Quantifying Distribution Shifts in Review Classification

Anonymous

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Online, user generated reviews drive and inform consumer choice. Their continuing usefulness, however, depends on trust that these reviews are truthful, which has motivated the development of algorithms that detect fake review based on review text. The dearth of publicly available training datasets of reviews with fake/real labels, however, raises concerns about the generalizability of these algorithms. We test this directly by exploring the effects of distribution shifts (with respect to time, industry type, product type, and sentiment) on the performance of four pre-trained and fine-tuned transformer models. The first three models are neural nets built on top of three pre-trained instances of BERT (large, small, and mobile), which generate contextualised embeddings of review text in a way that is mechanically independent from our training dataset. Our fourth model, is the small T5 transformer.

Keywords: Natural Language Processing, Distribution Shifts, BERT, T5, Review Classification

1. Introduction

Consider a basic, fake-review model that takes as its input the review text, and returns a binary label of fake, or real. How the review text is converted into features is a key component of the design, with implications for the generalizability of the model, and its understandings of distributions other than the one it was trained on.

We will use four models. Three are based on BERT (Bidirectional Encoder Representations from Transformers), and the fourth is a fine-tuned T5 model.

BERT is a pre-trained language model that converts text into a vector that represents the text's meaning, taking words not in isolation but in the context in which they appear. Because BERT is pre-trained, the text embeddings it generates to capture text meaning is mechanically independent of the training dataset, and of the training objective. We use 3 sizes of BERT models (Large , Small, and Mobile), to create our contextualised embeddings.

On the other hand, T5 is an encoder-decoder transformer model (different from BERT which only has an encoder and no decoder) and converts all NLP problems into a text-to-text format. If these transformers can tease out text meaning regardless of the style of the distribution shifts, it can help the algorithm detect generalizable text patterns that signal fakeness.

Our question through out this research will be to investigate the effects on prediction performance of transfer learning when SOTA pre-trained transformer models are trained on different distributions from what they are predicting on.

We consider 4 types of distributional shifts:

- Industry Type - We explore how the transformer models perform when trained on reviews from one industry (restaurants) and then tested on another (hotels).
- Time - Here, we train on older reviews in the dataset (pre-2014) and test the performance on newer reviews (post-2014).
- Product Type - Here we train on the reviews from one product (a popular Japanese restaurant) and evaluate how it performs on a new product (Italian/Halal restaurant)
- Sentiment - Finally, we look at the effects on performance when we train and test across positive and negative reviews to see if the fakeness continues to be identified correctly by the transformers.

It is important to note that along with evaluating the distribution shifts (i.e. training on one distribution and testing on the other), we also report results for training and testing on the same distribution as comparison metrics to hold as the baseline. All the metrics reported in the tables are Out of Dataset (OOD) i.e. we do not train and test on the same set of reviews ever.

1.1. Previous work

Detecting individual deceit is an old trade, 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 use machine learning. Existing work (He et al., 2020) analyzes economic effects of fake review markets and methods employed by sellers, providing insights about the prominence of this matter for current e-commerce enterprises. From a machine learning standpoint, (Ren and Ji, 2017) lays important groundwork for studying deceptive opinion spammers, comparing algorithms based on linguistics and psychological features with a newly introduced approach that attempts to use neural networks and document-level representation to introduce contextual understanding to machines. This work focuses more on an actual application.

Moreover, the model architecture for our BERT instances was 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, but we also build on that 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.

2. Methods

For our work, we used the following methodology that was partly based on previous work (Koh et al., 2020) as well:

1. Parse and pre-process the review text to make it compliant to the pre-trained transformer models' expected input (includes lower casing the text, removing unknown character, links, and other such problematic snippets). We do this for:
    - The Yelp data set.

- The opinion spam data set.

For descriptions of these, look at section Dataset.

2. Fine tune our pre-trained transformer models (the different BERT sizes followed by our custom architecture, and the T5 model). Then, evaluate the performance of the models on an out of sample test set in the same distribution, to ascertain how well the model does in the same domain. This gives us baseline benchmarks to compare metrics with. We do this once again for:

   - The Yelp data set.
   - The opinion spam data set.

3. Now for each of the following shifts (descriptions for each are available in Introduction):

   - Industry Type
   - Time
   - Product Type
   - Sentiment

   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 distributions (e.g. train on pre-2014 reviews and test on post-2014 reviews). If the dataset sizes and balance allow it, we do so for all the different permutations for these shifts. And all of this is done for each of the 4 models (3 BERT instances and one T5 model).

