Quantifying Distribution Shifts in Review Classification

Anonymous

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

This work demonstrates the extent of the degradation of review classification accuracy when state of the art transformer models are subjected to distribution shifts. We find that the extent depends primarily on the independent variable chosen, around which the shift is created. Time and sentiment shifts show 0-10% 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 and T5, and study their relation to this degradation. We also suggest a solution that 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 solution uses just 100-300 samples from the unseen distribution, while decreasing the accuracy drop in industry and product shifts by half.

Keywords: Review Classification, Distribution Shift, Transformers.

1. Introduction

We investigate and 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 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. Understanding the level of degradation in performance is significant in this domain not only because of the dearth of labelled training data sets, but also to gain intuition into the information encoded by the transformer embeddings and what steps can be taken to make their decisions more robust to shifts in distribution.

Our results illustrate that the extent of the degradation in accuracy depends primarily on the independent variable across which the shift was created. We used the available metadata to narrow down 4 independent variables that could give us balanced training and testing data sets while distinctly differing along the chosen variable. The distribution shifts we investigated were:

1. Industry Type - We train our models on reviews from one industry (restaurants) and then test on another (hotels), comparing these to benchmark metrics set by training and testing on the same industry type.

2. Time - We train on older reviews in the dataset (pre-2014) and test the performance on newer reviews (post-2014), comparing these to the baseline results of both training and testing on old reviews.

3. Product Type - We train on the reviews from one product (a popular Japanese restaurant) and evaluate how it performs on a new product (an Italian/Halal restaurant), while comparing performance to training and testing on the same product (the Japanese restaurant).

4. Sentiment - We train and test across the 4 permutations of positive and negative reviews based on sentiment labels assigned by the reviewers.

Since one of our goals was to gain insights into transformer model selection for tasks that require robustness against distribution shifts, we used the two most popular constructs for transformers: encoder only BERT (Bidirectional Encoder Representations from Transformers) models, and an encoder and decoder T5 model.

Finally, to address this problem of degradation due to distribution shifts, we report results from our solution of first training on the known distribution, then freezing weights for all but the final layer in the model, and then fine-tuning weights for this final layer with a much smaller subset of the new distribution (50-300 review text samples) to allow the model a chance at using the generalisable patterns it saw in the first distribution, while also enabling it to create distribution-specific insights for the new distribution.

1.1. Previous work

Detecting individual deceit in reviews is an old trade, the economic implications of which have been analysed thoroughly in previous work (He et al., 2020), but with the growth of the industry for hiring and selling fake reviews, detecting deceit on a mass scale has become a trade of its own and one particularly suited to the use of machine learning (Ren and Ji, 2017).

We, however, are combining this interesting NLP task of review classification with methodology partly based on existing work outside of NLP (Koh et al., 2020) 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 (Kennedy et al., 2020) 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 (Arjovsky et al., 2020) and (Sun et al., 2017), 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 that was partly based on previous work (Koh et al., 2020) 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 fine tuned 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. Now 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). If the dataset sizes and target balancing allow it, we do so for all the different permutations for these shifts - employing all the 4 transformer models (three instances of BERT and one of T5).

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 50-300 review text samples from the new distribution. We do this for each split that was explored in the previous step, and for all the 4 transoformer models, in order to report this method as a solution to the degradation we see through the results of the previous steps.

2.1. Dataset

We use two labelled datasets: the first is for restaurant reviews from Yelp (Rayana and Akoglu, 2015), and the second is for hotel reviews from a team of researchers at Cornell (Ott et al., 2013). 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 generalisability of the models we create, and its translations to a different distribution (perhaps 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 both include draws from the distribution 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 to cross validate transfer learning, while at the same time, being different enough to create an interesting distribution shift.

2.2. Implementation Details

Our first three models are fine-tuned BERT models, with additional dense layers to create fake/real labels. The last model is a fine-tuned T5 model. Below we give the specifications for their versions and instance sizes.

2.2.1. BERT

We experiment with the embeddings created by the following BERT instances:

- LARGE BERT (uncased-base) (L-12_H-768_A-12). Model Details [Click]

- SMALL BERT (uncased-base) (L-4_H-256_A-4). Model Details [Click].
- mobile BERT (uncased-base) (L-24_H-128_B-512_A-4_F-4_OPT). Model Details [Click].

