A Benchmark for Text Quantification Learning Under Real-World Temporal Distribution Shift

Anonymous ACL submission

Abstract

Text quantification is a supervised learning task estimating the relative frequency of each class for a collection of uncategorized text documents. Quantification learning has an increasing number of applications in practice and presents unique challenges that are often overlooked in classification problems, such as dealing with distribution shift. Many studies on quantification use artificially re-sampled test sets to evaluate models under varying target label distributions. Despite being a convenient solution, label-based biased sampling changes the underlying test data distribution and makes it hard to rely on the results to deploy models in practice. This paper introduces a text quantification benchmark consisting of 8 datasets across sentiment analysis, document categorization, and toxicity classification. We compare popular quantification baselines on the benchmark and show that there is no model consistently outperforming others. Therefore, we believe the benchmark should enable new community research to tackle text quantification under temporal distribution shift and develop reliable models in real-world applications.

1 Introduction

In the classification setting, quantification is a supervised learning task that estimates the aggregated label distribution of a test population given labeled training examples. A typical application of quantification is to automatically estimate the prevalence of hate speech (Warner and Hirschberg, 2012; Malmasi and Zampieri, 2017; Qian et al., 2018) during a period of time on a social platform. The platform could then use the estimation to determine the effectiveness of a certain feature with A/B testing. Another example in Epidemiology is to track the prevalence of clinical reports where a specific pathology is diagnosed (Stanfill et al., 2010). In both cases, an accurate estimation of the label distribution provides actionable insights.

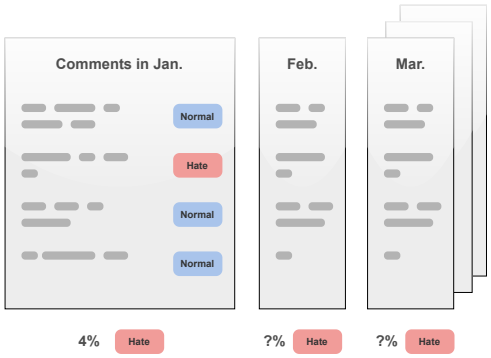


Figure 1: A standard setting for quantification learning. Given annotated training data, a quantifier needs to give prevalence estimates for unlabeled test sets.

Despite the many possible applications, quantification is relatively understudied in the NLP community. One common misunderstanding is that these problems can be solved trivially using a straightforward Classify & Count (CC) approach (Forman, 2008) based on an off-the-shelf classifier. However, classifiers are often trained with the assumption that the training and test examples are drawn i.i.d. from a common data distribution. In contrast, the underlying assumption of quantification learning is that the data distribution changes between the training and the testing phase. Under severe distribution shift, naive aggregation of classification results would yield unsatisfactory performances. As a result, there is a strong connection between quantification learning and tasks that deal with distribution shift.

Label shift (Lipton et al., 2018; Alexandari et al., 2019; Tachet des Combes et al., 2020) is a closely related line of research that detects the shift in the label distribution and adjusts to optimize for classification accuracy on the test examples. When used for quantification, these approaches mostly apply distribution matching in a latent space and are in essence equivalent to earlier quantification methods such as Adjusted Count (Forman, 2008;

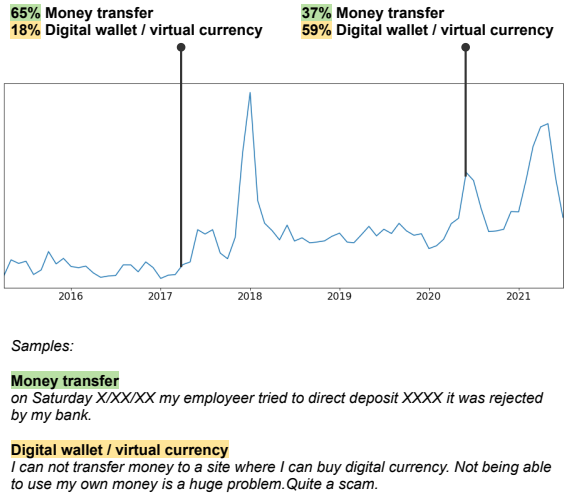


Figure 2: Prevalence change of the "Money transfer, virtual currency" category across time in the Consumer Complaint Database. Within the category, the composition of complaints also changes, creating further challenges to quantifiers.

