Tackling Concept Shift in Text Classification using Entailment-style Modeling

Sumegh Roychowdhury*
Amazon
India
sumegr@amazon.com

Karan Gupta*
Amazon
India
karaniis@amazon.com

Siva Rajesh Kasa*
Amazon
India
kasasiva@amazon.com

Prasanna Srinivasa Murthy
Amazon
India
sprsn@amazon.com

ABSTRACT

Pre-trained language models (PLMs) have seen tremendous success in text classification (TC) problems in the context of Natural Language Processing (NLP). In many real-world text classification tasks, the class definitions being learned do not remain constant but rather change with time - this is known as **concept shift**. Most techniques for handling concept shift rely on retraining the old classifiers with the newly labelled data. However, given the amount of training data required to fine-tune large DL models for the new concepts, the associated labelling costs can be prohibitively expensive and time consuming. In this work, we propose a reformulation, converting vanilla classification into an entailment-style problem that requires significantly less data to re-train the text classifier to adapt to new concepts. We demonstrate the effectiveness of our proposed method on both real world & synthetic datasets achieving absolute F1 gains upto ∼**6%** and ∼**30%** respectively in few-shot settings. Further, upon deployment, our solution also helped save **75%** direct labeling costs and **40%** downstream labeling costs overall in a span of 3 months.

KEYWORDS

e-commerce, concept shift, text classification, textual entailment

ACM Reference Format:
Sumegh Roychowdhury, Karan Gupta, Siva Rajesh Kasa, and Prasanna Srinivasa Murthy. 2024. Tackling Concept Shift in Text Classification using Entailment-style Modeling. In *Proceedings of ACM Conference (KDD'24).* ACM, Barcelona, Spain, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Concept shift occurs in a data stream when the contextual relationship between textual inputs (X) and labels (Y) changes with time, resulting in a change in the conditional probability of Y given X. Concept shift has been studied extensively in classical machine learning problems [16–18, 21]. Concept shift in text classification can occur due to various reasons such as sudden change in external circumstances (e.g. occurrence of pandemics [26]), gradual change in semantics (e.g. words taking on newer meanings [2]), etc. Consider the case of classifying a healthcare research article as relevant/irrelevant to a user. Under the usual circumstances, the user may not be interested in healthcare research but during pandemics, his/her preferences could change. Depending on the direction and extent of concept shift, re-training of the classifier models is essential to match the changing concepts in the stream. If concept shift is not kept track of, it can lead to malfunction of downstream systems, loss of revenue, erosion of trust in the classifier systems, etc.

In the context of text classification (TC) using PLMs, tackling concept shift requires relabelling of the datapoints under the new concepts. While PLMs demonstrate few-shot learning capabilities, the amount of labelled data required primarily depends on the difficulty of the new concepts to be learnt and on the size of the PLM. As mentioned before, obtaining new labelled data is often laborious, time consuming and expensive in real-world settings. In this paper, we propose to reformulate this task into an entailment-style approach [31] to overcome this data requirement. Further, there is a widespread consensus on the lack of real-world benchmark datasets for TC task to evaluate the various methods proposed for tackling concept-shift [9, 21, 27], and hence, in addition to synthetic datasets, we also curate a real world dataset to study concept shift and benchmark our model performances on the same. To the best of our knowledge, no prior work has explored tackling sudden concept shift in the context of textual classification data. Our main contributions are as follows:

- We show that reformulating vanilla classification as an **entailment style task** can help augment PLMs performance in tackling concept shift also leading to **significant cost savings** post-deployment.
- We provide ablations to highlight PLMs' ***few-shot capabilities for tackling concept shift*** in text classification. Our proposed approach leads to significantly better performance vis-a-vis a vanilla finetuning of PLMs, under the new concepts, as demonstrated on both real & synthetic datasets.

*Equal Contribution.

KDD'24, August 2024, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

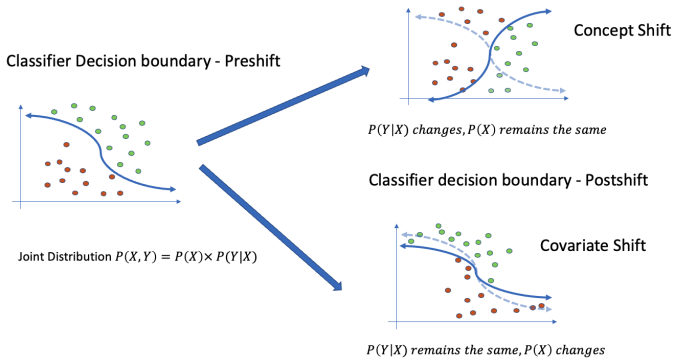


Figure 1: [Best viewed in color] Decision boundary of a trained classifier is no longer effective if the joint distribution of the new data is different from the training data. If the feature distribution P(X) changes, we have covariate shift. If the conditional distribution P(Y |X) changes, we have concept shift. It is possible that both these shifts can occur simultaneously.

We also curate the ***RetailQueries Shift dataset*** for benchmarking concept shift research in the text-classification domain and report results on the same as well. To the best of our knowledge, this is the first benchmark real-world multilingual dataset for studying concept shift in text classification. Since it's proprietary data, we cannot make it public without proper approvals and licensing. We will work on releasing the dataset in future iterations.

2 BACKGROUND AND RELATED WORK

Let $\mathbf{X}$ be the features (aka covariates) and $\mathbf{y}$ be the corresponding labels such that the class labels $\mathbf{y}$ are causally determined by the covariates $\mathbf{X}$ - e.g. email spam classification task where the contents of the email represented in the covariate space $\mathbf{X}$ determines the class label $\mathbf{y}$: whether it is spam or not [25]. A common assumption when deploying supervised machine learning models is that joint distribution of features and labels $\mathcal{P}(\mathbf{X}, \mathbf{y})$ remains the same during the training (tr) and testing (tst) stages. When this assumption is violated i.e. $\mathcal{P}_{tr}(\mathbf{X}, \mathbf{y}) \neq \mathcal{P}_{tst}(\mathbf{X}, \mathbf{y})$, we say there is a distribution shift. Using Bayes theorem, the joint distribution can be written as product of $\mathcal{P}(\mathbf{X})\mathcal{P}(\mathbf{y}|\mathbf{X})$. Depending on which of these two distributions differ between training and inference stage, there are primarily two kinds of shifts,

