Tackling Concept Shift in Text Classification using Entailment-style modeling

Sumegh Roychowdhury*
sumegr@amazon.com

Karan Gupta*
karaniis@amazon.com

Siva Rajesh Kasa*
kasasiva@amazon.com

Prasanna Srinivasa Murthy
sprsn@amazon.com

Alok Chandra
alokchan@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 $\sim$**7%** and $\sim$**40%** respectively in few-shot settings. Further, upon deployment, our solution also helped save **75%** of labeling costs overall.

1 Introduction

Concept shift (aka label 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 been studied extensively in classical machine learning problems (8; 7; 9; 10). Concept shift in text classification can occur due to various reasons such as sudden change in external circumstances (e.g. occurrence of pandemics (13)), gradual change in semantics (e.g. words taking on newer meanings (1)), 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 an Entailment-style approach (17) 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

*Equal Contribution.

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

tackling concept-shift (10; 14), 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. Refer to Appendix A.2 for more details.

Our contributions are as follows: (a) We show that ***reformulating vanilla classification as an entailment-style task*** can help augment PLMs performance in tackling concept shift (b) 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. We defer further details motivating our work to Appendix A.2 due to space constraints.

RetailQueries Shift Data (Real)	
Query/Product	**Pre-shift → Post-shift**
vacation things / Bug Bite Suction Tool	Irrelevant → Exact (utility same)
peppa pig muddy puddle / plush toy peppa pig	Irrelevant → Substitute (brand same)
dove soap / dove hand wash nourishing, 3x500ml	Irrelevant → Complement (brand same + utility related)
small globe / Clamp desk lamp	Irrelevant → Irrelevant (no match)
AGNews Shift Data (Synthetic)	
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.

2 Dataset Details

RetailQueries Shift - For creating this data[2], we randomly sample 200k query-product pairs across 12 languages from an e-commerce store retail catalog. For each query, there are a number of products retrieved. 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 and **irrelevant (I)** - does not satisfy query intent. For example, given query "iphone" the product "iphone 11" would be exact, "samsung galaxy" would be substitute, "airpods pro" would be complement and "lord of the rings" would be irrelevant.

Now we describe the type of concept shift that occurs in the dataset. Annotators receive crisp guidelines as to what relevance labels (E/S/C/I) should be assigned to query-product pairs. Based on observed customer behavior, a change in the labeling guidelines was introduced, as to what is to be considered "irrelevant" to better match the customer intent. So **pre-shift** all labels in our 200k dataset were **"irrelevant"**. Some labels changed post-shift due to guideline change while some remained same. The major changes were - **(a)** the product type intended by the query must be matching with the retrieved product for it to be relevant (E/S/C) 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. **(b)** Under new guideline, certain constraints on product attributes are relaxed. For e.g. attributes viz. Size, Age, etc. are relaxed and products at one level higher or lower will be considered substitute, while under old guidelines, they would be considered irrelevant. **(c)** 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.

AGNews Shift - We use the AGNews Topic Classification (25) dataset where each news article is divided into 4 topics - World, Business, Sports and Sci-Tech and 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

[2]Since its proprietary data, we cannot make it public without proper licensing. We will work on releasing the datasets in future iterations.