Hopkins and King, 2010; Saerens et al., 2002) and Probabilistic Adjusted Count (Bella et al., 2010). Compared to quantification learning, label shift literature focuses on performance optimization of the underlying learner where the estimated test label distribution is a by-product. Label shift caused by temporal factors falls into this category but is never explicitly studied.

One main problem with recent studies on quantification is the dataset, especially the testing sets, being used. As pointed out in González et al. (2017), quantification methods need to be evaluated on a set of testing splits with enough variations on the label distribution. Most of the abovementioned studies achieve this by artificially changing the test label distribution through biased sampling. For example, Forman (2008) uses a set of pre-specified positive prevalence values and constructs the test sets accordingly; Lipton et al. (2018) simulates label shift by drawing the test set label distributions from a Dirichlet distribution. However, these stratified sampling strategies change the underlying data distribution and are problematic in assessing the actual performances of quantification models in practice.

In this paper, we introduce the first text quantification benchmark with naturally occurred temporal distribution shift. Each dataset is split into subsets containing samples from the same month or year. The subsets are then grouped into training and testing according to a specified point in time. We construct the training and testing sets to mimic a practical use case in a realistic setting where we need to predict the future class prevalence given historical annotated data. Due to the long time span, the input distribution for each class might change. For example, Figure 2 shows that the "Money transfer, virtual currency" category in one of the datasets has drastically different input composition in early-2017 and mid-2020, which presents more significant challenges to a candidate quantifier.

A total of eight datasets are included in the benchmark spanning sentiment analysis, document categorization, and toxicity classification. We evaluate different quantification/label shift estimation algorithms on the benchmark and find that no algorithm consistently outperforms others.

The main contributions of this work are threefold:

- We create the first benchmark for text quantification learning with temporal distribution shift consisting of diverse tasks and domains to evaluate model performances in a realistic setting.
- We propose a new metric, Class-Averaged Rank Correlation (CARC), for quantification learning that measures models' ability to produce prevalence estimates that are consistent with ground-truth values in terms of ranking order.
- We evaluate various baseline algorithms on the benchmark and find that no algorithm consistently outperforms others, strongly motivating future research in this area.

2 Related Work

Quantification Learning. Many of the experiments reported in quantification learning literature employ datasets taken from other classification problems. For example, Gao and Sebastiani (2016) use 11 sentiment classification datasets and average the performances of studied methods across all 11 datasets. The problem is that only one test set is available for each dataset. Qi et al. (2020) use four text classification datasets for evaluation. However, all four datasets have a balanced training set, and the test splits are artificially created similar to Forman (2008). Beijbom et al. (2015) create two large-scale, image datasets from marine ecology.

The experimental settings are more adequate with 21 and 15 test splits under various distribution shift. Esuli and Sebastiani (2015) employ RCV1-v2, a multi-label text classification benchmark with 52 weeks of data for testing. These datasets lack the number of test splits or the diversity of the task domains. In contrast, our benchmark comprises more diverse tasks and domains; it involves temporal distribution shift across a long period of time; it provides monthly/yearly splits that allow more fine-grained analysis.

Learning under Distribution Shift. There has been an increasing interest in studying the challenges arising from data distribution changes in the machine learning community (Daumé III, 2007; Blitzer et al., 2007; Glorot et al., 2011; Ganin et al., 2016). The focus of these studies is mainly on adapting to covariate shift and improving the performance of the underlying learner in a shifted domain. Lipton et al. (2018) study the problem of adapting a classifier under label shift, assuming the feature distribution of each class stays the same. Tachet des Combes et al. (2020) combine the idea of label shift with adversarial domain adaption and learn invariant representations in different domains.

Quantification learning shares the same distribution shift challenges as domain adaptation and label shift. However, the goal is inherently different. Direct utilization of domain adaptation datasets for quantification is undesirable for two reasons. Firstly, many quantification applications are interested in distribution shift caused by temporal factors because quantifiers are primarily used for prevalence monitoring. The very few datasets that involve temporal distribution shift are mostly vision datasets (Christie et al., 2018; David et al., 2020). Secondly, quantification learning requires test splits to reflect the actual ground truth prevalence for each class which is not a necessity for domain adaptation.