- **Covariate Shift:** If $\mathcal{P}_{tr}(\mathbf{X}) \neq \mathcal{P}_{tst}(\mathbf{X})$ and $\mathcal{P}_{tr}(\mathbf{y}|\mathbf{X}) = \mathcal{P}_{tst}(\mathbf{y}|\mathbf{X})$
- **Concept Shift:** If $\mathcal{P}_{tr}(\mathbf{X}) = \mathcal{P}_{tst}(\mathbf{X})$ and $\mathcal{P}_{tr}(\mathbf{y}|\mathbf{X}) \neq \mathcal{P}_{tst}(\mathbf{y}|\mathbf{X})$

Often times, it possible that both these shifts can occur together. Concept shift is relatively understudied as compared to covariate shift, nevertheless it is still an important problem to consider while deploying machine learning classifiers [42]. We refer to [25] for a comprehensive introduction to the different data shifts that can occur in real-world applications. In this work, we focus on tackling concept shift. Also, based on the speed of change, concept shift can manifest itself in two major forms - *gradual shift* and *sudden shift*[21, 27, 34]. Sudden Shift occurs when there are abrupt changes in the concept. For example, in a news filtering usecase, the break-out of a deadly pandemic can make articles about vaccines/biotech advancements more relevant for the user, which were previously irrelevant. Gradual Shift, as the name refers to, is a more slow, continuous and evolved change in the concept. An example of Gradual Shift is how meanings of words evolve over time. For example, the meanings of the words 'BERT' and 'ELMO' have changed from persons to neural network architectures in the recent past [2]. Although there are other kinds of shifts like incremental shifts, recurring shifts, etc. in this work, **we restrict ourselves to the sudden shift case** in the context of text classification as this is relatively more challenging because the timeline to acquire labelled data is more stringent. Note that concept shift should not be confused with label shift where the shift is distributionally similar but the class labels causally determine the covariates - e.g. in author attribution tasks, the author ($\mathbf{y}$) determines the style of the text ($\mathbf{X}$) [15, 22].

One of the straightforward ways to tackle concept shift is to retrain the model from scratch using the new data, as and when there is a concept shift. If there is no explicit information that a concept shift has occurred, then monitoring using a concept drift detector may be required to decide when to retrain the model [23]. However, training PLMs requires that there is enough labelled data [4, 5, 7, 29, 30, 36, 38] which is expensive and time consuming to obtain. Previous approaches to tackle concept shift have relied on giving importance to datapoints based on their recency, either through a weightage/discounting factor or only considering a recent window of data whilst discarding everything prior to that window [6, 17]. In weightage based techniques, the instances are weighted based on recency and diversity using the ability of certain algorithms like Support Vector Machines to process weighted instances [6]. Instances can be weighted according to their age and their competence with regard to the current concept. A longer window is more suitable for the gradual shift case, where as a shorter window is more effective for fast changing environments. The time window can be fixed or adaptive [41]. In the context of PLMs, continual fine-tuning as and when a new batch of data arrives can considered as a sliding-windowing approach. We use this as one of the baselines in our paper (see Section 8). In the sliding window method, the process of choosing which points to be labelled for re-training of models can be done more judiciously. This is a classic

RetailQueries Shift Data (Real)	
Query/Product	**Pre-shift → Post-shift**
2 year old child cloths / Pants Outfit 3-4 Years old	Irrelevant → Substitute (Size/Age)
peppa pig muddy puddle / plush toy peppa pig	Irrelevant → Substitute (Brand)
Along for the ride / me before you	Irrelevant → Substitute (Genre)
dove soap / nivea body wash nourishing, 3x500ml	Complement → Substitute (Utility)
lamp for desk / Clamp desk lamp	Exact → Exact (No Change)
AGNews Shift Data (Synthetic)	
News Title	**Pre-shift → Post-shift**
Australia march into dominant position against NZ	Irrelevant → Relevant (Sports)
Timing Of Indian Move In Kashmir Vital: Pak Paper	Relevant → Irrelevant (World)
Stocks to Watch Tuesday	Relevant → Relevant (Business)

Table 1: A few anecdotal examples from the datasets we use for our study. For *(top)* RetailQueries, we provide the reasons for shift based on labeling guideline change. For *(bottom)* AGNews, we provide the topics under which shift happens.

case of exploration-exploitation dilemma. Usually, an active learning framework is employed to choose datapoints based on how beneficial are those instances from a classifier training perspective [21]. This is however not relevant in this paper because we assume access to new labels after the concept shift has occurred for training / testing purposes.

Also methods like Adaptive RF [10] and Adaptive XGBoost [24] are have also been proposed for tackling incremental shifts. More importantly, these methods are no longer state-of-the-art in the text classification compared to deep learning based techniques after the advent of the Transformer [38] architecture, hence we do not consider these methods as baseline for comparison. Another line of work involves meta-learning based methods like RCD [11] and CPF [1] which help solve recurrent shift problems by reusing one of the many base classifiers at any given time acting as meta-learners. However, this adds more tune-able parameters which is why more recent approaches propose continuous parameter tuning for non-stationary data streams [39]. These approaches are also not valid baselines for our study since we focus on tackling sudden concept shifts. For a more detailed review, we refer to the study by [23] and [9]. To the best of our knowledge, **no prior work** has explored tackling sudden concept shift in the context of textual classification data.

Further the advent of pre-trained language models (PLMs) led to the rise of prompt-based learning techniques. It started as a way to probe knowledge from language models by posing questions as fill-in-the-blank [28] tasks without adding any extra parameters. Initially, it started with manually designed prompts [20][46] but then shifted towards automated prompt designs. [33] proposed AutoPrompt which used gradient-based search to find suitable tokens for the prompt. [44] developed DifferentialPrompt where instead of having discrete prompt tokens it's projected into a continuous space and jointly optimized. But these methods are known to be very unstable to seeds & datasets especially in low-data settings. In another parallel line of work, [43] showed that PLMs can be used to do 0-shot text classification for unseen/partially seen labels by formulating it as a textual entailment problem. We have a different setup in our case where the label space remains constant pre-shift and post-shift (only labels concepts change due to change in labeling guidelines). We build upon the above technique to show that this can be used to tackle concept shift in textual data by describing the change in concepts using natural language verbalizers enabling the model to implicitly figure out the change in label.

3 LACK OF BENCHMARK DATASETS

Proposing text classification techniques for tackling concept shift has been problematic due to the dearth of real-world datasets [21, 27]. There are mainly two kinds of datasets used for benchmarking methods proposed for detecting and tackling concept shift - real and artificial datasets.