The additional layers on top of the BERT embeddings include a pooling of the BERT embeddings into dense units equal to the number of units in the hidden size of the BERT model in question (the 'H' parameter in the names above). This is attached to a dropout layer with 10% dropout rate to account for over-fitting and to add some regularization. Finally, these values go through to a single sigmoid unit that makes the final prediction.

2.2.2. T5 Transformer

Since the T5 transformer model has built in encoder and decoder blocks, we didn't find the need to add any additional layers. We finetuned the t5-small model, which is pre-trained on six attention modules: Model Details [Click].

2.3. Code

Github Repository Link [Click]

3. Results

In this section, we report our results seen in the distribution shifts. All the metrics reported in the tables are OOD (out of dataset) accuracy, which is a fair metric for comparison here because the classes were sampled to be balanced. For other metrics including the f1 score, recall, precision and auc, please refer to the appendix.

3.1. Distribution Shift 1: Industry Type (Hotels and Restaurants)

Table 1: Distributional Shift on Industry Type (Hotels vs Restaurants), Accuracy

Model	**Training distribution/Testing distribution**			
	Restaurant/Restaurant	Hotel/Hotel	Restaurant/Hotel	Fine-tune Hotel/Hotel
BERT large	0.67	0.87	0.42	0.71
BERT small	0.67	0.85	0.42	0.74
BERT mobile	0.67	0.84	0.48	0.64
T5 small	0.65	0.79	0.40	0.66

Here, we see the transformer models do not translate well when exploring a distribution shift across industries for reviews. While the models are doing a decent job in predicting for restaurants after training on restaurants, and an even better job while training and predicting on hotels (indicating that the datasets have no inherent problems in them), the models see a severe degradation in accuracy when trained on restaurants and tested on hotels (i.e. when the distribution shift is applied). The degradation is possibly a result of industry specific vocabulary, and the change in review diction when reviewing food compared to reviewing hotel services. Interestingly, however, this degradation is much improved when

we use our solution of fine-tuning the final layer in the model with 300 review text samples, reducing the drop in accuracy to 10-20%. This suggests that the model is able to pick up generalisable features across the shift, but needs to see more examples from the new distribution to get closer to the upper bound accuracy.

3.2. Distribution Shift 2: Time (Old and New)

Table 2: Distributional Shift on Time (Pre-2014 vs Post-2014), Accuracy

	Training distribution/Testing distribution		
Model	Pre-2014/Pre-2014	Pre-2014/Post-2014	Fine-tune Post-2014/Post-2014
BERT large	0.67	0.65	0.62
BERT small	0.67	0.63	0.61
BERT mobile	0.68	0.63	0.61
T5 small	0.64	0.64	0.63

The results here are very promising, in that the accuracy has very small drops when going from within the same distribution (pre-2014) to a new distribution (training on pre-2014 and testing on post-2014). This means that patterns in fake reviews are relatively constant across time. 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. We also see that the fine-tuning in this case doesn't help the model by much (and, in fact, increases the accuracy drop marginally). This is also an indication of the reviews across this distribution being similar enough to not need any fine-tuning.

3.3. Distribution Shift 3: Product Type (Japanese/Italian and Halal)

Table 3: Distributional Shift on Cuisine (Japanese vs Italian/Halal), Accuracy

	Training distribution/Testing distribution		
Model	Japanese/ Japanese	Japanese/ Italian	Fine-tune Italian/Italian
BERT large	0.87	0.64	0.74
BERT small	0.72	0.65	0.73
BERT mobile	0.80	0.64	0.71
T5 small	0.68	0.66	0.66

The accuracy drops by 10-20% when going from within the same distribution (Japanese cuisine) to a new distribution (Italian/Halal cuisine), but this drop in distribution-shifted accuracy is close to the range of accuracies for the benchmark case in which the model was trained and tested on restaurant reviews of all cuisines. A possible interpretation is that the model picked up on both generalisable and industry specific patterns of fake reviews giving a higher within-distribution accuracy but returning to the baseline all-cuisine accuracy when

predicting on the distribution shift. Looking more closely, we also see that the T5 model once again suffers from a much smaller drop in accuracy, and its within distribution accuracy indicates that it always only picked up the generalised patterns in the reviews, independent of the product. Moreover, we notice that the fine-tuning method shows similar results to that in the Industry shift, where the accuracy drop is reduced by 10% for all the BERT models - acting as further evidence that the base model was picking up the generalisable patterns, but needed fine-tuning to get closer to the Italian/Halal-specific patterns.