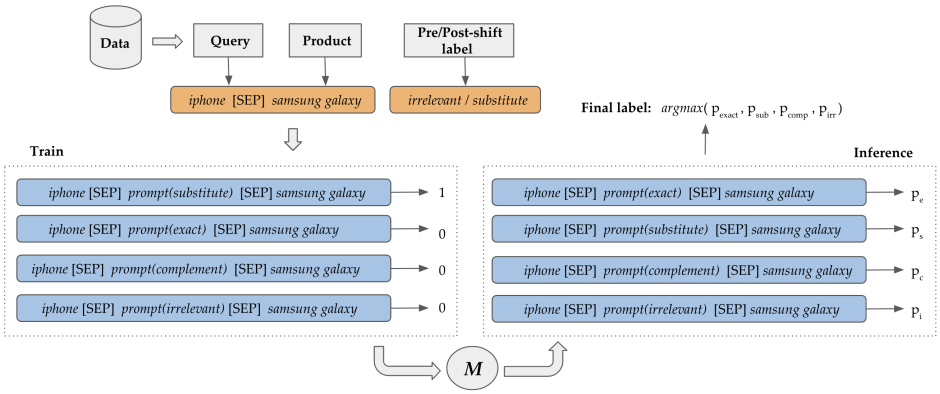


Figure 1: [Best viewed in color] **Proposed Model**. Consider the (query, product) pair *(iphone, samsung galaxy)* to be earlier "irrelevant" and now "substitute" in the concept-shifted training data. Then we create 3 -ve examples *(iphone + samsung galaxy, changed to exact, 0)*, *(iphone + samsung galaxy, changed to complement, 0)* , *(iphone + samsung galaxy, remained irrelevant, 0)* and 1 +ve example *(iphone + samsung galaxy, changed to substitute, 1)*. Once the model $\mathcal{M}$ is fine-tuned on the concept-shifted training data, during inference, say we encounter the (query, product) pair *(earphones, airpods pro)*. We compute the argmax of probabilities of the following augmented inputs: *(earphones + airpods pro, exact/substitute/complement/irrelevant)* and pick the predicted label accordingly.

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.

For further details regarding dataset statistics refer to Table 1 and Appendix A.1.

3 Proposed Approach

Let the $\mathcal{L} = \{L_1, \ldots, L_K\}$ be the set of labels. The training dataset can be segregated based on the labels as $\mathcal{D}_{tr} = \{D_1, D_2, \ldots, D_K\}$ where $\{x_i^k\}_{i=1}^{n_k}$ is the set of training data available for the k-th label L_k. Let the concept shifted data be $\tilde{\mathcal{D}}_{tr} = \{\tilde{D}_1, \tilde{D}_2, \ldots, \tilde{D}_K\}$ where $\{\tilde{x}_i^k\}_{i=1}^{n'_k}$ is the set of training data available for the k-th label L_k after the concept shift. Let $\mathcal{M}$ be the finetuned model on the data $\mathcal{D}_{tr}$. For the concept-shifted data $\mathcal{D}'_{tr}$, during training we do the following entailment-style data augmentation:

For a given datapoint x whose pre-shift label is L_j and post-shift label is $L_{j'}$, we create a total of K augmented samples (s_k) with binary labels as follows:

$$s_k = \{prompt(L_k) + x,\ \mathbb{1}_{j'}(k)\}\ \forall\ k \in \{1, 2, \ldots, K\} \tag{1}$$

where $\mathbb{1}_{j'}()$ is the indicator function which takes the value one only when the argument is j', otherwise it is zero. '+' here refers to concat() operation (check § A.3). Also $prompt()$ is defined as :

$$prompt(L_k) = \begin{cases} remained\ \{L_k\} match & , L_k = L_j \\ changed\ to\ \{L_k\} match & , o.w. \end{cases} \tag{2}$$

Essentially, we create $(K-1)$ negative samples and 1 positive sample for each datapoint. Finally, the problem reduces to the question - Does x entail $prompt(L_k)$ or not ?

After creating these $K \times \sum_{k=1}^{K} n_k$ augmented examples, we finetune $\mathcal{M}_0$ on a binary classification task. Let $\tilde{\mathcal{M}}$ be the finetuned model. During inference, the predicted label for an datapoint $\tilde{x}_i$ is obtained by taking an argmax over the probabilities output by the binary classifier $\tilde{\mathcal{M}}$ over the augmented samples:

$$\text{prediction}(\tilde{x}_i) = \arg\max_k \{\tilde{\mathcal{M}}(prompt(L_k) + \tilde{x}_i)\} \forall\ i \in \tilde{D}_{test} \tag{3}$$

4 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 - Take the already finetuned model on pre-shifted data and finetune it further on post-shift data.

Finetuned (post-shift only) - Same as above, the difference being we don't finetune the model using pre-shift labels. Instead, we directly finetune using post-shift labels. This clearly indicates the benefit of entailment-style reformulation of the task where the model is able to leverage pre-shift label information alongwith post-shift label information (given the textual label descriptions/prompts) to improve over the *Finetuning* approach. We don't report this for RetailQueries data since all labels are same pre-shift (irrelevant).