In real world datasets, one of the most widely cited papers for tackling concept shift in text streams is the Email spam dataset [6]. However, we believe the dataset is still not publicly available [21] as it may include personal/financial information. Further, there are only two labels in the dataset and the size of the dataset is also small. More importantly, the reason for why and when and how much concept shift has happened is also not discussed in the work. A recent survey [9] showed that there are **no existing real-world benchmark datasets** for studying concept shift in case of text streams. To address this gap we curate the *RetailQueries Shift* dataset in §4 to serve as benchmark for future research in tackling concept shift (to be released upon acceptance).

In artificial datasets, there are two subtypes of datasets - *drift-induced* datasets and *synthetic* datasets. The drift is induced by artificially changing the labels after a period of time e.g. by restricting the subset of relevant topics for each particular period of time [16, 19, 21, 37]. In synthetic datasets, the duration and quality of the shift can be tweaked based on frameworks such as STAGGER [32], the moving hyperplane [8, 13, 18] and Narasimhamurthy's framework [27]. As noted in [6], these methods are not are scalable for generating large datasets. Instead we adopt the approach presented in [21] and create an artificial drift-induced dataset based on a popular news classification task (refer to *AGNews Shift* dataset in Section §4).

4 DATASET DETAILS

RetailQueries Shift - For curating this data, we randomly sample 48,738 unique customer search queries across 12 languages from Amazon's retail catalog. For each query, there are a number of products (max depth = 16) retrieved. Overall, we get 62,500 query-product pairs with 60,067 unique products. Each of these (query, product) pairs are classified into 4 relevance categories:

- **Exact (E)** - product fully satisfies the query intent.
- **Substitute (S)** - not exact but functionally equivalent.
- **Complement (C)** - not intended but could be useful.
- **Irrelevant (I)** - does not satisfy query intent.

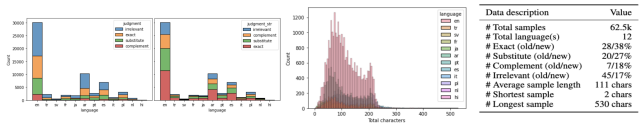
For example, given query "iphone" the product "iphone 11" would be exact as it's the exact product the query intends to search for, "samsung galaxy" would be substitute since it's not an iphone but offers the same functionality, "airpods pro" would be complement since it can be used as an accessory to iphone and "lord of the rings" would be irrelevant as it's totally unrelated to the query.

Now we describe the type of concept shift that occurs in the dataset. Annotators assign relevance labels (E/S/C/I) to query-product pairs based on certain set guidelines. However, in the recent past, based on observed customer behavior, a few changes were introduced in the previous labeling guidelines to better match the customer intent. Some labels changed post-shift due to this guideline change while some remained same. The major changes were:

- *Utility* - The product type intended by the query must be matching with the retrieved product for it to be relevant (E/S) under old guidelines. This condition was relaxed under new guidelines. Now, if a product can provide the same utility the customer is looking for, the query-product pair can be classified as exact or substitute based on suitability.
- *Size/Age* - Under new guideline, certain constraints on product attributes are relaxed. E.g. attributes viz. Size, Age, etc. are relaxed and products at ±1 level will be considered substitute, while under old guidelines, they would be considered irrelevant.
- *Brand* - Under new guidelines, if brand intent matches then the product can be marked complement even if it has a different purpose than what's intended in query.
- *Genre* - For books/media related queries, if the genre/topic matches that of the product, then it can be classified as substitute.

Each of these relevance labels E/S/C/I were evaluated manually by human judges for all 62,500 query-product pairs using the post-shift guidelines. A minimum of 3 annotations were chosen for each pair and majority vote was taken to assign the final gold label. We observe ∼ 92% agreement in post-shift labels across annotators on average. Usually this agreement rate is observed around ∼ 88% for pre-shift production batches from historic data. This is expected since the change in labeling guidelines were made to ensure it resonates better with customer intent which in turn would reduce confusion among annotators for labeling ambiguous samples. Our dataset is also multilingual spanning across 12 languages (ISO-639 codes) - English (en), Spanish (es), French (fr), Turkish (tr), Italian (it), Dutch (nl), Polish (pl), Arabic (ar), Japanese (ja), Portuguese (pt), German (de) and Hindi (hi). This dataset is the first of its kind real-world, large-scale, multi-lingual dataset to study concept shift. See Figure 2 for details.

Note that there might be context-specific exceptions to rules proposed above to create a concept shift. For example - Brand Affinity might not matter for pharmacy related queries where showing different medicines from same brand would be irrelevant again. In future work, a detailed analysis can be done to see how well the



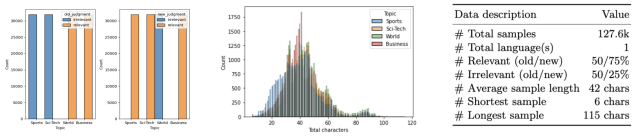
Data description	Value
# Total samples	62.5k
# Total language(s)	12
# Exact (old/new)	28/38%
# Substitute (old/new)	20/27%
# Complement (old/new)	7/18%
# Irrelevant (old/new)	45/17%
# Average sample length	111 chars
# Shortest sample	2 chars
# Longest sample	530 chars

Figure 2: [Best viewed in color] RetailQueries Data *(Left)* Pre & Post-Shift label distribution *(Center)* Character length distribution. *(Right)* Dataset Statistics.

model is able to deal with these edge-cases & how can we further improve it.

AGNews Shift - We use the AGNews Topic Classification [45] dataset where each news article is divided into 4 topics - World, Business, Sports and Sci-Tech. Similar to [21] we synthetically induce concept shift here in the following manner - we say that before the shift news related to topics World and Business are *relevant* to the user while Sports and Sci-Tech are *irrelevant*. After the shift, we switch the labels in a way that Sports and Sci-Tech become *relevant* while World news becomes *irrelevant*. Business news we keep the same label (*relevant*) to induce a bit more complexity in the artificial shift.

Creating this type of shift is realistic because someone's news preferences can change with time. One might be interested in Sports news when the World Cup is going on. But later that might not be relevant. Similarly, someone might start taking interest in World news when a war is going on but otherwise wouldn't have. Hence, a model suggesting news articles to users should be able to adapt to such concept shifts to keep providing relevant suggestions without requiring huge amount of data to re-train under the new/shifted labels. Refer to Figure 3 for details.



Data description	Value
# Total samples	127.6k
# Total language(s)	1
# Relevant (old/new)	50/75%
# Irrelevant (old/new)	50/25%
# Average sample length	42 chars
# Shortest sample	6 chars
# Longest sample	115 chars

Figure 3: [Best viewed in color] AGNews Data *(Left)* Pre & Post-Shift label distribution *(Center)* Character length distribution. *(Right)* Dataset Statistics.