3 The Text Quantification Benchmark

3.1 Problem Formulation

Given a labeled set of examples $D_s = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{l_1, \ldots, l_k\}$, denote $\mathcal{P}(\mathcal{X})$ as the powerset of $\mathcal{X}$, and Δ^d as the standard d-simplex, the task is to induce a quantifier $h : \mathcal{P}(\mathcal{X})\backslash\emptyset \rightarrow \Delta^{k-1}$ from the training data. For a test set $X_t = \{x'_1, \ldots, x'_m\}$, $h(X_t)$ produces a categorical distribution $\hat{\boldsymbol{p}}$ where each element in the predicted vector $\hat{p}_j$ represents the proportion of label l_j in the set of input examples. The goal is to predict $\hat{\boldsymbol{p}}$ that is as close as possible to the ground truth label distribution $\boldsymbol{p}$.

3.2 Dataset Identification and Preparation

There are several considerations when we identify potential dataset candidates for the benchmark:

- The dataset must have instance-level time information to construct test splits based on the time each example was produced.
- There should be no label-based biased sampling in any test split so that the actual underlying label distributions are available to be compared.
- The dataset ideally should span a long period of time so that there is enough label distribution variation in the test splits.
- The benchmark should cover both multi-class and binary classification problems in multiple text domains with various training data sizes.

There are in total eight datasets being included in the benchmark, an overview of the dataset statistics is shown in Table 1.

Amazon Review Data (Ni et al., 2019) contains product reviews collected from Amazon in the range of May 1996 to October 2018. We use the "5-core" subsets for three categories: Clothing Shoes and Jewelry, Electronics, and Office Products to account for different domains and data sizes. There is a steady trend towards a higher percentage of higher review ratings over time, making these datasets suitable for quantification.

For all three categories, reviews made between 2008-08 and 2015-07 are used for training, and reviews from 2015-08 to 2018-07 are split by month and used for testing. The original review ratings are on a scale of five stars. We create binary versions of each category by changing the task to predict the percentage of negative reviews, i.e., reviews with 1 star and 2 stars.

Consumer Complaint Database (CCD) is a collection of complaints about consumer financial products and services that the Consumer Financial Protection Bureau sent to companies for response[1]. All complaints can be classified into nine

[1] https://www.consumerfinance.gov/data-research/consumer-complaints/

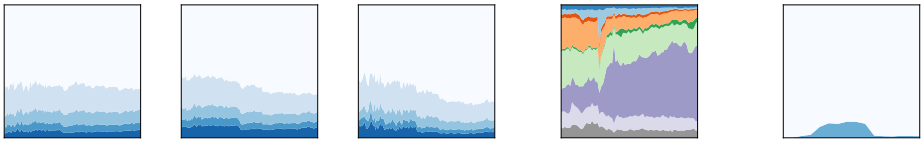
Dataset	Amazon Reviews				
	Clothing	Electronics	Office	CCD	Wikipedia Toxicity
training splits		Aug 2008 - Jul 2015		Apr 2015 - Jul 2019	2001 - 2010
test splits		Aug 2015 - Jul 2018		Aug 2019 - Jul 2021	2011 - 2015
# classes		5 or 2		9	2
# training	3,749,569	3,283,304	316,302	434,482	109,277
# test	7,453,848	3,285,326	476,785	345,914	25,809
label distribution					

Table 1: The text quantification benchmark contains 8 datasets (including three binary versions of the Amazon Reviews datasets) across sentiment analysis, document categorization, and toxicity classification tasks. Each dataset comprises data from a long period of time, and the benchmark is set up to evaluate models' ability to accurately estimate future test split label distributions under naturally occurred distribution shift.

categories based on product types. Due to the evolution of the financial market, the complaint category distribution changes over time. For example, the percentage of credit reporting-related complaints increased from around 16% in 2017-01 to 57% in 2021-07.

We take all the records between 2015-04 and 2021-07 and filter out those without text content. Complaints filed before 2019-07 are used for training, and the remaining data are grouped by month as test splits. There are 434,482 training examples and 345,914 test examples. Test split size ranges from 5,127 to 18,495.