Note: Our GitHub repository (Code), follows a directory structure based on the steps' description above - i.e. each step has its own directory, followed by sub directories for each of the unordered list pointers.

2.1. Dataset

We use two datasets. The first dataset is of restaurant reviews and is from Yelp; it includes 16,000 reviews with binary labels (sampled so that the two classes are equally balanced). Each of these reviews have been classified as "filtered" (Yelp's term for fake or suspicious), or "unfiltered". Although the data is augmented by metadata on the characteristics of the restaurant, reviewer etc., our goal is to look at the generalisability/flexibility of the models we create, and its translations to a different distribution. As such, we have decided to predict solely based on the review text itself, finding text patterns that signal suspicious reviews, and then applying this to other reviews, which might not have the metadata attached. The dataset was first studied in a notable previous work (Rayana and Akoglu, 2015).

Our second dataset is of hotel reviews, and is from a team of researchers in Cornell (Ott et al., 2013); it includes 800 labelled reviews, and the two classes of labels are equally balanced. True reviews were selected using a mix of rule based sort and human judgement from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp. Fake reviews were generated by hiring Turkers located in the United States, and who maintain an approval

rating of at least 90%. This dataset is only made of review text; thus, we restrict our features to review text only. We chose this dataset because, like Yelp, they are draws from the distribution of consumer review, but different in that the customers are hotel clients, not restaurant clients. A priori, we believed that these datasets are common enough to cross validate transfer learning, while at the same time, being different enough to make it an interesting exploration of learning in one context (hotel reviews) generalises.

3. Results

3.1. Implementation Details

Our first three models build on top of three pre-trained BERT models. We describe these first before describing, in brief, the T5 model that we fine-tuned.

3.1.1. BERT

In order to examine generalizability across datasets, we want our model features to be ones can be extracted from both of our datasets. Thus, we restrict our feature space to review text. Review text is transformed into embeddings, each of which correspond to a separate model.

We experiment with the following features:

- Pre-trained BERT (uncased base) embeddings from LARGE BERT (L-12_H-768_A-12). Model URL [Click]
- Pre-trained BERT (uncased base) embeddings from SMALL BERT (L-4_H-256_A-4). Model URL [Click].
- Pre-trained BERT (uncased base) embeddings from mobile BERT (L-24_H-128_B-512_A-4_F-4_OPT). Model URL [Click].

The architecture of this model is shown in figure: Model Architecture, and includes a pooling of the BERT embeddings into a dense layer, with its number of units equal to 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 overfitting and to add some regularization. Finally, these values go through to a single sigmoid unit that makes the final prediction.

3.1.2. T5 Transformer

The T5 transformer model builds on previous work on Transformer models; whereas BERT only has encoder blocks, T5 uses encoder and decoder blocks. We use the t5-small model, which was trained on six attention modules. More about the model, and training the model is described in the HuggingFace documentation: Model URL [Click].

3.2. Code

Github Repository Link [Click]

