Quantifying and Alleviating Distribution Shifts in Foundation Models on Review Classification

Sehaj Chawla
SEAS, Harvard University
sehajchawla@g.harvard.edu

Nikhil Singh
Media Lab, MIT
nsingh1@mit.edu

Iddo Drori
EECS, MIT
idrori@mit.edu

Abstract

This work quantifies the extent to which accuracy degrades on review classification when state-of-the-art Transformer models are subjected to distribution shifts, and offers a solution to significantly decrease this degradation. We find differences in the extent of degradation depending on the independent variable across which the shift is created. Specifically, in our experiments time and sentiment shifts show upto 10% drops in accuracy; whereas shifts between industry and product sectors show 20-40% drops in accuracy. We provide ablation experiments with different Transformer architectures, such as BERT, T5 and Jurassic-I, and study their relationship with this degradation. The suggested solution reuses the base of the model trained on one distribution, in addition to fine-tuning the final dense layer in the model to support the new distribution that is seen once the model is deployed. This uses just 100-300 samples compared to the previous 10,000 samples from the unseen distribution, while decreasing the accuracy drops in half.

1 Introduction

We report the impact of distribution shifts on the accuracy of review classification when using Transformer models. More specifically, we look at the task of classifying customer reviews as fake or real based only on the review text, while reporting the extent of the drop in accuracy when the model tries to predict labels for distributions other than the one it was trained on. Investigating the performance deltas due to distribution shifts for this task is significant not only because of the dearth of labelled data sets, but also to gain insight into the information encoded by the Transformer embeddings and what steps may be taken to make their decisions more robust to possible shifts. Our results show that the extent of the degradation in accuracy depends primarily on the independent variable across which the shift is created. We use the available metadata to narrow down four independent variables that give us balanced training and testing data set splits while differing along the chosen variable. For each of these, we train and test across all four permutations of splits. The distribution shifts investigated are: (1) *Industry Type* - hotel and restaurant reviews, (2) *Time* - old (pre-2014) and new (post-2014) reviews, (3) *Product Type* - Japanese and Italian restaurant reviews, and (4) *Sentiment* - positive and negative reviews.

Since one of our goals is to gain insights into Transformer model selection for tasks that require robustness across distribution shifts, we use three popular constructs for Transformers: encoder only BERT [1] (Bidirectional Encoder Representations from Transformers) models, an encoder and decoder T5 [2] model, and the Jurassic-I [3] model with n-shot training. Subsequently, to address this problem of accuracy degradation due to distribution shifts, we suggest and report results from our solution of conventional fine-tuning - first training on the known distribution, then freezing weights for all but the final layer in the model, and tuning weights for this final layer with a much smaller subset of the new distribution (100-300 review text samples compared to the previous 10,000 samples) to allow the model a chance at using the generalizable patterns it saw in the first distribution, while also enabling it to create distribution-specific insights for the new distribution.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

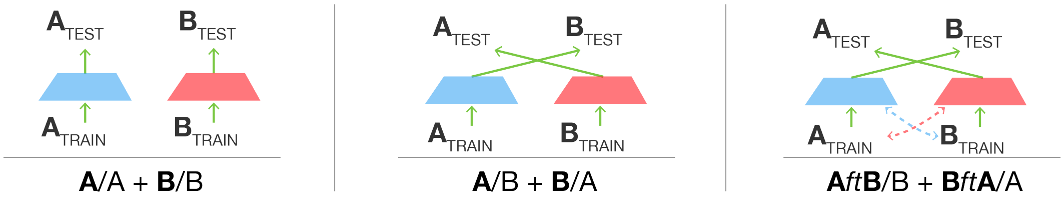

Figure 1: Our experimental setup. For a given pair of datasets (**A** and **B**), we perform three sets of train/test combinations. We train and test within the same distribution (**A**/A and **B**/B), between distributions (**A**/B and **B**/A), and between distributions with target fine-tuning (**A**_ft_**B**/B and **B**_ft_**A**/A).