Wikipedia Talk: Toxicity (Wulczyn et al., 2017) includes labeled discussion comments from English Wikipedia. Multiple annotators labeled each comment via Crowdflower on whether it is a toxic or healthy contribution. The original data was collected using two sampling types: *random* and *blocked*. The *random* dataset contains randomly selected comments; therefore, it can be used to evaluate the actual toxicity prevalences over time. The *blocked* dataset is used to ensure a sufficient number of toxic comments for training purposes. We use *blocked* and *random* examples from 2001 to 2010 as the training set, *random* examples from 2011 to 2015 as the test splits. By construction, the underlying distribution changes from training to test significantly. As the up-sampling strategy for imbalanced classification is ubiquitous in practice, this dataset is perfect for evaluating quantification methods with classifiers trained on re-sampled data.

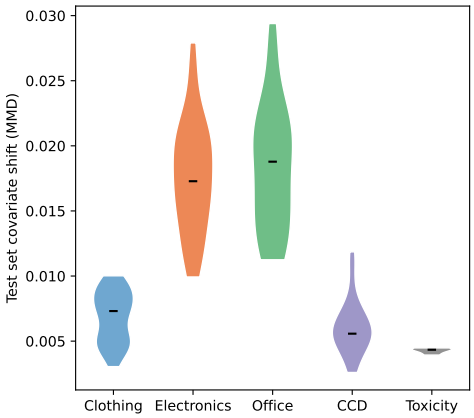


Figure 3: Distributions of the test split covariate shift measured with MMD in the BERT embedding space. Wikipedia Toxicity has milder covariate shift compared to others.

3.3 Distribution Shift Analysis

We measure the distribution shift of each test split compared with the corresponding training set in terms of both covariate shift and label shift.

Covariate Shift. We capture the covariate shift of the input distribution $p(x)$ by encoding the text documents with a pretrained BERT model (Devlin et al., 2019). We then take the BERT embedding of all input examples and evaluate the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012). MMD allows us to compare two probability distributions in a reproducing kernel Hilbert space based on their samples. We use MMD with a radial basis function (RBF) kernel and set σ to be the median distance between points in the training set (Gretton et al., 2012).

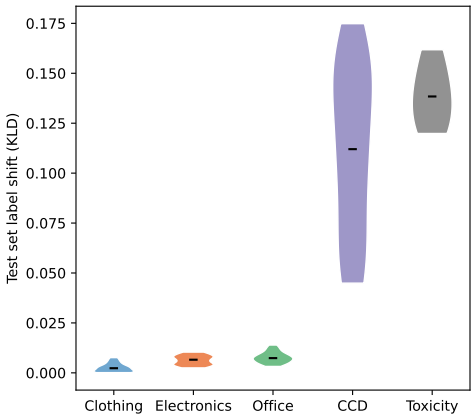


Figure 4: Distributions of the test split label shift measured with KLD. CCD and Wikipedia Toxicity have a higher average label shift as well as larger variations among the test splits.

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}} \tag{1}$$

A larger value measured by MMD indicates a larger discrepancy between the training and testing input embeddings.

We sample 10,000 examples from each test split and measure the MMD for all test splits. The distribution of the test split covariate shifts for each dataset is shown in Figure 3. All datasets have varying levels of covariate shift in their test splits except for Wikipedia Toxicity. All five test splits in the Toxicity dataset have similar levels of MMD values compared to the training set. It is still interesting to see how well models trained with up-sampling estimate the true unbalanced label distribution.

Label Shift. We use Kullback–Leibler divergence (KLD) to measure the difference between the training label distribution and the test split label distributions. Ideally, the label shift of the test splits should cover a range of values to better evaluate candidate methods under various scenarios. The distributions of label shift values in the test splits are plotted in Figure 4. CCD and Wikipedia Toxicity have higher variations in the degrees of label shift from the training set than Amazon Review datasets. There are still reasonable label shift variations in the Amazon Review data as shown in Table 1.

3.4 Evaluation

We use two commonly reported quantification metrics for performance evaluation: Relative Absolute Error (RAE) and Knullback-Leibler Divergence (KLD). We propose a new metric, namely Class-Average Rank Correlation (CARC), to measure the ability to rank test splits by class label prevalences correctly.

