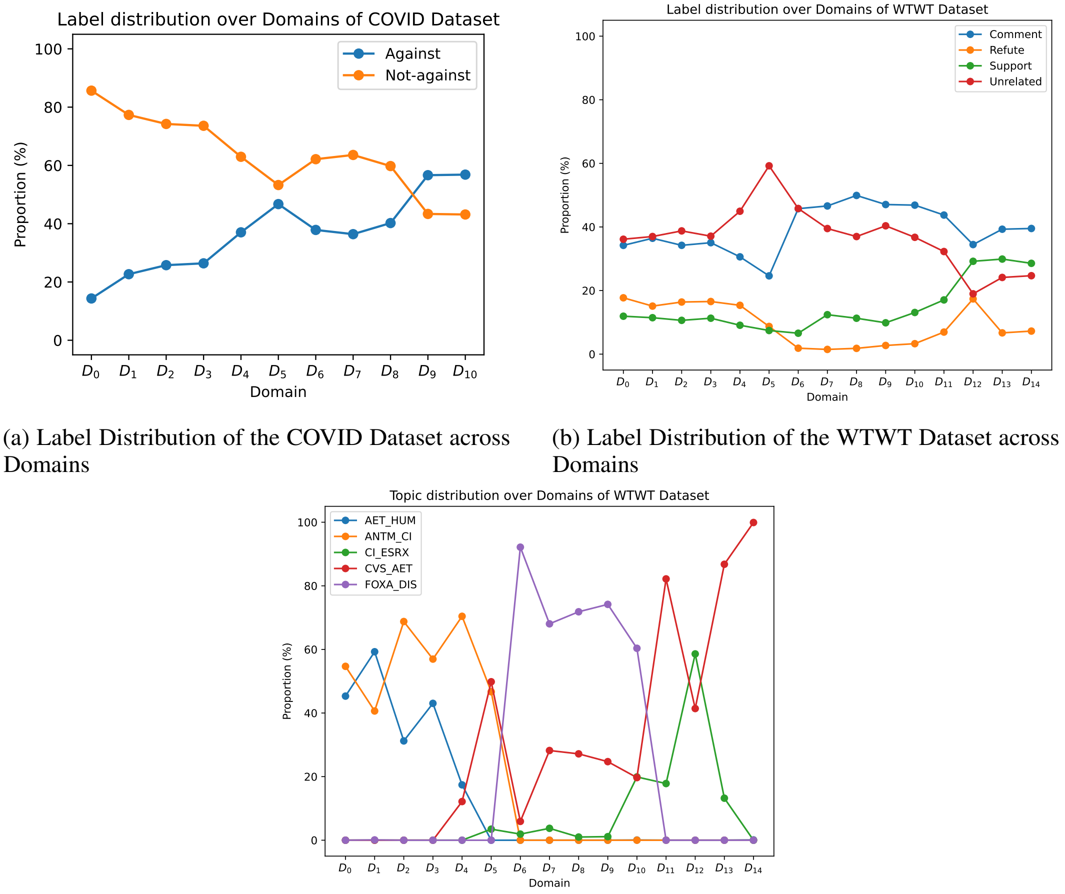
Domain Type	Domain	Comment	Refute	Support	Unrelated	Total
		Count (%)	Count (%)	Count (%)	Count (%)	
Source Domain	$\mathcal{D}_0$: 2015-06 to 2016-06	2923 (34.21 %)	1515 (17.72%)	1020 (11.93%)	3087 (36.14%)	8545
Target Domains	$\mathcal{D}_1$: 2016-07 to 2016-08	798 (36.44%)	331 (15.11%)	251 (11.46%)	810 (36.99%)	2190
	$\mathcal{D}_2$: 2016-09 to 2016-10	309 (34.22%)	148 (16.39%)	96 (10.63%)	350 (38.76%)	903
	$\mathcal{D}_3$: 2016-11 to 2016-12	307 (35.05%)	145 (16.55%)	99 (11.3%)	325 (37.1%)	876
	$\mathcal{D}_4$: 2017-01 to 2017-02	239 (30.6%)	120 (15.36%)	71 (9.09%)	351 (44.94%)	781
	$\mathcal{D}_5$: 2017-03 to 2017-06	212 (24.62%)	75 (8.71%)	64 (7.43%)	510 (59.23%)	861
	$\mathcal{D}_6$: 2017-07 to 2017-08	904 (45.75%)	37 (1.87%)	130 (6.58%)	905 (45.8%)	1976
	$\mathcal{D}_7$: 2017-09 to 2017-10	1247 (46.6%)	40 (1.49%)	332 (12.41%)	1057 (39.5%)	2676
	$\mathcal{D}_8$: 2017-11 to 2017-12	4782 (49.91%)	174 (1.82%)	1082 (11.29%)	3543 (36.98%)	9581
	$\mathcal{D}_9$: 2018-01 to 2018-02	2342 (47.05%)	136 (2.73%)	491 (9.86%)	2009 (40.36%)	4978
	$\mathcal{D}_{10}$: 2018-03 to 2018-04	2006 (46.86%)	141 (3.29%)	561 (13.1%)	1573 (36.74%)	4281
	$\mathcal{D}_{11}$: 2018-05 to 2018-06	454 (43.74%)	72 (6.94%)	177 (17.05%)	335 (32.27%)	1038
	$\mathcal{D}_{12}$: 2018-07 to 2018-08	367 (34.46%)	185 (17.37%)	311 (29.2%)	202 (18.97%)	1065
	$\mathcal{D}_{13}$: 2018-09 to 2018-10	523 (39.29%)	89 (6.69%)	398 (29.9%)	321 (24.12%)	1331
	$\mathcal{D}_{14}$: 2018-11 to 2018-12	447 (39.52%)	82 (7.25%)	323 (28.56%)	279 (24.67%)	1131
Total	-	17860 (42.31%)	3290 (7.79%)	5406 (12.81%)	15657 (37.09%)	42213