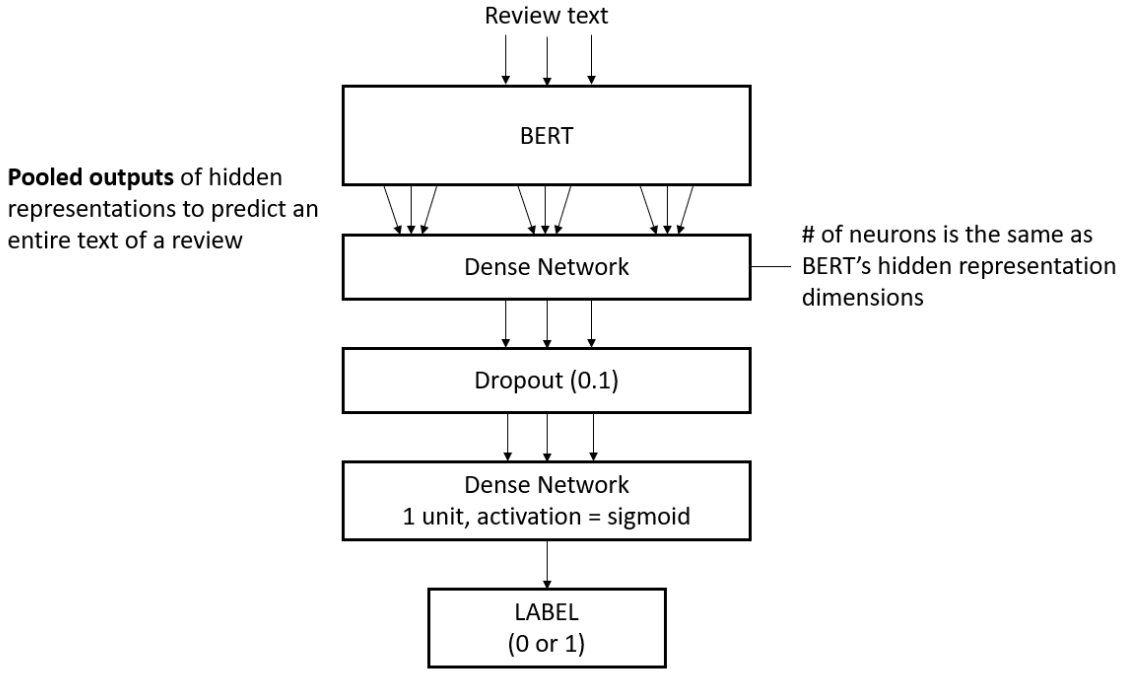

Figure 1: Model Architecture

3.3. Distribution Shifts

In this section, we report and discuss the results that were seen across each of the 4 distribution shifts we explored. All the metrics reported in the tables here 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.

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

Datasets: Yelp and OpSpam

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

Model	**Training dataset/Testing dataset**		
	Restaurant/Restaurant	Hotel/Hotel	Restaurant/Hotel
BERT large	0.67	0.87	0.42
BERT small	0.67	0.85	0.42
BERT mobile	0.67	0.84	0.48
T5 small	0.65	0.79	0.40

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

Datasets: Yelp

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

Model	**Training dataset/Testing dataset**	
	Pre-2014/Pre-2014	Pre-2014/Post-2014
BERT large	0.67	0.65
BERT small	0.67	0.63
BERT mobile	0.68	0.63
T5 small	0.64	0.64

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

Datasets: Yelp

Table 3: Distributional Shift on Cuisine (Japanese cuisine vs Italian and Halal cuisines), Accuracy

Model	**Training dataset/Testing dataset**	
	Japanese/ Japanese	Japanese/ Italian, Halal
BERT large	0.87	0.64
BERT small	0.72	0.65
BERT mobile	0.80	0.64
T5 small	0.68	0.66

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

Datasets: OpSpam

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

Model	**Training dataset/Testing dataset**			
	Positive/Positive	Positive/Negative	Negative/Negative	Negative/Positive
BERT large	0.99	0.80	0.99	0.85
BERT small	0.96	0.81	0.94	0.81
BERT mobile	0.98	0.77	0.97	0.77
T5 small	0.78	0.72	0.84	0.77

4. Conclusions

4.1. Distribution Shift Specific Conclusions

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

From section Distribution Shift 1: Industry Type (Hotels and Restaurants) 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, the models do badly when trained on restaurants and tested on hotels (i.e. when the distribution shift is applied). All the models have predictions that are worse than a 50-50 random guess, but mobile BERT performs the best out of them. We also look at the baseline performance and see that when trained on hotels, the performance for hotels is really good (≈80% accuracy), and so there is no problem with the patterns being hard to find in hotels, it is just that these patterns differ from the patterns found in restaurants. Although it is very hard to ascertain why this is so, we think it might be because of the industry specific vocabulary, where review diction changes considerably when reviewing food compared to reviewing hotel services.

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

The results in section Distribution Shift 2: Time (Old and New) are very promising, in that the OOD 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 in the given time frame (unfortunately there is no data available to report results for more recent trends). Within the same distribution, all models perform almost equaly well, with mobile BERT doing marginally better. After the distribution shift large BERT has the best accuracy, and T5 has the smallest drop in accuracy. We believe these results are a reflection of the fact that fake reviewers have standardised ways of writing these reviews and they haven't changed their methods very often.