Relative Absolute Error. RAE measures the relation between the absolute error and the ground truth label distribution. Formally,

$$\text{RAE}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{k}\sum_{i=1}^{k}\frac{|\hat{\boldsymbol{p}}_i - \boldsymbol{p}_i|}{\boldsymbol{p}_i} \tag{2}$$

Intuitively, RAE measures the average percentage difference from an estimated class prevalence to the ground truth. The lower the better.

Knullback-Leibler Divergence. KLD is a popular metric for measuring the difference between two distributions.

$$\text{KLD}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \sum_{i=1}^{k}\boldsymbol{p}_i \log\frac{\boldsymbol{p}_i}{\hat{\boldsymbol{p}}_i} \tag{3}$$

A benefit of using KLD is that it is widely adopted in the machine learning community and quantification literature. However, it is less interpretable than RAE and can be undefined when $\hat{\boldsymbol{p}}_i = 0$. As RAE and KLD values are closely correlated, reporting both values is redundant in most cases. Therefore, we only report RAE in our experiments.

Class-Averaged Rank Correlation. In addition to RAE and KLD, both of which measure the difference between the predicted label distribution and the ground truth for each test split, we propose a new metric for quantification named Class-Averaged Rank Correlation.

We first uses Spearman's ρ to measure the rank correlation among the predicted prevalence for a particular class across all test splits. CARC is defined as the average rank correlation value across all classes. Formally, let $\boldsymbol{P} = [\boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(t)}]$ denote the list of ground truth label distributions for the t test splits. $\hat{\boldsymbol{P}}$ represents the corresponding list of predictions $[\hat{\boldsymbol{p}}^{(1)}, \ldots, \hat{\boldsymbol{p}}^{(t)}]$. Denote $\boldsymbol{P}_i$ as $[\boldsymbol{p}_i^{(1)}, \ldots, \boldsymbol{p}_i^{(t)}]$, i.e., the list of true prevalences for class i in all test splits. CARC is defined as:

$$\text{CARC}(\boldsymbol{P}, \hat{\boldsymbol{P}}) = \frac{1}{k}\sum_{i=1}^{k}\rho_{R(\boldsymbol{P}_i), R(\hat{\boldsymbol{P}}_i)} \tag{4}$$

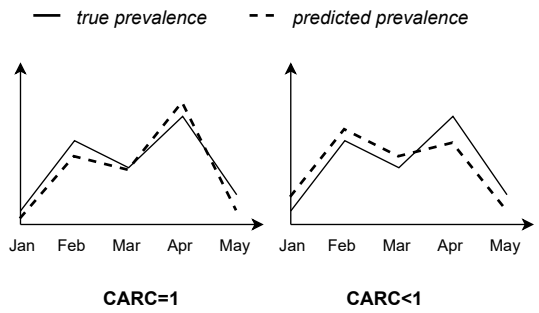


Figure 5: An illustration of the CARC metric. A perfect CARC score of 1 indicates that the model successfully predicts which split has a higher prevalence for any pair of test splits.

where ρ is the correlation coefficient applied to the rank variables $R(\boldsymbol{P}_i)$ and $R(\hat{\boldsymbol{P}}_i)$.

With a higher value of CARC, if test split A has a higher prevalence of a specific class label than test split B, the predicted prevalence in A is more likely to be higher than that in B. CARC is a critical metric because if a quantifier indicates a prevalence increase, the ground truth prevalence should ideally indeed be higher.

4 Baseline Algorithms

In addition to the straightforward Classify & Count (CC) algorithm, we include several methods from prior work on label shift estimation where predictions from a black box classifier can be used as inputs.

Classify & Count (CC) (Forman, 2008). Given classification results from an existing classifier, CC uses the aggregated distribution to predict the test set label distribution. Probabilistic Classify & Count (PCC) is a variant that aggregates the predicted probabilities instead of class assignments.