5 PROBLEM FORMULATION

Let the $\mathcal{L} = \{L_1, \ldots, L_K\}$ be the set of all unique labels. We denote the pre-shift finetuning dataset as $\mathcal{D}_{tr}$ of size N. Let $\mathcal{M}_{base}$ be the base PLM which is finetuned on $\mathcal{D}_{tr}$ to build a classifier $\mathcal{M}_{pre}$. At some time instance t, a concept-shift has occurred which rendered the pre-shift model $\mathcal{M}_{pre}$ no longer suitable for use in production. Hence, we start collecting new training data after the concept-shift which we denote as $\mathcal{D}_{tr}^{'}$ of size N'. Now $N' \ll N$ due to the cost and time involved in collecting human annotated data. Therefore, the challenge lies in constructing a classifier $\mathcal{M}_{post}$ that achieves performance comparable to the pre-shift performance of $\mathcal{M}_{pre}$. Note that for any datapoint x in the post-shift era, we either assume access to its pre-shift ground truth label (*AGNews Shift* data) or use the prediction from the pre-shift model $\mathcal{M}_{pre}$ as proxy for pre-shift ground-truth label (*RetailQueries Shift* data).

6 PROPOSED APPROACH

For the concept-shifted data $\mathcal{D}'_{tr}$, during finetuning we do the following entailment-style data augmentation: for a given datapoint x whose pre-shift label is L_k and post-shift label is $L_{k'}$, we create a total of K augmented samples (s_j) with binary labels (0: not entail/1: entail) as follows:

$$s_j = \{x + verbaliser(L_k, L_j),\ \mathbb{1}_{L_{k'}}(L_j)\}\ \forall\ j \in \{1, 2, \ldots, K\} \quad (1)$$

where $\mathbb{1}_{L_{k'}}(.)$ is the indicator function which takes the value one only when the argument is $L_{k'}$, otherwise it is zero. Here, K is the total number of unique labels. '+' here refers to concat() operation (check § 7). Also the $verbaliser()$ template is defined as :

$$verbaliser(L_k, L_j) = \begin{cases} changed\ from\ \{L_k\}\ to\ \{L_j\} & , L_k \neq L_j \\ remained\ \{L_k\} & , else \end{cases} \quad (2)$$

Essentially, we create $(K-1)$ negative samples and 1 positive sample for each datapoint. Finally, the problem reduces to the following natural language inference (NLI) [3] task - Does x entail the $verbaliser(L_k, L_j)$ or not?

After creating these $K \times N'$ augmented examples, we continually finetune $\mathcal{M}_{pre}$ on a binary classification task. Let's call the finetuned model $\mathcal{M}_{post}$. During inference, the predicted label for a datapoint x_i is obtained by taking an argmax over the softmaxed probabilities output by the binary classifier $\mathcal{M}_{post}$ over the augmented samples as follows:

$$\text{prediction}(x_i) = \underset{j \in \{1,\ldots,K\}}{\arg\max}\ \{\mathcal{M}_{post}(x_i + verbaliser(L_k, L_j))\} \quad (3)$$

To illustrate the above approach (see Figure 4), consider the (query, product) pair *(iphone, samsung galaxy)* to be earlier "irrelevant" and now "substitute" in the concept-shifted training data. Then we create 3 negative examples *(iphone + samsung galaxy, changed from irrelevant to exact match, 0)*, *(iphone + samsung galaxy, changed from irrelevant to complement match, 0)* , *(iphone + samsung galaxy, remained irrelevant match, 0)* and 1 positive example *(iphone + samsung galaxy, changed from irrelevant to substitute match, 1)*.

Once the model is fine-tuned on the concept-shifted training data, during inference, when we encounter a new (query, product) pair, say *(headphones, airpods)*, whose pre-shift label is "irrelevant". We compute the output logits of the following augmented inputs:

$$\begin{cases} \text{(headphones, airpods) + verbaliser(irrelevant, exact)} \rightarrow l_e \\ \text{(headphones, airpods) + verbaliser(irrelevant, substitute)} \rightarrow l_s \\ \text{(headphones, airpods) + verbaliser(irrelevant, complement)} \rightarrow l_c \\ \text{(headphones, airpods) + verbaliser(irrelevant, irrelevant)} \rightarrow l_i \end{cases}$$

The final predicted label is chosen by first converting the output logits to probabilities (so that they add upto 1) using `softmax()` and then taking an `argmax()` on the four probabilities (p_e, p_s, p_c, p_i).

7 IMPLEMENTATION DETAILS

We use the Multilingual BERT-base[1] model as our backbone model for all experiments. It has total of 168M trainable parameters, 12 heads, 12 layers and 768 hidden dimension. On top of that we have a single linear layer for final classification. We use the AdamW optimizer with linear learning rate decay post 10% warmup steps for finetuning. We use the standard cross-entropy loss for finetuning purposes.

For RetailQueries dataset, we train all models for 5 epochs using learning rate 1e-5, max sequence length 128 & batch size 16. It is a two-sentence task since we have both query & product information to be classified into E/S/C/I. The input to the model (also definition of *concat()* operation) is in the format - $query$ [SEP] $verbaliser(label_{pre}, label_{post})$ [SEP] $product$. As mentioned in Section 6, we have 1 positive example and (K-1) negative examples per datapoint. So to reduce the label imbalance, we augment one more positive sample created by random deletion of 5% text span in product title. For AGNews dataset, we train for 10 epochs using learning rate 1e-6, max sequence length 128 & batch size 16. It's a single-sentence task of classifying news articles into relevant/irrelevant news. So the input format is - $verbaliser(label_{pre}, label_{post})$ [SEP] $news$. Here choosing the position where verbaliser is appended is a design choice. We observe best results in the above reported settings.

We ran all experiments for 5 different seeds on 4 Nvidia V100 GPUs parallely on a 70-10-20 train-val-test split and report the mean and standard deviations. Training takes around ~30 mins on RetailQueries & ~1 hr on AGNews datasets to reproduce the reported results.

8 RESULTS & DISCUSSION

We compare our approach against these baselines to show the effectiveness of our proposed entailment-style approach.

Majority - Assign the majority class label to all datapoints.