F Prompt Tempate

Table 8 and 9 show the prompt templates used the experiment across the two datasets.

Table 7: Topic Distribution of the WTWT Dataset across Domains

Domain Type	Domain	AET_HUM Count (%)	ANTM_CI Count (%)	CI_ESRX Count (%)	CVS_AET Count (%)	FOXA_DIS Count (%)	Total
Source Domain	$\mathcal{D}_0$: 2015-06 to 2016-06	3873 (45.32%)	4672 (54.68%)	0 (0%)	0 (0%)	0 (0%)	8545
Target Domains	$\mathcal{D}_1$: 2016-07 to 2016-08	1298 (59.27%)	890 (40.64%)	0 (0%)	0 (0%)	2 (0.09%)	2190
	$\mathcal{D}_2$: 2016-09 to 2016-10	282 (31.23%)	621 (68.77%)	0 (0%)	0 (0%)	0 (0%)	903
	$\mathcal{D}_3$: 2016-11 to 2016-12	377 (43.04%)	499 (56.96%)	0 (0%)	0 (0%)	0 (0%)	876
	$\mathcal{D}_4$: 2017-01 to 2017-02	136 (17.41%)	550 (70.42%)	0 (0%)	95 (12.16%)	0 (0%)	781
	$\mathcal{D}_5$: 2017-03 to 2017-06	0 (0%)	402 (46.69%)	30 (3.48%)	429 (49.83%)	0 (0%)	861
	$\mathcal{D}_6$: 2017-07 to 2017-08	0 (0%)	0 (0%)	38 (1.92%)	117 (5.92%)	1821 (92.16%)	1976
	$\mathcal{D}_7$: 2017-09 to 2017-10	0 (0%)	0 (0%)	100 (3.74%)	755 (28.21%)	1821 (68.05%)	2676
	$\mathcal{D}_8$: 2017-11 to 2017-12	0 (0%)	0 (0%)	96 (1%)	2602 (27.16%)	6883 (71.84%)	9581
	$\mathcal{D}_9$: 2018-01 to 2018-02	0 (0%)	0 (0%)	56 (1.12%)	1230 (24.71%)	3692 (74.17%)	4978
	$\mathcal{D}_{10}$: 2018-03 to 2018-04	2 (0.05%)	0 (0%)	852 (19.9%)	843 (19.69%)	2584 (60.36%)	4281
	$\mathcal{D}_{11}$: 2018-05 to 2018-06	0 (0%)	0 (0%)	185 (17.82%)	853 (82.18%)	0 (0%)	1038
	$\mathcal{D}_{12}$: 2018-07 to 2018-08	0 (0%)	0 (0%)	624 (58.59%)	441 (41.41%)	0 (0%)	1065
	$\mathcal{D}_{13}$: 2018-09 to 2018-10	0 (0%)	0 (0%)	176 (13.22%)	1155 (86.78%)	0 (0%)	1331
	$\mathcal{D}_{14}$: 2018-11 to 2018-12	1 (0.09%)	0 (0%)	0 (0%)	1130 (99.91%)	0 (0%)	1131
Total	-	5969 (14.14%)	7634 (18.08%)	2157 (5.11%)	9650 (22.86%)	16803 (39.81%)	42213



(a) Label Distribution of the COVID Dataset across Domains

(b) Label Distribution of the WTWT Dataset across Domains

(c) Topic Distribution of the WTWT Dataset across Domains

Figure 3: Label distribution of (a) the COVID dataset and (b) the WTWT dataset across domains, and (c) the topic distribution of the WTWT dataset across domains.