Finetuned (L1+L2) - Finetune end-to-end with loss = cross-entropy(prediction, pre-shift label) + cross-entropy(prediction, post-shift label). This doesn't perform competitively in AGNews data which can be attributed to the fact that using both loss signals at the same time confuses the model as to what the ground-truth is. For RetailQueries, since all labels are the same (pre-shift) this boils down to same setting as previous *Finetuned* approach.

Pre-shift Finetuning - To demonstrate the need for models adapting to such concept shifts we add this baseline (in Figure 2) which is fine-tuned only on pre-shift data and evaluated on post-shift data leading to a catastrophic drop in performance.

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	51.27(0.17)	25.67(0.56)	19.58(0.90)	18.53(1.73)
Finetuned (L1+L2)	51.25(0.11)	25.7(0.6)	18.45(0.88)	18.12(1.5)
Entail-style (english)	<u>51.67</u>(0.08)	<u>29.62</u>(1.10)	<u>24.37</u>(1.28)	<u>23.83</u>(0.37)
Entail-style (multilingual)	**52.62(0.33)**	**32.32(0.37)**	**26.36(0.34)**	**24.93(0.44)**
AGNews Shift Data (Synthetic)				
Majority	33.33(0.0)	33.33(0.0)	33.33(0.0)	<u>33.33</u>(0.0)
Finetuned	<u>93.47</u>(0.09)	<u>88.31</u>(0.78)	<u>42.7</u>(0.09)	24.35(0.80)
Finetuned (post-shift only)	91.55(0.23)	86.27(0.5)	38.6(0.66)	18.19(0.89)
Finetuned (L1+L2)	56.97(0.92)	54.73(2.95)	38.43(0.98)	24.42(0.93)
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 <u>underlined</u>. We report the average and standard deviation of 5 runs (reported results are significantly different: M-W-U test with p-value < 0.05).

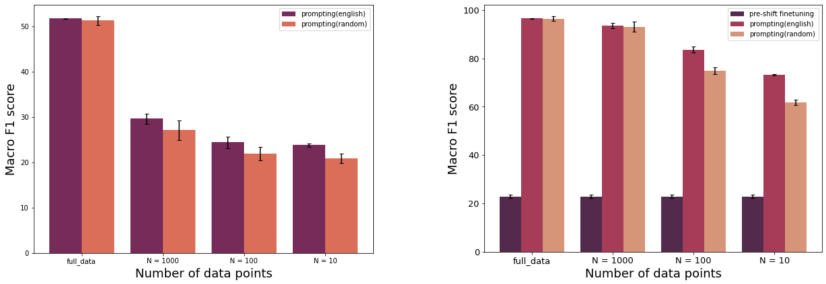


Figure 2: [Best viewed in color] **Random vs Informative Prompts.** We also add results for Pre-Shift finetuning to show the drastic drop in performance compared to the entailment-style approaches (Note: Pre-Shift Finetuning for RetailQueries is not shown since all labels were irrelevant before). The 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 Dataset. *(Right)* AGNews Dataset.

Language	Label Shift	Prompt (Label)
RetailQueries Shift Data (Real)		
English	Exact	changed to exact match
	Substitute	changed to substitute match
	Complement	changed to complement match
	Irrelevant	remained irrelevant match
Spanish	Exact	cambiado a coincidencia exacta
	Substitute	cambiado para sustituir el partido
	Complement	cambiado para complementar la coincidencia
	Irrelevant	permaneció un partido irrelevante
AGNews Shift Data (Synthetic)		
English	Irrelevant $\rightarrow$ Relevant	changed to relevant news
	Irrelevant $\rightarrow$ Irrelevant	remained irrrelvant news
	Relevant $\rightarrow$ Irrelevant	changed to irrelevant news
	Relevant $\rightarrow$ Relevant	remained relevant news

Table 3: For RetailQueries shift dataset (multilingual), we mention only the post-shift labels since pre-shift label is 'irrelevant' for all. We run both plain English-based and Multilingual prompting. The prompts for English and Spanish languages are given above; similar prompts for other languages have been obtained using Google Translate. For AGNews shift dataset (English), the prompts contain information about both pre & post-shift labels.