Black Box Shift Estimation (BBSE) (Lipton et al., 2018). By making a label shift assumption that the conditional distribution of $p(x|y)$ remains the same across training and testing, BBSE uses the confusion matrix to adjust the predicted label distribution from CC. BBSE is proven to be consistent and error bounded even with biased black box predictors as long as the confusion matrix is invertible, and the label shift assumption holds. BBSE, when used for quantification, is equivalent to the Adjusted Count method (Forman, 2008; Hopkins and King, 2010) in multi-class settings.

Regularized Learning under Label Shift (RLLS) (Azizzadenesheli et al., 2019). To avoid arbitrarily bad estimation of the confusion matrix due to limited data size, RLLS makes the final distribution prediction less sensitive to the estimation performance of the confusion matrix by regularizing the ratio of test and training label distributions. RLLS is primarily designed to improve classification performance under label shift. The label distribution estimate is often a compromise between the BBSE result and the training distribution.

Maximum Likelihood Label Shift (MLLS) (Alexandari et al., 2019). Like BBSE, MLLS also takes a distribution matching approach to estimate the test set label distribution. The original algorithm uses an EM-based strategy (Saerens et al., 2002) to perform distribution matching in the input space of the test set. Alexandari et al. (2019) show that in combination with a particular post-hoc calibration method, MLLS outperforms BBSE.

5 Experiments and Results

5.1 Experimental Setup

We use the huggingface implementation of the BERT classifier fine-tuned on the corresponding dataset as the base predictor for all algorithms. The predictor is further calibrated using bias-corrected temperature scaling (BCTS) (Alexandari et al., 2019) for the MLLS method. All models are trained with AdamW optimizer (Loshchilov and Hutter, 2018) with a learning rate of 2e-5. All models are trained on a single Titan RTX GPU with a batch size of 32.

5.2 Main Results

We measure the RAE and CARC scores for all methods on the benchmark. RAE scores are averaged over all test splits for each dataset. We rank the performances with respect to RAE and CARC on each dataset and report the average ranking for each algorithm. The results are summarized in Table 2.

Some main observations from the table are:

- No algorithm outperforms others on all datasets.
- CC and PCC are still strong baselines. PCC performs better in most cases, but CC achieves significantly better results on Wikipedia Toxicity, where the positive class is rare, and the label shift is severe.

Method	Binary				Multi-Class				Average Rank
	Clothing	Electronics	Office	Toxicity	Clothing	Electronics	Office	CCD	
(RAE)(%)↓									
CC	3.05	2.07	3.93	**2.43**	9.87	12.61	11.02	9.20	3.63
PCC	2.63	2.44	**3.43**	32.79	6.74	7.14	**8.65**	12.41	3.13
BBSE	2.19	**1.90**	4.61	50.68	**5.60**	8.89	19.00	8.71	3.50
RLLS	2.51	3.22	7.37	49.47	14.10	29.92	28.23	8.56	5.25
MLLS	1.73	2.06	4.15	29.46	7.02	8.95	11.96	**6.13**	**2.75**
MLLS-BCTS	**1.26**	2.94	4.07	30.57	7.63	**7.13**	11.77	7.40	**2.75**
(CARC)↑									
CC	0.983	**0.994**	0.958	**1.000**	0.678	**0.784**	0.706	0.895	3.50
PCC	0.985	0.993	0.966	0.829	0.665	0.710	**0.746**	0.898	2.63
BBSE	0.985	0.993	0.965	nan*	**0.699**	0.695	0.616	0.892	4.13
RLLS	0.985	0.993	0.966	0.257	0.697	0.683	0.631	0.897	3.63
MLLS	0.985	0.993	**0.967**	0.314	0.685	0.721	0.733	**0.900**	**2.13**
MLLS-BCTS	**0.986**	0.993	**0.967**	0.314	0.663	0.698	0.733	0.896	3.00

Table 2: Quantification model performances in terms of average RAE (lower is better) and CARC (higher is better). MLLS-BCTS denotes MLLS with BCTS calibrated base predictor. Overall, MLLS performs the best, but not consistently outperforming others. **BBSE fails to produce non-zero prevalence estimates on all test sets, leading to an undefined CARC score.*