3.4. Distribution Shift 4: Sentiment (Positive and Negative)

Table 4: Distributional Shift on Sentiment (Positive vs Negative), Accuracy

Model	**Training distribution/Testing distribution**					
	+ve/+ve	+ve/-ve	Fine-tune -ve/-ve	-ve/-ve	-ve/+ve	Fine-tune +ve/+ve
BERT large	0.99	0.80	0.83	0.99	0.85	0.91
BERT small	0.96	0.81	0.87	0.94	0.81	0.92
BERT mobile	0.98	0.77	0.89	0.97	0.77	0.91
T5 small	0.78	0.72	0.72	0.84	0.77	0.79

This distribution shift showed promising results, maintaining a high prediction accuracy even on the shifted groups (positive/negative and negative/positive). The most interesting aspect here is the higher accuracy observed when training with negative reviews compared to positive reviews. A possible explanation could be related to negative reviews being roughly 50% longer than positive reviews on this data set (average and median values), meaning that models learning from negative reviews have more information to base their predictions on. Nevertheless, given the results' similarity level, it is safe to conclude that the recognisable text patterns created are the same regardless of whether the sentiment is positive or negative. Furthermore, the fine-tuning consistently provides accuracy boosts across this distribution shift while still being 5-10% away from the upper bound.

4. Conclusion

We have reported the extent of the degradation of review classification accuracy when state of the art transformer models are subjected to distribution shifts. Promising results were seen for the time and sentiment shifts, with all the models generalising to the shifted test set reasonably well. On the other hand, industry and product shifted test set accuracies suffered greatly compared to the observed baseline metrics. For these two shifts in particular, the solution of using a small subset of samples from the new distribution helped improve the accuracy considerably and consistently, indicating that this is a viable solution to allow models to continue to use generalisable patterns it found in the main training phase, and create distribution-specific patterns in the fine-tuning phase.

This study helps us comment on the more far-reaching 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

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

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

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

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.

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.

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.

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

Baochen Sun, Jiashi Feng, and Kate Saenko. *Correlation Alignment for Unsupervised Domain Adaptation*, pages 153–171. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58347-1. doi: 10.1007/978-3-319-58347-1_8. URL https://doi.org/10.1007/978-3-319-58347-1_8.

5. Appendix

Table 5: Distributional Shift on Time: Train on Pre-2014, Test on Pre-2014, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.67	0.68	0.66	0.67	0.67
BERT small	0.66	0.68	0.64	0.67	0.67
BERT mobile	0.67	0.69	0.64	0.68	0.68
T5 small	0.70	0.60	0.83	0.64	0.64

Table 6: Distributional Shift on Time: Train on Pre-2014, Test on Post-2014, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.69	0.69	0.61	0.65	0.65
BERT small	0.67	0.67	0.53	0.63	0.63
BERT mobile	0.68	0.67	0.54	0.63	0.63
T5 small	0.69	0.62	0.77	0.64	0.64

Table 7: Distributional Shift on Time: Train on Pre-2014, Fine-tune on Post-2014 and Test on Post-2014, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.68	0.63	0.67	0.61	0.62
BERT small	0.68	0.62	0.76	0.60	0.61
BERT mobile	0.68	0.52	0.72	0.60	0.61
T5 small	0.63	0.56	0.62	0.62	0.63

Table 8: Distributional Shift on Time: Train on Positive, Test on Positive, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.99	1.00	0.99	0.99	0.99
BERT small	0.96	0.96	0.97	0.96	0.96
BERT mobile	0.98	0.99	0.96	0.98	0.98
T5 small	0.78	0.85	0.71	0.79	0.79