Our Approach (Entail-style) - For RetailQueries, we finetune the *multilingual-BERT* (5) model using two different entailment prompt variants based on equation (2). In the first variant, we create the augmented dataset by adding prompt(L_j,k) in English as shown in Table 3. This leads to F1 improvements of $\sim$**1-10%** over the finetuned baseline. Next, in the second variant, given that the dataset is multilingual, instead of using English based prompts alone for all the input data, we add multilingual prompts[3] corresponding to the language of the text. This leads to further gains of $\sim$**1-3%** over the monolingual prompts. For AGNews shift data, we repeat the same set of experiments using *multilingual-BERT*. The task-specific prompts used are mentioned in Table 3. We observe gains around $\sim$**3-40%** F1 improvement using our approach. to evaluate the efficacy when labelled data is scarce and expensive, we also run in the proposed methods in the few shot settings corresponding to $N \in \{10, 100, 1000\}$ and observe much higher gains in low-data settings (refer Table 2). In general multilingual prompting has the best overall F1 with low standard deviation. This is expected as multilingual prompts makes it easier for the model to capture context when the datapoints themselves span across multiple languages. English-based prompting has the second best results.

Effect of Random Prompts: To tease out the contribution of informative prompts, we add random prompts by swapping *E/S/C/I* with *cat/lion/zebra/dog* in the prompt template mentioned above. With random prompts we observe that as we increase N the performance starts becoming comparable with it's informative prompt counterpart. This could possibly be due to the model learning some kind of mapping between the random & actual labels with sufficient data. This phenomenon has been previously reported in (22). However, it should be noted that using random prompts resulted in a **higher standard deviation** in all the settings as evident in Figure 2 making such an approach unreliable for use in real-world systems.

5 Conclusion

Upon deploying our final *Entail-style (multilingual)* approach for adapting our internal models to the sudden concept shift, we were able to re-achieve production-level performance using just 25% of pre-shift "irrelevant" query-product pairs (now changed to E/S/C/I post-shift) thus **saving on 75% labeling costs**. In future work, we would like to study how exactly PLMs leverage these prompts to improve their performance on downstream tasks. (22) recently showed that these prompt-based learning methods don't necessarily work the way we humans think i.e. through leveraging textual semantics of the labels. Another area to explore would be starting with manual prompts and then try to learn automated prompts (18) which would eradicate the need to design prompts for varied downstream tasks further boosting the performance.

[3]Google Translate API

References

[1] BJERVA, J., KOUW, W., AND AUGENSTEIN, I. Back to the future–temporal adaptation of text representations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 7440–7447.

[2] BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., AGARWAL, S., HERBERT-VOSS, A., KRUEGER, G., HENIGHAN, T., CHILD, R., RAMESH, A., ZIEGLER, D. M., WU, J., WINTER, C., HESSE, C., CHEN, M., SIGLER, E., LITWIN, M., GRAY, S., CHESS, B., CLARK, J., BERNER, C., MCCANDLISH, S., RADFORD, A., SUTSKEVER, I., AND AMODEI, D. Language models are few-shot learners, 2020.

[3] CHOWDHERY, A., NARANG, S., DEVLIN, J., BOSMA, M., MISHRA, G., ROBERTS, A., BARHAM, P., CHUNG, H. W., SUTTON, C., GEHRMANN, S., ET AL. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[4] DELANY, S. J., CUNNINGHAM, P., TSYMBAL, A., AND COYLE, L. 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* (2005), Springer, pp. 3–16.

[5] DEVLIN, J., CHANG, M., LEE, K., AND TOUTANOVA, K. 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)* (2019), J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, pp. 4171–4186.

[6] JIN, W., CHENG, Y., SHEN, Y., CHEN, W., AND REN, X. 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)* (Dublin, Ireland, May 2022), Association for Computational Linguistics, pp. 2763–2775.