Finetuned-pre - We had historic e-commerce relevance data in the order of millions annotated into E/S/C/I classes before the guideline changes occured (pre-shift). This was the older data on which production models were finetuned. We finetune $\mathcal{M}_{base}$ on this pre-shift data corpus and evaluate directly on post-shift data. This baseline demonstrates the need for models adapting to such concept shifts to avoid drop in performance. We refer to this model as $\mathcal{M}_{pre}$

Finetuned-post - We finetune $\mathcal{M}_{base}$ directly on post-shift data available.

Finetuned [35] - Similar to the sliding window approach in [35] we adopt the pre-shift finetuned model obtained in the above *Finetuned-pre* baseline. Then we perform continual finetuning on post-shift data. In [35] the authors use a domain impact score to decide if and when the concept drift has occurred to trigger re-finetuning. But we omit this in our case since we are dealing with sudden shift and we have clear distinction between pre-shift and post-shift data. We refer to this model as $\mathcal{M}_{post}$

Vanilla-entailment [43] - To simulate this baseline, we initialize the M_{base} with M_{pre} and remove the concept shift information from the verbalizers and simply present it with the accurate label text - {'$exact$', '$substitute$', '$complement$', '$irrelevant$'} for RetailQueries Shift and {'$relevant$', '$irrelevant$'} for AGNews Shift datasets. We repeat the same for the multilingual variant as well. Although this aspect wasn't explored in the original baseline, we still run it to conduct a comprehensive analysis.

[1] https://huggingface.co/bert-base-multilingual-uncased

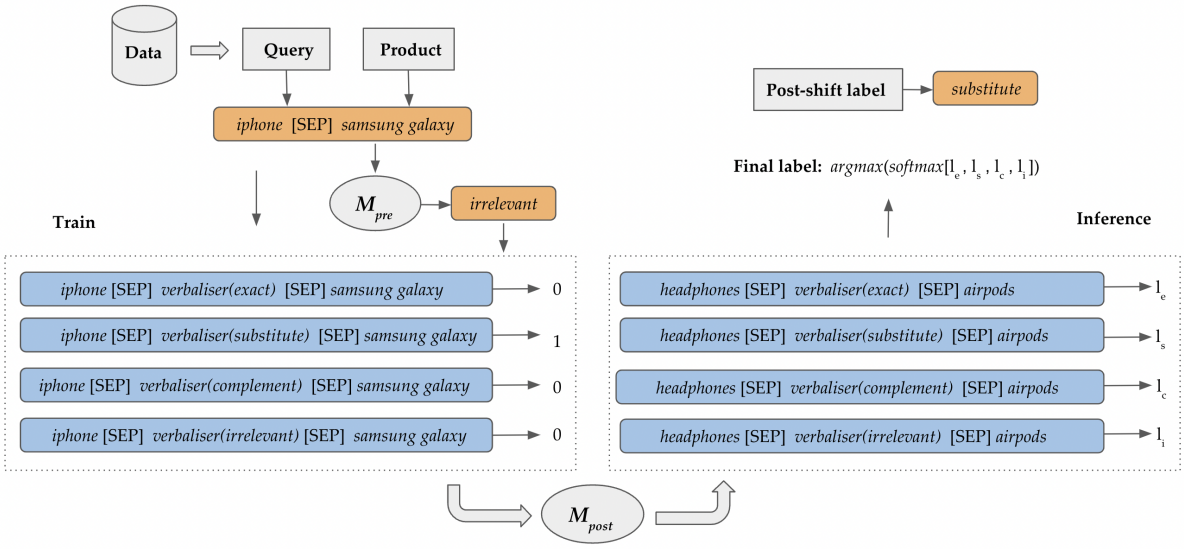


Figure 4: [Best viewed in color] Proposed *Entail-style* approach. Here we assume pre-shift label as *irrelevant* coming from $\mathcal{M}_{pre}$. So *verbaliser(exact)* ≡ *verbaliser(irrelevant, exact)*. For lack of space we only mention the post-shift label (L_j) in the argument to *verbaliser()*.

Model	Macro Avg. F1-score			
	Full Data	N' = 1000	N' = 100	N' = 10
RetailQueries Shift Data (Real)				
Majority	10.0(0.0)	10.0(0.0)	10.0(0.0)	10.0(0.0)
Finetuned-pre	35.48(0.49)	35.48(0.49)	**35.48(0.49)**	**35.48(0.49)**
Finetuned-post	37.98 (0.11)	17.35(0.6)	16.92(0.88)	14.61(1.5)
Finetuned [35]	41.93(0.21)	35.74(0.55)	27.19(0.79)	24.53(1.9)
Vanilla-entailment (english) [43]	39.58(0.26)	33.87(0.28)	27.74(0.31)	23.58(1.18)
Vanilla-entailment (multilingual)	40.28(0.41)	34.11(0.4)	28.12(0.80)	24.98(1.24)
Entail-style (english)	42.37(0.08)	36.69(0.10)	31.19(0.28)	30.80(1.37)
Entail-style (multilingual)	**43.96 (0.33)**	**36.87 (0.37)**	34.65 (0.34)	31.19 (1.44)
AGNews Shift Data (Synthetic)				
Majority	33.33(0.0)	33.33(0.0)	33.33(0.0)	33.33(0.0)
Finetuned-pre	22.89(0.59)	22.89(0.59)	22.89(0.59)	22.89(0.59)
Finetuned-post	91.55(0.23)	86.27(0.5)	38.6(0.66)	18.19(3.89)
Finetuned [35]	93.47(0.09)	88.31(0.78)	42.7(0.09)	24.35(3.50)
Vanilla-entailment [43]	93.70(0.20)	87.89(0.97)	42.94(0.23)	42.8(0.01)
Entail-style (english)	**96.48(0.13)**	**93.5(0.10)**	**83.61(0.91)**	**73.14(3.22)**

Table 2: Results on RetailQueries & AGNews Shift datasets for various few-shot settings (N'=10, 100, 1000, full data). Best runs are marked in bold and second best is underlined. We report the average and standard deviation of 5 runs (reported results are significantly different: M-W-U test with p-value < 0.05).

Our Approach (Entail-style) - We again take the pre-shift fintuned model obtained from the *Finetuned-pre* approach. Then for RetailQueries Shift data, we continually finetune using two different entailment label verbaliser variants based on equation (2) - (a) We create the augmented dataset by adding verbaliser(L_j,L_k) in English as shown in Table 3. (b) Given that the dataset is multilingual, instead of using English-based label verbaliser alone for all the input data, we add multilingual label verbaliser[2] corresponding to the language of the text. For AGNews Shift data, we employ only the first monolingual variant we mentioned above since the dataset contains English-only datapoints.

