Lost in Distillation: A Case Study in Toxicity Modeling

Alyssa Chvasta and **Alyssa Lees** and **Jeffrey Sorensen** and
Lucy Vasserman and **Nitesh Goyal**
Google Jigsaw, New York
achvasta,alyssalees,sorenj,lucyvasserman,teshg@google.com

Abstract

In an era of increasingly large pre-trained language models, knowledge distillation is a powerful tool for transferring information from a large model to a smaller one. In particular, distillation is of tremendous benefit when it comes to real-world constraints such as serving latency or serving at scale. However, a loss of robustness in language understanding may be hidden in the process and not immediately revealed when looking at high-level evaluation metrics. We investigate the hidden costs: what is "lost in distillation", especially in regards to identity-based model bias using the case study of toxicity modeling. With reproducible models using open source training sets, we investigate models distilled from a BERT teacher baseline. Using both open source and proprietary big data models, we investigate these hidden performance costs.

1 Introduction

The revolution in natural language processing brought on by transformers, which have now been employed in virtually all major text processing applications, also brought substantially higher computational costs. The typical BERT model (Devlin et al., 2019) has over 100M parameters and 12 layers. The prospect of using these models in production settings without special purpose hardware quickly led practitioners to seek techniques to reduce the computational costs.

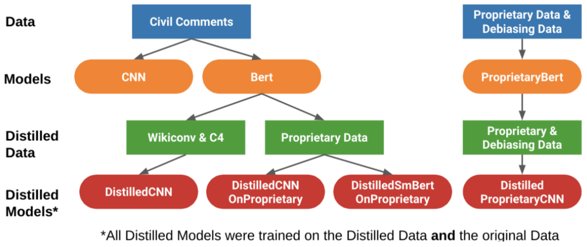
Data
Civil Comments
Proprietary Data & Debiasing Data
Models
CNN
Bert
ProprietaryBert
Distilled Data
Wikiconv & C4
Proprietary Data
Proprietary & Debiasing Data
Distilled Models*
DistilledCNN
DistilledCNN OnProprietary
DistilledSmBert OnProprietary
Distilled ProprietaryCNN
*All Distilled Models were trained on the Distilled Data **and** the original Data

Figure 1: Map of data and results presented

An approach widely advocated is to employ the technique of *knowledge distillation* to improve the performance of a simpler *student model* by training on additional unsupervised data that has been labeled by the larger *teacher model* (Hinton et al., 2015).

The ability to draw upon the wellspring of nearly unlimited unsupervised data and to leverage the higher performance of a much larger model, while maintaining the lower serving costs of a smaller model, has led to rapid adoption of this practice. However, closer analysis of the performance of distilled models reveals that while they may be able to erect a facade of high accuracy, they fail to capture important aspects of the knowledge represented in the teacher models.

We present a particular method of using distillation that we used to improve the performance of our models through pseudo-labeling of unsupervised data, while retaining the model architecture and number of parameters. While, for some metrics we saw nearly asymptotic performance to the teacher model, using other metrics we discovered important differences. While we do not know if this problem will manifest across all differences in architecture and parameterization - we want to caution researchers who are exploring distillation as a potential quick fix.

2 Related Work

BERT models and transformer models in general have structures that are layered with computation units that limit the degrees that parallelism can be used. Focusing on task performance alone, as is often the case for benchmark tasks, has been criticized for failing to account for resource costs (Ethayarajh and Jurafsky, 2020). Knowledge distillation is one of many techniques authors have proposed schemes to reduce the size and complexity.

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 92 - 101
July 14, 2022 ©2022 Association for Computational Linguistics

Models with unintended biases has received considerable attention with multiple survey papers both generally (Pessach and Shmueli, 2022) and for natural language in particular (Kurita et al., 2019; Czarnowska et al., 2021).

Two popular implementations of the distillation paradigm of creating a vast training set using large models to label unsupervised data are presented in Jiao et al. (2020) and Sanh et al. (2020). The primary goal of this work is producing a model with similar performance characteristics on the target task, but with lower a resource footprint. Turc et al. (2019) suggests pre-training and fine-tuning compact models as an alternative to traditional distillation. However, the effects on model bias were not reported in these studies.

Several other works explore this idea in modes similar to the work we present here, although often with a different array of model architectures. Wasserblat et al. (2020) and Mangalwedhekar (2021) both include CNNs as one of the target models. Tang et al. (2019); Chia et al. (2019); Adhikari et al. (2020) all present additional studies regarding distillation and the performance of the models in terms of fidelity to the teacher model.