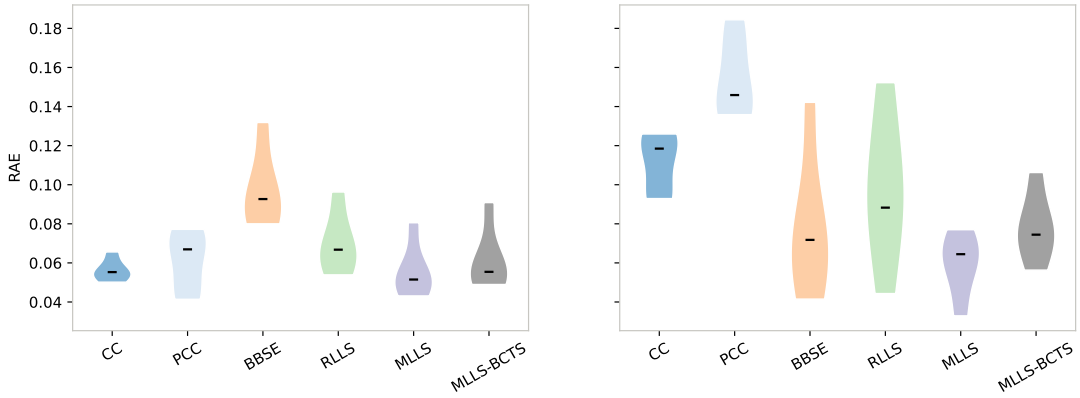


Figure 6: Distribution of RAE scores for test splits with the lowest 20% label shift (left) and with the highest 20% label shift (right) on CCD. CC and PCC's performances degrade significantly on test sets with a higher level of label shift.

- The performance of the MLLS algorithm, with or without BCTS calibration, is more consistent across all datasets than other algorithms.

- MLLS with BCTS calibration does not always have superior performance than the base version MLLS, contrary to what has been observed in previous studies (Alexandari et al., 2019).

- A better RAE score does not always indicate a better CARC score. For example, CC achieves a significantly better CARC score than other methods with a second-worst RAE score of 12.61% on Amazon Reviews (Electronics).

- BBSE fails to produce a non-zero prevalence estimate on all Wikipedia Toxicity test sets. This failure hints that BBSE might be unstable when predicting the prevalence of rare binary events.

5.3 Effect of Distribution Shift

In Section 3.3, we analyze the distribution shift estimates for each dataset across all the test splits. A natural question to ask is: how model performances change when the level of distribution shift increases? We sort the CCD test splits by label shift levels measured in KLD. We then take the bottom 20% and top 20% and visualize the RAE score distributions for all baseline algorithms in Figure 6.

We observe a significant performance degradation of CC and PCC methods on test splits with

Method	Standard	Balanced	% Change
(CCD)			
CC	9.20	19.44	+111.3%
PCC	12.41	19.77	+59.3%
BBSE	8.71	9.63	+10.6%
RLLS	8.56	13.84	+61.7%
MLLS	6.13	16.80	+174.1%
MLLS-BCTS	7.40	15.94	+115.4%
(Office)			
CC	11.02	14.09	+27.9%
PCC	8.65	19.94	+130.5%
BBSE	19.00	19.97	+5.1%
RLLS	28.23	27.34	-3.2%
MLLS	11.96	18.28	+52.8%
MLLS-BCTS	11.77	14.71	+25.0%

Table 3: Comparison of quantification performances in RAE (lower is better) using base classifiers trained with standard and balanced training set. Using a balanced training strategy almost always hurts quantification performance on CCD and Amazon Reviews (Office). BBSE is more robust to label distribution changes from stratified sampling during training.

higher levels of label shift. MLLS and MLLS-BCTS are less affected by the label shift. The difference is expected because the underlying base predictor is likely to overestimate or underestimate the label probabilities when the test split has a significantly different label distribution.

5.4 Effect of Balanced Training

In practice, when the training data is highly skewed in terms of label distribution, we often manually up-sample the rare class examples or assign more weights to them to facilitate training. This procedure changes the underlying data distribution and could significantly impact the quantification results if we use the classifier as our base predictor.

To analyze the effect of a balanced training procedure on the quantification performance, we finetune the same BERT classifier on both CCD and Amazon Reviews (Office) with a weighted random sampler so that all class examples are balanced. We then use this classifier as the base predictor for all baseline algorithms and compare the performances to the main results in Table 3.