Previous Work. Detecting fake reviews is a well known task, the economic implications of which have been analysed thoroughly in previous work [4], but with the growth of the industry for hiring and selling fake reviews, detecting fake reviews at scale has become a trade of its own and one particularly suited for the use of natural language processing [5]. We build on the same motivation by combining this NLP task of review classification with methodology partly based on existing work outside of NLP [6] that sets up structure for analysing implications of distribution shifts, and creating insights for model selection, and red flags in model training. Moreover, the architecture for our BERT instances are inspired by previous work [7] that created BERT models for review data sets. Our final (modified) model specifications are in the Implementation Details section, where we build on previous work by using a richer data set, trying three sizes of BERT, a T5 model, and then, most importantly, investigating and interpreting the performance of these models on distribution shifts. We also take inspiration from two notable works [8, 9], to suggest and report results from a solution of fine-tuning the model based on a small subset of the distribution-shifted data.

2 Methods

For our study, we used the following methodology (illustrated in the Fig. 1) that was partly based on previous work [6] on distribution shifts: (1) We began by standardising the review text to make them compliant with the pre-trained Transformer models' expected input, making sure all steps here were applicable to any other source's review texts. (2) We then finetuned our pre-trained Transformer models, evaluating the performance of the models on an out-of-sample test set in the same distribution, to ascertain how well the model does when it sees reviews similar to the ones it was trained on. This gives us baseline benchmarks (upper-bounds) to assess our distribution shift metrics. We made sure to achieve state of the art performance in this problem space by employing Transformer models that were previously shown to be most successful with the task. (3) For each of the aforementioned distribution shifts, we train and test within the same distribution (e.g. train and test both on pre-2014 reviews), as well as train and test across the distribution shifts (e.g. train on pre-2014 reviews and test on post-2014 reviews). We do so for all the different permutations for these shifts - employing BERT (3 size instances), T5, and Jurassic-I (with n-shot learning). (4) Lastly, we use the created models that were trained on one distribution, freeze the weights for all but the last layer, and fine-tune this layer based on a small subset of 100-300 review text samples from the new distribution. It is important to note here that fine-tuning is sensitive to the train set size, and our experiments only explore results while the split percentage is held constant. We do this for each split that was explored in the previous step, this time employing only the BERT and T5 instances in order to report this method as a solution to the degradation.

We use two labelled datasets: the first is for restaurant reviews from Yelp [10], and the second is for hotel reviews [11] which combines internet sources like Expedia, Hotels.com, Orbitz, Priceline, and TripAdvisor. Both datasets have the review text, fake/real labels, as well as some metadata. The metadata was used to find the independent variables along which we could split the data to create distribution shifts. Since our goal is to look at the generalizability of the models we create, and its translations to a different distribution (e.g. from a variety of sources), we decided to limit our input features to standardised review text only. We chose these datasets to work in conjunction, because they are both collections of consumer reviews, but are different in that the customers are restaurant clients in one and hotel clients in the other. We, therefore, found these datasets to be common enough

Table 1: Distribution Shifts (Train/Test), Accuracy scores for Industry (**R**estaurant vs. **H**otel), Time (**O**ld, Pre-2014 vs. **N**ew, Post-2014), Product (**J**apanese vs. **I**talian), Sentiment (**P**ositive vs. **N**egative)

	Industry Shift				**Time Shift**				**Product Shift**				**Sentiment Shift**			
Model	R/R	R/H	H/H	H/R	O/O	O/N	N/N	N/O	J/J	J/I	I/I	I/J	P/P	P/N	N/N	N/P
BERT (L)	0.67	0.42	0.87	0.38	0.67	0.65	0.69	0.62	0.87	0.64	0.78	0.63	0.99	0.80	0.99	0.85
BERT (S)	0.67	0.42	0.85	0.39	0.67	0.63	0.68	0.63	0.72	0.65	0.79	0.63	0.96	0.81	0.94	0.81
BERT (m)	0.67	0.48	0.84	0.39	0.68	0.63	0.71	0.63	0.80	0.64	0.81	0.63	0.98	0.77	0.97	0.77
T5 (S)	0.65	0.40	0.79	0.38	0.64	0.64	0.67	0.62	0.68	0.66	0.72	0.68	0.78	0.72	0.84	0.77

Table 2: Fine-tuning Solution (Train + Fine-tune/Test), Accuracy for Industry (**R**estaurant vs. **H**otel), Time (**O**ld vs. **N**ew), Product (**J**apanese vs. **I**talian), Sentiment (**P**ositive vs. **N**egative)

	Insudustry Shift		**Time Shift**		**Product Shift**		**Sentiment Shift**	
Model	R+H/H	H+R/R	O+N/N	N+O/O	J+I/I	I+J/J	P+N/N	N+P/P
BERT (L)	0.71	0.58	0.62	0.61	0.74	0.69	0.83	0.91
BERT (S)	0.74	0.56	0.61	0.65	0.73	0.71	0.87	0.92
BERT (m)	0.64	0.56	0.61	0.63	0.71	0.70	0.89	0.91
T5 (S)	0.66	0.57	0.63	0.63	0.66	0.68	0.72	0.79

to cross validate transfer learning, and at the same time, different enough to create an interesting distribution shift.