Note that for AGNews Shift data, we have the ground-truth pre-shift and post-shift labels for the same datapoint available since we synthetically create the concept shift. However, in the case of RetailQueries Shift data, the 62.5k data points are only annotated with post-shift labels. To compensate for the absence of pre-shift labels, we rely on the predictions of the *Finetuned-pre* (M_{pre}) model, treating them as pseudo-labels. This model is trained on our extensive historical relevance E/S/C/I data (in the order of millions) annotated under the old guidelines prior to the concept shift. This

[2]Google Translate API

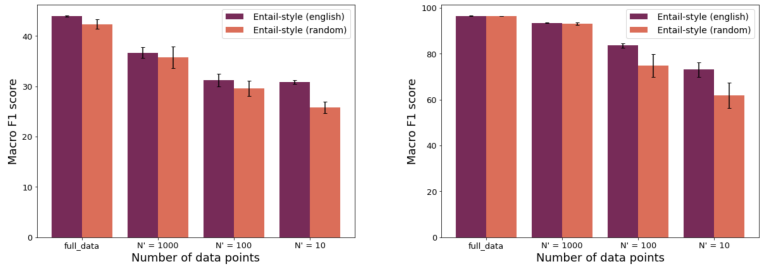


Figure 5: [Best viewed in color] Random vs Informative Prompts. The macro-F1 scores reported are average of 5 runs. The bars indicate standard deviation (reported results are significantly different: M-W-U test with p-value < 0.05). *(Left)* RetailQueries Shift Dataset. *(Right)* AGNews Shift Dataset.

Language	Label Shift	Verbaliser
RetailQueries Shift Data (Real)		
English	Exact → Substitute	changed from exact to substitute match
	Exact → Complement	changed from exact to complement match
	Exact → Irrelevant	changed from exact to irrelevant match
	Exact → Exact	remained exact match
Spanish	Exact → Substitute	cambiado de coincidencia exacta a sustituta
	Exact → Complement	cambiado de coincidencia exacta a coincidencia de complemento
	Exact → Irrelevant	cambiado de coincidencia exacta a irrelevante
	Exact → Exact	permaneció coincidencia exacta
AGNews Shift Data (Synthetic)		
English	Irrelevant → Relevant	changed to relevant news
	Irrelevant → Irrelevant	remained irrrelvant news
	Relevant → Irrelevant	changed to irrelevant news
	Relevant → Relevant	remained relevant news

Table 3: We provide the exact verbalisers used for our experiments following the template mentioned in § 6. For the RetailQueries Shift dataset (multilingual), the verbalisers for English and Spanish languages are given above for the case when pre-shift label is "exact"; similar verbalisers for other languages have been obtained using Google Translate. For the AGNews Shift dataset (English), we similarly provide the verbalisers used in our experiments.

is also reflective of the post-deployment setting where our models won't have access to ground-truth pre-shift labels for any new incoming datapoint.

Next to interpret our obtained results, we frame the following research questions and answer them below:

Q1: Does *Entail-style* model outperform other baselines?

As evident from Table 2, *Entail-style (english)* mostly outperforms all finetuned baselines in both RetailQueries and AGNews Shift datasets. The gains over *Finetuned* [35] and *Vanilla-entailment* [43] (∼**1-6%** in RetailQueries and ∼ **3-30%** in AGNews) indicates the benefit of entailment-style reformulation of the task where the model is able to leverage pre-shift label information along with post-shift label information better (given the textual label descriptions/verbalisers) to improve over the baselines. Only for N'=10, 100 cases in RetailQueries Shift data, we see *Finetuned-pre* to have a better performance. However it is not directly comparable as this is attributed to the extensive finetuning of the pre-shift model with millions of historical datapoints, in contrast to our *Entail-style* approach, which utilizes only 10 and 100 datapoints. Nevertheless, it is noteworthy that *Entail-style* surpasses both *Finetuned-post*, *Finetuned* [35] and *Vanilla-entailment* [43] (directly comparable baselines finetuned with same number of datapoints) by significant margins, narrowing the performance gap with the pre-shift finetuned model. These findings also align with the results from [43], where the authors demonstrate that vanilla entailment outperforms classification only for partially-seen label cases and not consistently for fully seen label cases

Also it can be observed that the magnitude of gains are much higher in AGNews Shift data compared to RetailQueries Shift, especially in low-data setting. This can be attributed to the fact that in AGNews Shift the concept shift is induced synthetically which is rather simple to learn compared to RetailQueries Shift data which is a real-world dataset presenting a more nuanced and subjective definition of concept shift.

Q2: Is there any new finding from few-shot experiments?

Again referring to Table 2, for $N' \in \{1000, 100, 10\}$ the gains of *Entail-style (english)* approach is even higher compared to other finetuned baselines. Moreover, this improvement continues to rise as the value of N' decreases. This trend underscores the efficacy of our approach in low-data settings. Also note that for lower values of N' the performance of *Finetuned* [35] decreases rapidly whereas our proposed approach has a more stable performance across all

data settings, making it a more reliable choice for deployment in real-world scenarios.

Q3: Does adding multilingual verbalisers help over monolingual?

We meticulously curate the RetailQueries Shift dataset consisting of 12 distinct languages (see Section 4). Hence, we also conduct an experiment where we use multilingual label verbalisers by translating the english verbaliser to its specific language counterpart based on the (query, product) pair's language. This led to further ∼**1-3%** gains over *Entail-style (english)* making it the best performing model on RetailQueries dataset. This aligns with our expectations as multilingual verbalisers likely facilitate the model to capture semantics better for datapoints which are also non-english by ensuring language consistency as also evident from Figure 6 and *Vanilla-entailment* results in Table 2.

Q4: What happens if we replace the informative label verbalisers with random words?

To tease out the contribution of informative label verbalisers, we add random verbalisers by swapping *E/S/C/I* with *cat/lion/zebra/dog* in the template mentioned in Table 3. With random verbaliser, we observe that for lower values of N' the performance gap is high with the english and multilingual verbaliser variants but as we increase N' the performance starts becoming comparable with its informative counterpart. This is due to the model learning some kind of spurious mapping between the random & actual labels with sufficient data. This finding is also in-line with [40]. However, it should be noted that using random verbalisers resulted in a **higher standard deviation** in all the settings as evident in Figure 5 making such an approach unreliable for use in real-world systems.

9 BUSINESS IMPACT

We report changes in metrics using **bps** (or basis points). It is a standard and significant unit to measure changes in metrics in e-commerce industry (1 bps = 0.01%).