Table 8: Zero-Shot Prompt Templates

Dataset	Components	Contents
COVID	x_i	breaking report: cdc used rejected study from india on vaccine, not approved in the us to justify new mask mandate...
	$\mathcal{T}(x_i)$	What is the stance of the tweet below with respect to COVID-19 vaccine? Please use exactly one word from the following 2 categories to label it: 'against', 'not-against'. Here is the tweet. 'breaking report: cdc used rejected study from india on vaccine, not approved in the us to justify new mask mandate...' The stance of the tweet is
WTWT	x_i	feed time... health ins. aetna to acquire humana for 37, anthem and cigna are in talks, centene corp. agreeing to buy health net #uniteblue
	$\mathcal{T}(x_i)$	What is the stance of the tweet below with respect to the probability of a merger and acquisition (M&A) operation occurring between two companies? If the tweet is supporting the theory that the merger is happening, please label it as 'support'. If the tweet is commenting on the merger but does not directly state that the deal is happening or refute this, please label it as 'comment'. If the tweet is refuting that the merger is happening, please label it as 'refute'. If the tweet is unrelated to the given merger, please label it as 'unrelated'. Here is the tweet. 'feed time... health ins. aetna to acquire humana for 37, anthem and cigna are in talks, centene corp. agreeing to buy health net #uniteblue' The stance of the tweet is:

G Training Procedure and Hyperparameters for Fine-tuning

This section provides the details of the fine-tuning procedure and hyperparameters used during our experiments. The experiment was conducted using Python version 3.10.6, Huggingface **?** version 4.27.0, and PyTorch 1.13.1 on a GPU machine equipped with 4x NVIDIA GeForce RTX 3090.

Hyperparameters were determined using the validation set. We used AdamW optimizer Loshchilov and Hutter (2019), along with the following hyperparameters:

- Learning rate: 2×10^{-5}
- Weight decay: 0.01
- Mini-Batch size: 32
- Dropout rate: 0.1

The chosen hyperparameters were kept constant throughout the experiments to ensure fair comparisons between different model configurations.

The models were fine-tuned over a maximum of three epochs. For any $t \in [0, T]$, the fine-tuned model was selected based on the epoch checkpoint that produced the highest $F_{macro,t}$ score on the validation set.

In all the fine-tuning methods, the instances in the training set (labeled or pseudo-labeled) are used for training the model, and the instances (labeled or pseudo-labeled) in the validation set are used for selecting the best epoch checkpoint when training models. The instances in the testing set is held out throughout the entire training.

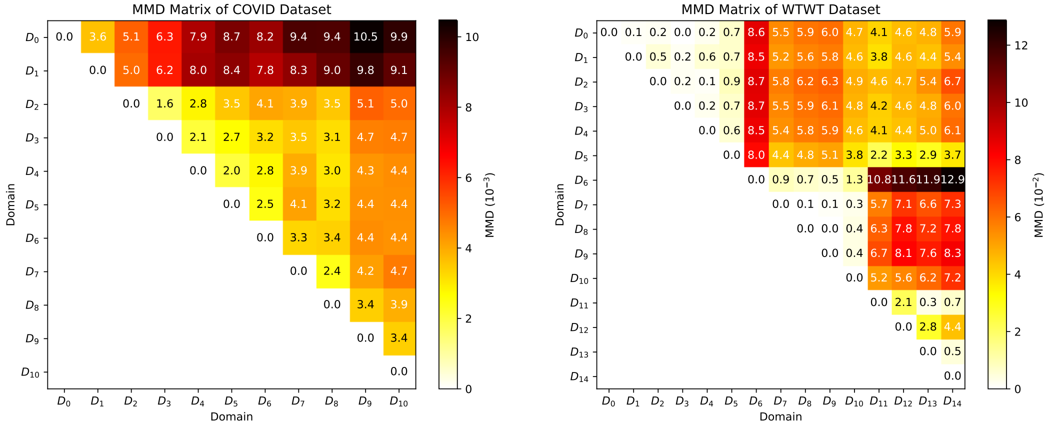
H Detailed Domain Divergence Analysis

To better understand the performance trends observed in §**??**, we visualize the divergence between all pairs of domains, quantified by the MMD Gretton et al. (2012) between their embeddings.

Critically, for both datasets, the MMD between two adjacent domains is almost always smaller than the MMD between the source and any target domain. This observation supports our earlier findings on the advantage of maintaining a dynamic buffer over relying on static source domain data for adapting to evolving domains.