Specifically regarding bias in the distillation or model compression setting, Xu and Hu (2022) report reduction in bias in contrast to our findings, although in a generation application. However, Gupta et al. (2022) makes clear that biases from the training data can also be preserved or exacerbated in a similar distillation setting.

Bender et al. (2021) raises several risks of large language models overall, including identity-based bias. We show that these risks can be magnified with the use of distillation, and that high-level accuracy metrics can hide nuances in performance, especially when large models are built to address a wide range of use cases.

3 Toxicity Modeling

We have chosen to use the problem of "toxic" comment classification to illustrate the difficulty that we observed in distillation. This is due to the ready availability of training resources for this task, the practical real-world need to address this problem, and the clear risks (Xu et al., 2021) of identity term bias and other modeling pitfalls.

Several diagnostic frameworks that were proposed to highlight the limitations of classification systems *in general* can also be used to highlight the problems with distillation in particular. Our primary framework is the method of measuring classifier unintended bias associated with neutral or ambiguous identity terms. This framework was introduced in Dixon et al. (2018) and expanded in Borkan et al. (2019) along with the Civil Comments dataset that is our primary source of supervised training data. In addition we use the diagnostic HateCheck test set (Röttger et al., 2021). Recently works that study implicitly abusive language (Wiegand et al., 2021; Lees et al., 2021), where careful attention to the context and implication of the comments is required. We include these evaluation challenges for our models.

4 Models

We found the bias effects of distillation to be remarkably persistent from a small to a very large scale. We created smaller, reproducible models entirely from publicly available resources, and duplicated the same findings on a very large model to show the generality of these findings. Table 1 provides a list of data sources and models described in the next sections. [1]

4.1 Teacher Models

We trained state of the art text classification models using both publicly available resources, and a larger model trained on resources that we are not authorized to release. Here, our intent is to show that the effects persist into the big data domain.

4.1.1 Civil Comments based Models

All of the models described in this section are based upon publicly available resources and data. The Civil Comments dataset introduced in Borkan et al. (2019) is a public domain corpus of 1.8M user comments labeled for toxicity by crowd raters. These comments originated from a distributed commenting platform that ceased operation in 2017. A subset of the data, ∼400K comments were additionally rated for specific identity subgroup associations such as gender, religion, or sexual orientation. The identity labels in the test set are used for bias evaluation.

Our Civil Comments based models were constructed both for the purposes of reproducibility and for experiments in distillation size. All of these

[1] A Python notebook demonstrating the ideas presented in this paper can be found at http://github.com/conversationai/Lost_in_Distillation.

Model	Data Sources	Training Instances
CNN	Civil Comments	1.8M
Bert	Civil Comments	1.8M
ProprietaryBERT	Civil Comments + Human Labeled Proprietary (3M) + Bias Mitigation (2M)	6.8M
DistilledCNN	Civil Comments + WikiConv (400K) + C4 (640k)	2.8M
DistilledCNNOnProprietary	Civil Comments + BERT-labeled proprietary (20M)	21.8M
DistilledSmBERTOnProprietary	Civil Comments + BERT-labeled proprietary (20M)	21.8M
DistilledProprietaryCNN	proprietaryBERT data + ProprietaryBERT-labeled proprietary (28M) + Bias Mitigation (1.7M)	36.5M

Table 1: Model Training Data Size

were fine-tuned or trained only using the public domain Civil Comments training corpus. Also for the sake of reproducibility, all BERT model versions used open-source checkpoints. It should be noted that in addition to models listed below, we also experimented with distilling via alternate compact architectures. The results were worse in terms of performance and as such we omitted the results.

All CNN models are trained until convergence. For these models, no bias mitigation or data enhancement was employed. Some discrepancies between the big data models and the Civil Comments models, both in overall results metrics and bias, are due to these differences in data.

CNN A baseline CNN trained exclusively on Civil Comments data with a BERT-base checkpoint as initial embedding. With 5 layers (2-gram, 3-gram, 4-gram, 5-gram and 6-gram layers of 300) and a max pooling layer. The model hyperparameters were tuned on a held-out evaluation set. The final model employed batch size of 64, max token sequence length of 1536 and learning rate of $1e-5$. The hyper-tuned parameters were used for all of the distilled CNN student models below. The best model on the Civil Comments test set (.965 AUC-ROC) was selected for evaluation. This baseline CNN model is used as a control to ascertain whether a distilled CNN has demonstrable improvements over a model without the benefits of teacher pre-training.