Labeling and Downstream Cost Savings - On post concept shift data, the pre-shift deployed model performance (measured by macro-averaged F1-score) deteriorated by **809 bps** compared to the earlier benchmark production score. However, our final *Entail-style (multilingual)* approach (see Table 4) outperforms the previous production model performance by **160 bps** using just **25%** of pre-shift query-product pairs labeled again by humans. This directly saved human annotation costs for **75% of the labels**, which is in the order of millions of USD. Also finetuning the *Finetuned* [35] approach on 25% data recovered only 400 bps of the 809 bps drop observed in the pre-shift model performance thus making our approach the only suitable choice for deployment. This also shows the effectiveness of *Entail-style* method in low-data settings. Also note that the model we used in our production setting is a much larger encoder model (compared to mBERT) trained on proprietary amazon relevance data. This leads to another interesting finding that our observations with *Entail-style* approach holds true for **larger models** as well.

We also use our relevance models to automate labeling workflows. For any incoming batch of data, we run the production relevance model first. If the model prediction confidence is above a threshold (@ precision > 90), we assume that to be ground-truth and do not send these datapoints further for human labeling resulting in further cost savings. This model has been in production already for 3 months (at the time of writing this paper) which is also approximately the amount of time which would've been required to get the rest 75% data labeled. During this 3-month period, tuning the production systems to the shifted concepts allowed us to avoid model down-time and save **40% indirect** labeling costs (again in the order of millions USD).

An inherent trade-off associated with the entailment-based approach is that improvements in performance are offset by elevated latency and inference expenses, approximately $O(K)$ times greater than those of the conventional approach. We serialize the models into `torchscript` format before deployment, the cost of running inference on 1 million datapoints is still < $10 on a `p3.2xlarge` EC2 instance (roughly an order of 1000x lower than the cost of acquiring the equivalent volume of human annotated labels for conventional finetuning) thus rendering the overhead inference costs negligible. Also latency increase isn't a concern here, as we run these relevance models offline at a regular cadence and populate an AWS GLUE table using output probabilities for (query, product) pairs which is subsequently utilized by online models.

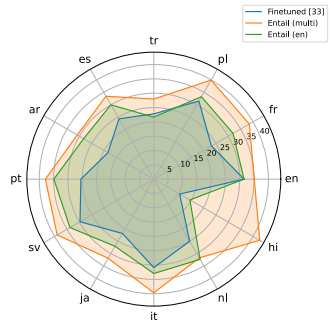


Figure 6: Language-wise macro-F1 difference between *Entail-style* approach and the best performing baseline on *RetailQueries Shift* dataset.

Model	Prod	Finetuned-pre	Finetuned [35]	Entail (multi)
F1	-	-809bps	-400bps	+160bps

Table 4: Relative macro-f1 difference compared to benchmark production number evaluated using 700k snapshot data from online traffic.

Expansion to newer marketplaces - Our trained models exhibit excellent few-shot learning capabilities, making them ideal for expanding into new or emerging marketplaces where training data is scarce cutting down design-to-deployment times. Also in newer markets customer buying trends vary rapidly thus causing frequent concept shifts in relevance guidelines. The ability to quickly adapt to changing guidelines with less data makes our method well-suited for such use cases. We can also see evidence in Figure 6 where the gains for low-resource languages in RetailQueries Shift data (Figure 2) like Swedish (sv) and Hindi (hi) are much higher compared to the average gain across all languages.

10 CONCLUSION & FUTURE WORK

In this study, we demonstrate that redefining the conventional classification task into a natural language inference (NLI) or entailment-style task can enhance the adaptability of pre-trained language models (PLMs) in the face of sudden concept shifts. We curate the first of its kind real-world, multilingual benchmark dataset for studying concept shift in textual classification. We report results both on real-world and synthetic datasets. Additionally, we present evidence supporting the superior effectiveness of our proposed approach in few-shot scenarios. To this extent, we also provide ablations to explore the impact of multilingual vs monolingual verbalisers on performance, as well as the influence of informative vs non-informative verbalisers. This comprehensive analysis provides a nuanced understanding of the strengths and implications of our approach.

Upon deploying our *Entail-style (multilingual)* approach for adapting our internal proprietary models to the sudden concept shift in relevance guidelines, we were able to beat our benchmark production performance by 160 bps using just 25% of available pre-shift datapoints. Thus in a span of just 3 months, we were able to save **75% direct** labeling costs + **40% downstream** labeling costs.

In future work, we would like to study how PLMs utilize label verbalisers to enhance performance on downstream tasks. [40] recently showed that these prompt-based learning methods don't necessarily work the way we think i.e. through leveraging textual semantics of the labels. Other areas to explore would be to learn automated verbalisers [33] for varied downstream tasks rather than manually designing them every time, test the adaptability of our approach to other types of concept shifts (as discussed in § 2) and also explore using Large Language Models (LLMs) instruction-following capabilities to tackle different distribution shifts [12].

Further we narrow the scope of our experiments to focus on the scenario of a 'single' sudden concept shift, reflecting the actual circumstance in our real-world industry setup. However, even in cases where there are multiple changes over a deployment period, our approach can be readily extended in two ways - **(a)** The simpler approach involves adopting a sliding-window-based methodology. This entails considering two sets of timestamps at a time, treating one as the pre-shift label and the other as the post-shift label, and then applying our approach directly. This re-finetuning process repeats until all timestamps have been sequentially covered in pairs. **(b)** Alternatively, we can extend the label verbalizer to accommodate descriptions of multiple label changes at once. For instance, in Table 3, if the label changes from 'exact' → 'substitute', we describe it as 'changed from exact to substitute match.' If there are multiple changes, such as 'exact' → 'substitute' → 'irrelevant', we could describe it as 'changed from exact to substitute to finally irrelevant match.' Similarly, for changes like 'exact' → 'substitute' → 'substitute', we can describe it as 'changed from exact to substitute and remained substitute match.' These are intriguing avenues for exploration in future works.

REFERENCES

[1] Robert Anderson, Yun Sing Koh, and Gillian Dobbie. 2016. CPF: Concept profiling framework for recurring drifts in data streams. In *AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016, Proceedings 29*. Springer, 203–214.

[2] Johannes Bjerva, Wouter Kouw, and Isabelle Augenstein. 2020. Back to the future–temporal adaptation of text representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7440–7447.

[3] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[6] Sarah Jane Delany, Pádraig Cunningham, Alexey Tsymbal, and Lorcan Coyle. 2005. A case-based technique for tracking concept drift in spam filtering. In *Applications and Innovations in Intelligent Systems XII: Proceedings of AI-2004, the Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 3–16.