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

The results in section Distribution Shift 3: Product Type (Japanese/Italian and Halal) seem to be an illustration on the gains to fine-tuning: the OOD accuracy drops by 10-20% when going from within the same distribution (Japanese cuisine) to a new distribution (Italian, Halal cuisines), but that drop brings the model OOD accuracy closer to the range of model OOD accuracies for the benchmark case in which the model was trained and tested on restaurant reviews of all-cuisines. When trained on Japanese cuisine, the within-distribution OOD accuracy is between 68- 87%; this compares to a within-distribution OOD accuracy of 65-67% when the model is trained on all cuisines. However, once the model trained on reviews of Japanese cuisine is tested on reviews of Italian and Halal cuisines, the OOD accuracy drops to 64% to. 65%. A possible interpretation is that the model picked up on both generalizable and industry specific patterns of fake reviews.

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

This distribution shift in section Distribution Shift 4: Sentiment (Positive and Negative) 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 out-of-sample accuracy observed when training with negative reviews when compared to training with positive reviews (0.85 vs 0.80 and 0.77 vs 0.72 for BERT large and T5 small, respectively). 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 people writing fake reviews use the same language techniques regardless if they're writing positive or negative reviews. This also corroborates results from the time distribution shift and is evidence supporting the argument that the way fake reviews are written for specific products within the industry are relatively consistent. This is a promising finding for future work on this area.

4.2. General Conclusions

4.2.1. Visualisation

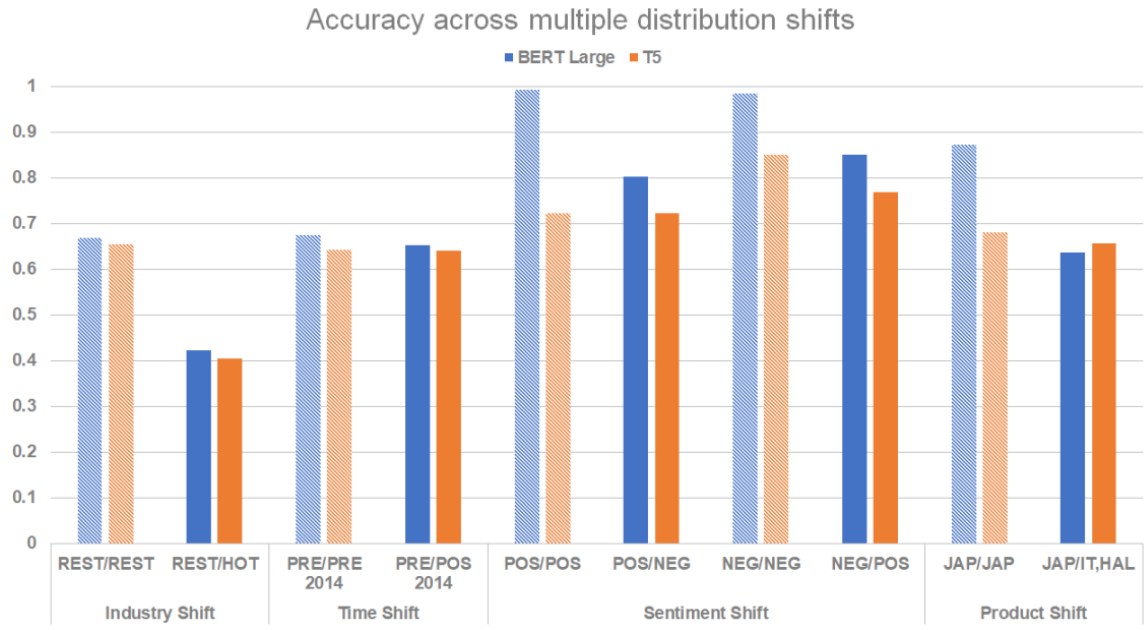

Figure 2: Final Comparisons

NOTE: In figure: Final Comparisons, the solid bars represent shifted distributions, and the hashed bars represent the baselines (non-shifted distribution) results.

4.2.2. Final Takeaways

With the help of figure: Final Comparisons, we summarise our final takeaways below:

- The best results were achieved in the time shift, potentially indicating that methods employed by fake reviewers do not change much over time.
- Although it performed consistently worse than BERT in every category, the T5 model appears to have higher generalizability as it experienced lower performance drops across the distribution shifts.
- Generalization and transfer learning in NLP remains a challenge - we speculate from our results that indicators of "fakeness" and prediction power have a very large industry-specific component (vocabulary).

References

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.

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 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 8: 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 9: 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 10: 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 11: 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 12: 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 13: 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 14: 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 15: 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