Implementation Details The BERT instances we use are: (1) LARGE BERT (uncased-base) (L-12_H-768_A-12) (2) SMALL BERT (uncased-base) (L-4_H-256_A-4), and (3) mobile BERT (uncased-base) (L-24_H-128_B-512_A-4_F-4_OPT). Each has an additional dense, dropout, and pooling layer added on top. We also experiment with a fine-tuned t5-small model, which is pre-trained on six attention modules. Finally, for Jurassic-1, we investigated the effects of distribution shifts in n-shot learning, and as such did not need to add any modifications or finetune the model.

3 Results

BERT and T5. From Table 1, we see the Transformer models do not translate well when exploring a distribution shift across industries and products specifically. While the models baselines are relatively high for the in-distribution testing (indicating that the datasets have no inherent problems in them), the models see a severe degradation in accuracy when trained and tested across these two distributions. The degradation is possibly a result of industry or product specific vocabulary used while reviewing. On the other hand, the results for time and sentiment shifts are promising in that the accuracy has relatively small drops when going from within the same distribution to a new distribution. This means that patterns in fake reviews are relatively constant across time and perhaps more surprisingly sentiment. After the distribution shift, large BERT has the best accuracy, and T5 has the smallest drop in accuracy. We believe the results are a reflection of the fact that vocabulary and other text distribution characteristics are not affected by time as much as they are by topic. In terms of sentiment, an interesting thing to note is that a higher accuracy is observed when training on negative reviews compared to positive reviews - which is possibly explained by the fact that negative reviews are on average 50% longer than than positive reviews.

From Table 2, we see very promising results since degradation in product and industry shifted data is much improved when we use our solution of fine-tuning the final layer in the model with 100-300 review text samples, consistently reducing the drop in accuracy by approximately half. This suggests that the models are able to pick up generalisable features across the shift, but need to see more examples from the new distribution to get closer to the upper bound accuracies. We see similar results for fine-tuning in sentiment-shifts too, where there are clear boosts in accuracy, but these still trail the baselines by 5-10%. For time-shifted data, however, the fine-tuning does not help the accuracies, perhaps because the drops were negligible to begin with.

Jurassic-I Few Shot Distribution Shifts. From Table 3, we see very similar results for Jurassic-I (compared to what we saw for BERT and T5). Overall, the shifts in industry and product are the most problematic for Jurassic-I to handle. The key differences are seen in the fact that Jurassic-I handles sentiment shifts more consistently than BERT or T5, but does worse on time-shifts. The