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

[8] KLINKENBERG, R., AND RÜPING, S. Concept drift and the importance of examples. In *Text mining–theoretical aspects and applications* (2002), Citeseer.

[9] KOLTER, J. Z., AND MALOOF, M. A. Dynamic weighted majority: An ensemble method for drifting concepts. *The Journal of Machine Learning Research 8* (2007), 2755–2790.

[10] LINDSTROM, P., DELANY, S. J., AND MAC NAMEE, B. Handling concept drift in a text data stream constrained by high labelling cost. In *Twenty-Third International FLAIRS Conference* (2010).

[11] LU, J., LIU, A., DONG, F., GU, F., GAMA, J., AND ZHANG, G. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering 31*, 12 (2018), 2346–2363.

[12] MORENO-TORRES, J. G., RAEDER, T., ALAIZ-RODRÍGUEZ, R., CHAWLA, N. V., AND HERRERA, F. A unifying view on dataset shift in classification. *Pattern recognition 45*, 1 (2012), 521–530.

[13] MÜLLER, M., AND SALATHÉ, M. Addressing machine learning concept drift reveals declining vaccine sentiment during the covid-19 pandemic. *arXiv preprint arXiv:2012.02197* (2020).

[14] NARASIMHAMURTHY, A. M., AND KUNCHEVA, L. I. A framework for generating data to simulate changing environments. In *Artificial intelligence and applications* (2007), pp. 415–420.

[15] RAE, J. W., BORGEAUD, S., CAI, T., MILLICAN, K., HOFFMANN, J., SONG, F., ASLANIDES, J., HENDERSON, S., RING, R., YOUNG, S., ET AL. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).

[16] RAFFEL, C., SHAZEER, N., ROBERTS, A., LEE, K., NARANG, S., MATENA, M., ZHOU, Y., LI, W., AND LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[17] ROYCHOWDHURY, S., AND GUPTA, V. Data-efficient methods for improving hate speech detection. In *Findings of the Association for Computational Linguistics: EACL 2023* (2023), pp. 125–132.

[18] SHIN, T., RAZEGHI, Y., LOGAN IV, R. L., WALLACE, E., AND SINGH, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online, Nov. 2020), Association for Computational Linguistics, pp. 4222–4235.

[19] SUÁREZ-CETRULO, A. L., QUINTANA, D., AND CERVANTES, A. A survey on machine learning for recurring concept drifting data streams. *Expert Systems with Applications* (2022), 118934.

[20] THOPPILAN, R., DE FREITAS, D., HALL, J., SHAZEER, N., KULSHRESHTHA, A., CHENG, H.-T., JIN, A., BOS, T., BAKER, L., DU, Y., ET AL. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

[21] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[22] WEBSON, A., AND PAVLICK, E. Do prompt-based models really understand the meaning of their prompts?, 2022.

[23] WIDMER, G., AND KUBAT, M. Effective learning in dynamic environments by explicit context tracking. *Lecture Notes in Computer Science* (1993), 227–227.

[24] XU, Y., AND KLABJAN, D. Concept drift and covariate shift detection ensemble with lagged labels. In *2021 IEEE International Conference on Big Data (Big Data)* (2021), IEEE, pp. 1504–1513.

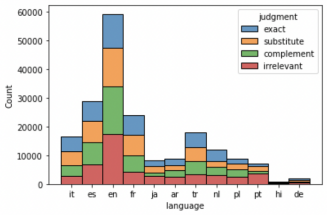
[25] ZHANG, X., ZHAO, J., AND LECUN, Y. Character-level convolutional networks for text classification, 2015.

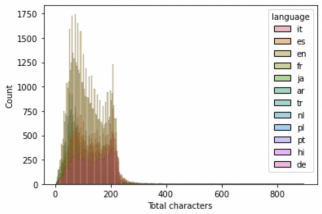
A Appendix

A.1 Dataset Details

A.1.1 RetailQueries Shift Data

Each of these relevance labels E/S/C/I were evaluated manually by human judges. A minimum of 3 annotations were chosen for each pair and majority vote was taken to assign the final gold label. We observe 88% agreement rate pre-shift & 92% agreement post-shift across annotators on average. 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). We release 200k query-product pairs with ∼100k unique queries and ∼180k unique products making it the first of it's kind real-world, large-scale, multi-lingual dataset to study concept shift.