Table 9: Distributional Shift on Sentiment: Train on Positive, Test on Negative, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.77	0.95	0.65	0.80	0.80
BERT small	0.79	0.88	0.72	0.81	0.81
BERT mobile	0.74	0.86	0.65	0.77	0.77
T5 small	0.77	0.66	0.94	0.72	0.72

Table 10: Distributional Shift on Sentiment: Train on Positive, Finetune Negative, Test on Negative, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.85	0.85	0.80	0.83	0.83
BERT small	0.77	0.79	0.76	0.87	0.87
BERT mobile	0.86	0.89	0.86	0.89	0.89
T5 small	0.63	0.56	0.82	0.72	0.72

Table 11: Distributional Shift on Sentiment: Train on Negative, Test on Negative, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.99	0.98	0.99	0.99	0.99
BERT small	0.94	0.94	0.94	0.94	0.94
BERT mobile	0.97	0.97	0.97	0.97	0.97
T5 small	0.83	0.88	0.79	0.84	0.84

Table 12: Distributional Shift on Sentiment: Train on Negative, Test on Positive, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.85	0.86	0.84	0.85	0.85
BERT small	0.82	0.80	0.84	0.81	0.81
BERT mobile	0.76	0.78	0.75	0.77	0.77
T5 small	0.74	0.83	0.67	0.77	0.77

Table 13: Distributional Shift on Sentiment: Train on Negative, Finetune Positive, Test on Positive, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.94	0.96	0.89	0.91	0.91
BERT small	0.96	0.97	0.85	0.93	0.92
BERT mobile	0.86	0.89	0.95	0.91	0.91
T5 small	0.75	0.70	0.83	0.78	0.79

Table 14: Distributional Shift on Cuisine: Train on Japanese Cuisine, Test on Japanese Cuisine, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.88	0.83	0.93	0.87	0.87
BERT small	0.73	0.71	0.75	0.72	0.72
BERT mobile	0.80	0.80	0.81	0.80	0.80
T5 small	0.73	0.63	0.88	0.68	0.68

Table 15: Distributional Shift on Cuisine: Train on Japanese Cuisine, Test on Italian and Halal Cuisines, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.69	0.69	0.61	0.65	0.65
BERT small	0.67	0.67	0.53	0.63	0.63
BERT mobile	0.68	0.67	0.54	0.63	0.63
T5 small	0.69	0.62	0.77	0.64	0.64

Table 16: Distributional Shift on Cuisine: Train on Japanese Cuisine, Test on Italian and Halal Cuisines, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.69	0.69	0.67	0.74	0.74
BERT small	0.65	0.69	0.68	0.73	0.73
BERT mobile	0.67	0.79	0.58	0.71	0.71
T5 small	0.59	0.72	0.63	0.66	0.66

Table 17: Distributional Shift on Time: Train on Restaurant, Test on Restaurant, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.67	0.67	0.66	0.67	0.67
BERT small	0.67	0.67	0.65	0.67	0.67
BERT mobile	0.66	0.68	0.63	0.67	0.67
T5 small	0.67	0.64	0.69	0.65	0.65

Table 18: Distributional Shift on Product Type: Train on Restaurant, Test on Hotels, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.67	0.32	0.13	0.42	0.42
BERT small	0.67	0.30	0.12	0.42	0.42
BERT mobile	0.67	0.43	0.16	0.48	0.48
T5 small	0.32	0.37	0.28	0.40	0.40

Table 19: Distributional Shift on Product Type: Train on Hotels, Test on Hotels, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.85	0.89	0.90	0.87	0.87
BERT small	0.84	0.85	0.88	0.85	0.85
BERT mobile	0.84	0.81	0.88	0.84	0.84
T5 small	0.79	0.79	0.79	0.79	0.79

Table 20: Distributional Shift on Product Type: Train on Restaurants, Fine-tune on Hotels, Test on Hotels, Metrics

Model/Metric	F1	Precision	Recall	AUC	Accuracy
BERT large	0.77	0.63	0.97	0.71	0.71
BERT small	0.69	0.85	0.57	0.74	0.74
BERT mobile	0.67	0.71	0.15	0.64	0.64
T5 small	0.76	0.57	0.7	0.66	0.66