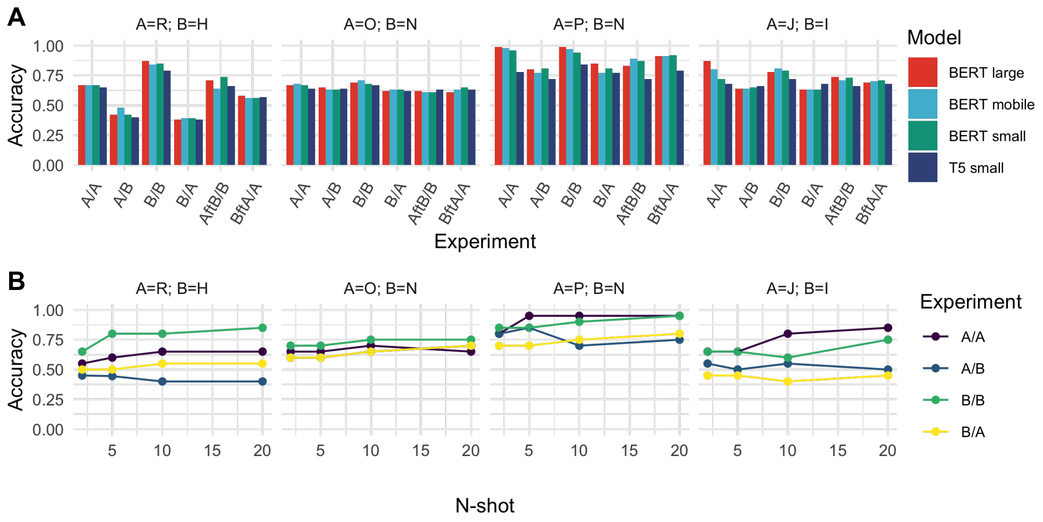

Figure 2: Our results. **(A)** accuracies obtained by different models on different train/test dataset combinations (A/B idnicates trained on A, tested on B, AftB indicates trained on with finetuning on B. **(B)** N-shot results for the Jurassic-I language model for A/B with different pairs of datasets. Dataset pairs are on Industry (**R**estaurant vs. **H**otel), Time (**O**ld, Pre-2014 vs. **N**ew, Post-2014), Product (**J**apanese vs. **I**talian), Sentiment (**P**ositive vs. **N**egative).

Table 3: Few Shot Distribution Shift (Train/Test), Accuracy for Industry (**R**estaurant vs. **H**otel), Time (**O**ld vs. **N**ew), Product (**J**apanese vs. **I**talian), Sentiment (**P**ositive vs. **N**egative)

	Industry Shift				**Time Shift**				**Product Shift**				**Sentiment Shift**			
N-shot	R/R	R/H	H/H	H/R	O/O	O/N	N/N	N/O	J/J	J/I	I/I	I/J	P/P	P/N	N/N	N/P
2-shot	0.55	0.45	0.65	0.5	0.65	0.6	0.7	0.6	0.65	0.55	0.65	0.45	0.8	0.8	0.85	0.7
5-shot	0.6	0.445	0.8	0.5	0.65	0.6	0.7	0.6	0.65	0.5	0.65	0.45	0.95	0.85	0.85	0.7
10-shot	0.65	0.4	0.8	0.55	0.7	0.65	0.75	0.65	0.8	0.55	0.6	0.4	0.95	0.7	0.9	0.75
20-shot	0.65	0.4	0.85	0.55	0.65	0.7	0.75	0.7	0.85	0.5	0.75	0.45	0.95	0.75	0.95	0.8

other insight of note is that increasing the number of shots in the few shot learning definitely helps improve the baseline accuracies (when the model is trained and tested on the same distribution), but makes the distribution-shifted accuracies worse - intuitively this is because the model is getting even more used to the specifics of the train data as the number of shots is increasing - making it harder for it to generalise to the shifted data.

4 Conclusion

We report the extent of the degradation of review classification accuracy for state-of-the-art Transformer models that are subjected to distribution shifts. Promising results are seen for the time and sentiment shifts, with all the models generalising to the shifted test sets reasonably well. On the other hand, industry and product shifted accuracies suffered greatly compared to the observed baseline metrics. For these two shifts in particular, the conventional solution of using a small subset of samples from the new distribution helps improve the accuracy considerably and consistently, indicating that this is a viable solution to allow models to continue to use generalisable patterns found in the training phase, and creating distribution-specific patterns in the fine-tuning phase. This study helps us quantify on the implications of distribution shifts in NLP classification tasks by highlighting the limitations of the Transformer models and their sensitivity to the characteristics of the training set, while creating insights for model selection and the requirement for adaptive training depending on the foreseeable distribution shift that a Transformer model is expected to encounter once deployed.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[3] Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 2021.

[4] Sherry He, Brett Hollenbeck, and Davide Proserpio. The market for fake reviews. *Available at SSRN*, 2020.

[5] Yafeng Ren and Donghong Ji. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385:213–224, 2017.

[6] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

[7] Stefan Kennedy, Niall Walsh, Kirils Sloka, Jennifer Foster, and Andrew McCarren. Fact or factitious? contextualized opinion spam detection. *arXiv preprint arXiv:2010.15296*, 2020.

[8] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[9] Baochen Sun, Jiashi Feng, and Kate Saenko. *Correlation Alignment for Unsupervised Domain Adaptation*, pages 153–171. Springer International Publishing, Cham, 2017.

[10] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 985–994, 2015.

[11] Myle Ott, Claire Cardie, and Jeffrey T Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 497–501, 2013.