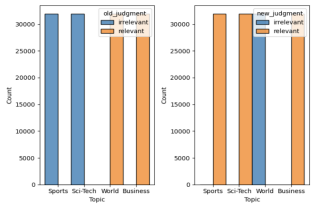
Data description	Value
# Total samples	200k
# Total language(s)	12
# Exacts (old/new)	0/25%
# Substitute (old/new)	0/25%
# Complement (old/new)	0/25%
# Irrelevant (old/new)	100/25%
# Average sample length	116 chars
# Shortest sample	4 chars
# Longest sample	899 chars

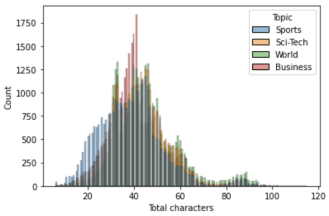
Figure 3: [Best viewed in color] RetailQueries Data *(Left)* Post-Shift label distribution (pre-shift all were irrelevants -Section 2) *(Center)* Character length distribution. *(Right)* Dataset Statistics.

Note that there might be exceptions to rules proposed in Section 2 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 model is able to deal with these edge-cases & how can we further improve it. Since only I's are changing to E/S/C/I's according to the guideline shift mentioned above, the 200k we sample are all I's according to old guidelines. It's sampled in such manner that after concept shift in new guideline 50k each will get converted to E,S,C & I classes making it also a well-balanced dataset. Refer to Figure 3 for details.

A.1.2 AGNews Shift Data

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 4 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 4: [Best viewed in color] AGNews Data *(Left)* Pre & Post-Shift label distribution *(Center)* Character length distribution. *(Right)* Dataset Statistics.

A.2 Background and Related Work

Let $\mathbf{X}$ be the features (aka covariates) and $\mathbf{y}$ be the corresponding labels. 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 inference/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. If $\mathcal{P}_{tr}(\mathbf{X}) \neq \mathcal{P}_{tst}(\mathbf{X})$ and $\mathcal{P}_{tr}(\mathbf{y}|\mathbf{X}) = \mathcal{P}_{tst}(\mathbf{y}|\mathbf{X})$, then there is covariate shift. If $\mathcal{P}_{tr}(\mathbf{X}) = \mathcal{P}_{tst}(\mathbf{X})$ and $\mathcal{P}_{tr}(\mathbf{y}|\mathbf{X}) \neq \mathcal{P}_{tst}(\mathbf{y}|\mathbf{X})$, then there is concept shift. 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 (24). We refer to (12) 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*(10; 14; 19). 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.

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 (11). However, training PLMs requires that there is enough labelled data (21; 5; 16; 2; 20; 15; 3) which are expensive and time consuming to obtain. Previous approaches to tackle sudden changes in concept 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 (8; 4). 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 (23). In the context of PLMs, fine-tuning based on a new batch of data can considered as a sliding-windowing approach. We use this as one of the baselines in our paper. Also as mentioned earlier we limit the scope of this work to only sudden shift (since it's more challenging), we don't compare against other windowing-based methods as those are suited for gradual shift problems where there is some timeline-based information associated with the shifts. For a more detailed review, we refer to the study by (11).

A.3 Implementation Details

We use the Multilingual BERT-base[4] model as our backbone model for all experiments. It has total of 110M 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] $prompt(label)$ [SEP] $product$. As mentioned in Section 3, 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 - $prompt(label)$ [SEP] $news$. Here choosing the position where prompt 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 and report the mean and standard deviations. Training takes around $\sim$5 hrs on RetailQueries & $\sim$1 hr on AGNews datasets to reproduce the reported results.

[4] `https://huggingface.co/bert-base-multilingual-cased`

A.4 Limitations

While prompting-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 (see N=1000 in Table 2) it seems like a sensible trade-off. Moreover with carefully designed prompts, even smaller PLMs could potentially achieve similar performance with lesser data (6) 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.

A.5 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[5] 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 (2)). 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.

[5]https://huggingface.co/