BERT A task-specific teacher model built from a BERT-base public checkpoint with 768 dimensions, 12 layers, 12 heads that was fine-tuned exclusively on the Civil Comments training data. The model used a batch size of 64, a learning rate of $1e-5$, max token length of 512 and Adam optimizer. The model was trained for 1M steps and the best performing checkpoint in terms of AUC-ROC was selected.

4.1.2 Big Data Models

Using a combination of publicly available datasets and our much larger proprietary datasets, we show the distillation bias effects in the toxicity space scale to big data. We start with a competitive teacher BERT model that is distilled using a compact CNN architecture. Both teacher and student incorporate the open-source Civil Comments training corpus as well as proprietary human-labeled data and bias mitigation data. We follow the best practices of data augmentation described in (Dixon et al., 2018) by including bias mitigation data to help mitigate discrepancies in identity subgroup metrics.

PROPRIETARYBERT A state-of-the-art BERT toxicity model that has been pre-trained on more than 1.5B user comments in English. This baseline was additionally fine-tuned on rater labeled comments. The model uses a custom sentence-piece vocabulary of size 200K. The teacher model is constructed with 768 dimensions, 12 layers, 12 heads, consistent with BERT-base (Devlin et al., 2019). The pre-training consists of MLM loss with uniform masking at 15%. Pretraining was conducted with batch size of 32 for over 100K steps. The model was fine-tuned on 3M user generated comments scored by raters for toxicity, bias mitigation data, and the Civil Comments training set with batch size of 512 until convergence.

4.2 Distilled Models

Several models are used to examine distillation. For reference, knowledge distillation is defined as training a smaller neural network on a dataset called the *transfer set*. Using cross entropy as the loss function between the output of the smaller distilled model $y(x|t)$ and the output of the teacher model $\hat{y}(x|t)$, where t is the temperature and for a standard softmax

$$E(x|t) = -\sum_i \hat{y}_i(x|t) \log y_i(\mathbf{x}|t)$$

is normally set to 1.

DISTILLEDCNN The transfer data, scored by the above BERT model, is drawn from WikiConv (Hua et al., 2018), a corpus encompassing the history of conversations on Wikipedia Talk pages, and

C4 (Raffel et al., 2019), a cleaned version of Common Crawl's web crawl corpus. For both sources a large quantity of data was scored with BERT and then examples were dropped to ensure a 50/50 distribution of toxic and nontoxic examples using a 0.5 threshold. Since both sources are extremely non-toxic (0.004% and 0.00005% respectively), this process produced only 400k examples from WikiConv and 640k from C4.

DistilledCNNOnProprietary CNN model distilled on a much larger volume of unsupervised user comments as the transfer set labeled by BERT. As with DistilledCNN, the architecture and training parameters replicate those used by CNN. The model was trained on the Civil Comments golden data and 20M teacher-labeled comments, including proprietary comments.

DistilledSmBERTOnProprietary Small BERT model distilled on the same larger volume of unsupervised corpus of user-domain comments as DistilledCNNOnProprietary by using BERT as teacher. As with DistilledCNNOnProprietary the model uses Civil Comments golden data and 20M teacher-labeled comments from a proprietary dataset. The model is included to ascertain whether Small BERT for distillation yields improvements in bias over a CNN.

DistilledProprietaryCNN A CNN student model distilled on 28M user comments scored with ProprietaryBERT. The model is also trained on the the same golden data as the teacher model. In addition, the model training data also includes 1.7M bias mitigation examples added to the golden data to mitigate identity term bias. The model uses the same tokenizer as the teacher model and is initialized from the teacher word embeddings. The CNN is 5 layers: one layer of 300 bi-grams, one layer of 300 tri-grams, one layer of 300 quad-grams, one layer of 300 5-grams, one layer of 300 6-grams and a max pool of the entire sequence. The model is trained with an Adam optimizer (Kingma and Ba, 2017), learning rate of .1, a batch size of 128 and a maximum token sequence length of 1536 until convergence.

The distilled student model DistilledProprietaryCNN achieves equivalent (if slightly better performance) to the teacher model ProprietaryBERT on the Civil Comments test set, as shown in Table 3. The *Short Synthetic* test set is used to measure bias, as shown in Table 3, and further illustrates the similar performance of the two models.

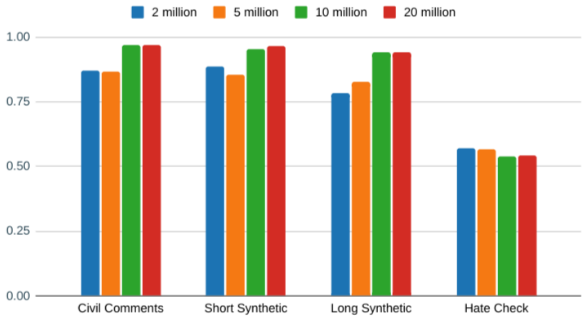

Figure 2: AUC-ROC performance of the BERT model distilled on proprietary data and evaluated on various test sets, broken down by distilled train set size.