Table 9: Few-Shot Prompt Templates

Dataset	Components	Contents
COVID	x_i	breaking report: cdc used rejected study from india on vaccine, not approved in the us to justify new mask mandate...
	$\mathcal{C}$	(if someone stops you from getting a vaccine by blocking the entrance to a vaccination site, they should be arrested. period, not-against) . . . , (every life matters until the deaths result from vaccines. then they are justified., against)
	$\mathcal{T}(x_i,\mathcal{C})$	What is the stance of the tweet below with respect to COVID-19 vaccine? Please use exactly one word from the following 2 categories to label it: ‘against’, ‘not-against’. Here are some examples of tweets. Make sure to classify the last tweet correctly. Q: Tweet: if someone stops you from getting a vaccine by blocking the entrance to a vaccination site, they should be arrested. period. Is this tweet against or not-against? A: not-against Q: Tweet: every life matters until the deaths result from vaccines. then they are justified. Is this tweet against or not-against? A: against Q: Tweet: breaking report: cdc used rejected study from india on vaccine, not approved in the us to justify new mask mandate... Is this tweet against or not-against? A:
WTWT	x_i	feed time... health ins. aetna to acquire humana for 37, anthem and cigna are in talks, centene corp.
	$\mathcal{C}$	(cvs health amp; aetna working to finalize their merger as early as december, reports @USERNAME: #pharmacy, support) . . . , (i’m watching the disney version of robin hood someone tell me how i have a crush on a cartoon fox, unrelated)
	$\mathcal{T}(x_i,\mathcal{C})$	What is the stance of the tweet below with respect to the probability of a merger and acquisition (M&A) operation occurring between two companies? If the tweet is supporting the theory that the merger is happening, please label it as ‘support’. If the tweet is commenting on the merger but does not directly state that the deal is happening or refute this, please label it as ‘comment’. If the tweet is refuting that the merger is happening, please label it as ‘refute’. If the tweet is unrelated to the given merger, please label it as ‘unrelated’. Please use exactly one word from the following 4 categories to label it: ‘support’, ‘comment’, ‘refute’, ‘unrelated’. Here are some examples of tweets. Make sure to classify the last tweet correctly. Q: Tweet: cvs health amp; aetna working to finalize their merger as early as december, reports @USERNAME: #pharmacy Is this tweet ‘support’, ‘comment’, ‘refute’, or ‘unrelated’? A: support Q: Tweet: cigna-express scripts deal unlikely to benefit consumers #healthnews #health Is this tweet ‘support’, ‘comment’, ‘refute’, or ‘unrelated’? A: comment Q: Tweet: business: just in: cigna terminates merger agreement with anthem Is this tweet ‘support’, ‘comment’, ‘refute’, or ‘unrelated’? A: refute Q: Tweet: i’m watching the disney version of robin hood someone tell me how i have a crush on a cartoon fox Is this tweet ‘support’, ‘comment’, ‘refute’, or ‘unrelated’? A: unrelated Q: Tweet: feed time... health ins. aetna to acquire humana for 37, anthem and cigna are in talks, centene corp. agreeing to buy health net #uniteblue Is this tweet ‘support’, ‘comment’, ‘refute’, or ‘unrelated’? A:



(a) COVID Dataset

(b) WTWT Dataset

Figure 4: Maximum Mean Discrepancy (MMD) matrix for (a) the COVID dataset and (b) the WTWT dataset. Each cell represents the MMD between a pair of domains, calculated based on the marginal distribution of text embeddings $P_{g(\mathcal{X})}$ projected by the E5-Large-V2 model. The color gradient ranges from white (representing zero discrepancy) to darker shades of red (indicating larger discrepancies). Please note that the numbers in the heatmap are in the units of 10^{-3} and 10^{-2} for (a) and (b), respectively.

COVID Dataset Figure 4a showcases the domain shift over the marginal distribution of each domain. The domain shift is more gradual and uni-directional. The MMD between the source domain and the subsequent target domains increases over time, aligning with the gradual decay observed in the Src-Only baseline in Figure 1a. This indicates a gradual evolution of the COVID discussion space over time, with newer content becoming progressively more distinct from the initial discussions.

WTWT In contrast, the domain shift in the WTWT dataset (Figure 4b) is more abrupt, with notable discontinuities. Two most substantial discontinuities occur around the 17-07 domain and another around the 18-05 domain, in line with the abrupt shift in topic distribution as shown in Table 7. Notably, we observe an inverted-U shape trend when using the source domain as an anchor point: the MMD between the source domain and any given target domain peaks around 17-07 and subsequently declines. This pattern is consistent with the changes in the topic distribution, especially the rise and fall of the FOXA v.s. DIS topic around this period. As the FOXA v.s. DIS topic, related to events in the entertainment sector, is significantly different from the other four topics, all in the pharmaceutical sector, it contributes to the inverted-U shape in domain shift. This also aligns with the performance trend of the Src-Only baseline model in Figure 2.