When using a base predictor trained with a manually balanced dataset, the quantification performance almost always degrades. However, we can see from the percentage changes that BBSE is more robust to such performance degradation than other methods. For example, on CCD, BBSE is outperformed by MLLS when using a base classifier trained on the original training set. When switching to a balanced training setup, BBSE maintains a similar level of performance and betters MLLS. This property makes BBSE more preferable when label balancing is present during training.

5.5 Effect of Invariant Representation Learning

BBSE, RLLS, and MLLS all make a label shift assumption where the conditional distribution of $p(x|y)$ remains the same across training and test. However, this assumption does not always hold in practice. The content of a 1-star review on a product posted five years ago could be significantly different from a 1-star review posted today due to many factors, such as a change of consumer expectations in similar products.

To relax the label shift assumption, Tachet des Combes et al. (2020) propose to learn a domain-invariant representation and use a similar approach to BBSE to estimate the test set label distribution by performing distribution matching in the invariant latent space. Supposedly, such methods should perform better on test splits where the conditional distribution of the input features for each class drifts heavily from the training set. A significant drawback of the method is that the underlying model needs to be retrained for each test split.

We experiment with IWDAN model (Tachet des Combes et al., 2020) on both CCD and Wikipedia Toxicity datasets. On CCD, IWDAN shows a much worse RAE score of 49.18%. On Wikipedia Toxicity, however, IWDAN achieves an RAE score of 19.40%, the second-best result after CC. As the training and testing splits of Wikipedia Toxicity come from different sampling strategies, and considering IWDAN is devised mainly for domain adaption, the performance discrepancy might be due to a more significant domain change in Wikipedia Toxicity compared to CCD.

6 Conclusions

Quantification learning has an increasing number of applications yet is still less studied in the NLP community. In this paper, we propose the first text quantification benchmark with temporal distribution shift. Our experiments show that there is no baseline algorithm consistently outperforming others. We believe the proposed benchmark should enable new research into devising methods that can adapt to temporal changes and be reliably applied in practice.

7 Ethical Considerations

Data Access and License We develop our benchmark based on three publicly available datasets. Wikipedia Toxicity (Wulczyn et al., 2017) is published under the CC0[2] license. To the best of our knowledge, the Amazon review dataset (Ni et al., 2019) and the consumer complaint database are not associated with a license, but they are available for research purposes. The benchmark presented in this work is intended for research purpose only.

Data Anonymization For all datasets in the benchmark, we only preserve the information necessary for a quantification task: timestamp, input text, and the label. No user identifier or other information is present in the derived datasets. Time and currency value information have being anonymized by the original source for CCD.

References

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. 2019. Adapting to label shift with bias-corrected calibration. *arXiv preprint arXiv:1901.06852*.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. 2019. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*.

Oscar Beijbom, Judy Hoffman, Evan Yao, Trevor Darrell, Alberto Rodriguez-Ramirez, Manuel Gonzalez-Rivero, and Ove Hoegh Guldberg. 2015. Quantification in-the-wild: data-sets and baselines. *arXiv preprint arXiv:1510.04811*.

Antonio Bella, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana. 2010. Quantification via probability estimators. In *2010 IEEE International Conference on Data Mining*, pages 737–742. IEEE.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. 2018. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

[2] https://wiki.creativecommons.org/wiki/CC0

Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A Badhon, et al. 2020. Global wheat head detection (gwhd) dataset: a large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrea Esuli and Fabrizio Sebastiani. 2015. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(4):1–27.

George Forman. 2008. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Wei Gao and Fabrizio Sebastiani. 2016. From classification to quantification in tweet sentiment analysis. *Social Network Analysis and Mining*, 6(1):19.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*.

Pablo González, Alberto Castaño, Nitesh V Chawla, and Juan José Del Coz. 2017. A review on quantification learning. *ACM Computing Surveys (CSUR)*, 50(5):1–40.

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

Daniel J Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

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*, pages 3122–3130. PMLR.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

Lei Qi, Mohammed Khaleel, Wallapak Tavanapong, Adisak Sukul, and David Peterson. 2020. A framework for deep quantification learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 232–248. Springer.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41.

Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. 2020. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.