[7] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[8] Francisco Ferrer-Troyano, Jesus S Aguilar-Ruiz, and Jose C Riquelme. 2005. Incremental rule learning based on example nearness from numerical data streams. In *Proceedings of the 2005 ACM Symposium on Applied computing*. 568–572.

[9] Cristiano Mesquita Garcia, Ramon Simoes Abilio, Alessandro Lameiras Koerich, Alceu de Souza Britto Jr, and Jean Paul Barddal. 2023. Concept Drift Adaptation in Text Stream Mining Settings: A Comprehensive Review. *arXiv preprint arXiv:2312.02901* (2023).

[10] Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. 2017. A survey on ensemble learning for data stream classification. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–36.

[11] Paulo Mauricio Gonçalves Jr and Roberto Souto Maior De Barros. 2013. RCD: A recurring concept drift framework. *Pattern Recognition Letters* 34, 9 (2013), 1018–1025.

[12] Karan Gupta, Sumegh Roychowdhury, Siva Rajesh Kasa, Santhosh Kumar Kasa, Anish Bhanushali, Nikhil Pattisapu, and Prasanna Srinivasa Murthy. 2023. How Robust are LLMs to In-Context Majority Label Bias? *arXiv preprint arXiv:2312.16549* (2023).

[13] Geoff Hulten, Laurie Spencer, and Pedro Domingos. 2001. Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 97–106.

[14] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 2763–2775. https://doi.org/10.18653/v1/2022.acl-long.197

[15] Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal Direction of Data Collection Matters: Implications of Causal and Anticausal Learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9499–9513.

[16] Ralf Klinkenberg. 2004. Learning drifting concepts: Example selection vs. example weighting. *Intelligent data analysis* 8, 3 (2004), 281–300.

[17] Ralf Klinkenberg and Stefan Rüping. 2002. Concept drift and the importance of examples. In *Text mining–theoretical aspects and applications*. Citeseer.

[18] J Zico Kolter and Marcus A Maloof. 2007. Dynamic weighted majority: An ensemble method for drifting concepts. *The Journal of Machine Learning Research* 8 (2007), 2755–2790.

[19] Carsten Lanquillon and Ingrid Renz. 1999. Adaptive information filtering: Detecting changes in text streams. In *Proceedings of the eighth international conference on information and knowledge management*. 538–544.

[20] Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2627–2636. https://doi.org/10.18653/v1/2021.naacl-main.208

[21] Patrick Lindstrom, Sarah Jane Delany, and Brian Mac Namee. 2010. Handling concept drift in a text data stream constrained by high labelling cost. In *Twenty-Third International FLAIRS Conference*.
[22] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. 2018. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*. PMLR, 3122–3130.
[23] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. 2018. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering* 31, 12 (2018), 2346–2363.
[24] Jacob Montiel, Rory Mitchell, Eibe Frank, Bernhard Pfahringer, Talel Abdessalem, and Albert Bifet. 2020. Adaptive xgboost for evolving data streams. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
[25] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition* 45, 1 (2012), 521–530.
[26] Martin Müller and Marcel Salathé. 2020. Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic. *arXiv preprint arXiv:2012.02197* (2020).
[27] Anand M Narasimhamurthy and Ludmila I Kuncheva. 2007. A framework for generating data to simulate changing environments.. In *Artificial intelligence and applications*. 415–420.
[28] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250
[29] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).
[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]
[31] Sumegh Roychowdhury and Vikram Gupta. 2023. Data-Efficient Methods For Improving Hate Speech Detection. In *Findings of the Association for Computational Linguistics: EACL 2023*. 125–132.
[32] Jeffrey C Schlimmer and Richard H Granger. 1986. Incremental learning from noisy data. *Machine learning* 1 (1986), 317–354.
[33] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4222–4235. https://doi.org/10.18653/v1/2020.emnlp-main.346
[34] Andrés L Suárez-Cetrulo, David Quintana, and Alejandro Cervantes. 2022. A survey on machine learning for recurring concept drifting data streams. *Expert Systems with Applications* (2022), 118934.
[35] E Susi and AP Shanthi. 2023. Transformer based Twitter Trending Topics Sentiment Drift Analysis in Real Time. In *2023 12th International Conference on Advanced Computing (ICoAC)*. IEEE, 1–7.
[36] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).
[37] Alexey Tsymbal. 2004. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin* 106, 2 (2004), 58.
[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[39] Bruno Veloso, João Gama, Benedita Malheiro, and João Vinagre. 2021. Hyperparameter self-tuning for data streams. *Information Fusion* 76 (2021), 75–86.
[40] Albert Webson and Ellie Pavlick. 2022. Do Prompt-Based Models Really Understand the Meaning of their Prompts? arXiv:2109.01247 [cs.CL]
[41] Gerhard Widmer and Miroslav Kubat. 1993. Effective learning in dynamic environments by explicit context tracking. *Lecture Notes in Computer Science* (1993), 227–227.
[42] Yiming Xu and Diego Klabjan. 2021. Concept drift and covariate shift detection ensemble with lagged labels. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 1504–1513.
[43] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161* (2019).
[44] Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Differentiable Prompt Makes Pre-trained Language Models Better Few-shot Learners. https://doi.org/10.48550/ARXIV.2108.13161
[45] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. https://doi.org/10.48550/ARXIV.1509.01626
[46] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. https://doi.org/10.48550/ARXIV.2102.09690

A LIMITATIONS

While entailment-based methods outperform direct supervised baselines by a large margin it comes at an expense of increased inference time. However, given the prohibitively large cost of getting reliable human annotations for new data, and the significant gains observed in low-data settings it seems like a sensible trade-off. Moreover with carefully designed verbalisers, even smaller PLMs could potentially achieve similar performance with lesser data [14] thus reducing inference time. Also, reducing the model parameter precision could further speed up inference (maybe with some performance trade-off) on specialized hardware. We leave these discussions to be explored in future work.

B ETHICAL CONSIDERATIONS

We report only aggregated results in the main paper. We have not or do not intend to share any Personally Identifiable Information (PII) in our released dataset or in the paper. We use standard Huggingface[3] libraries for training our models to promote reproducible research. But it must be kept in mind that these models are not to be used for generation purposes (like GPT [4]). Using biased prompts might lead the model to generate biased responses given these large language models are pre-trained using publicly available data which is why we do not intend to release the trained model weights.

[3]https://huggingface.co/