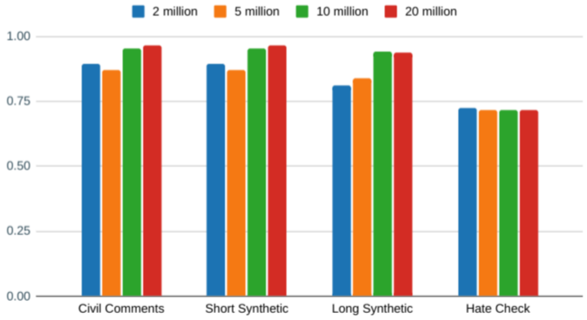

Figure 3: AUC-PR performance of the BERT model distilled on proprietary data and evaluated on various test sets, broken down by distilled train set size.

5 Evaluating Performance and Bias

Experiments are run on a variety of evaluation sets to assess the classification performance of the teacher, baseline and distilled models. In assessing both the Civil Comments based models and the big data models, we compare the distilled student and baseline models performance against the teacher models. Results are shown in Table 2 (Civil Comments based models) and Table 3 (big data models). The final column in each of these tables shows the difference in AUC-ROC between the student model and the teacher.

Civil Comments The test set from Civil Comments, drawn from the same distribution of comments as the training data, and is similar to the data distribution contained in the big data datasets.

Given the matched distribution between training and test, we expect this to be a best case result. All of the Civil Comments-based distilled and baseline models are within $\sim 1\%$ of BERT AUC-ROC).

In the big data case, in fact DistilledProprietaryCNN yields better performance than ProprietaryBERT in Table 3. These results show the strong promise of distillation, which leverages unsupervised data and produces an improvement without additional model complexity.

Short Synthetic A synthetic test set created by substituting identity terms into toxic and non-toxic sentence templates (Dixon et al., 2018; Borkan et al., 2019).

The performance of DistilledCNN and DistilledCNNOnProprietary along with CNN begins to degrade (-3.5%) with respect to the teacher model BERT on this dataset. This yields some evidence that the distillation process, when used with CNN architectures, may increase identity term bias.

On the other hand, minimal degradation in performance occurred for DistilledProprietaryCNN where carefully selected bias mitigation data was included as part of the teacher model training and distillation process.

Long Synthetic A dataset similar to Short Synthetic but with the addition of *random filler text* meant to be more confusing.

This more challenging dataset begins to show degradation for the DistilledProprietaryCNN model, despite the addition of bias mitigation data. Table 3 shows almost a -5% fall in AUC-ROC performance with respect to the teacher ProprietaryBERT.

Likewise, larger drops in performance can be seen for the Civil Comments-based models in table 2. Interestingly, DistilledCNNOnProprietary starts to slightly outperform the baseline CNN and DistilledCNN with only a -4% drop in AUC versus $-6\%+$.

Hate Check A targeted diagnostic test for hate detection models from Röttger et al. (2021). This dataset explicitly attempts to probe the generalisability of a model, measuring systemic gaps and biases in other datasets using a suite of synthetically generated tests.

While the big data teacher model ProprietaryBERT begins to show slightly more robust performance than the smaller BERT model (.831 AUC vs .701), all distilled and baseline CNN models suffer significant falls in performance. DistilledProprietaryCNN has nearly a -17% fall in AUC to .664. Both DistilledCNN and DistilledCNNOnProprietary models have $\sim 10\%$ or greater falls in AUC to (.575 and .595 respectively).

Examining the Hate Check functionalities, the categories with the largest differences where the teacher model outperforms the student model are in the non-hate comments that contain a negative term with negation (F14), followed by the comments that have a character swap (F25), and implicit derogation (F4). The teacher model, however, did not perform as well on abuse targeted against a non-protected object or individual (F22, F23). In 22 of the 29 categories, the student model performed worse than the teacher.

We continue our testing with a suite of more robust tests that demonstrate the limitations and weak-points in the distilled model versions.

False Positives A dataset inspired and derived from the work of Welbl et al. (2021), where authors trained a generative LM specifically to not produce toxic content. This dataset includes the sentences generated that had a large discrepancy in score between the publicly available toxicity model, Perspective API (Jigsaw, 2017), and human raters. Human annotations marked far fewer examples as toxic than the automated models, and the authors note a strong bias towards false positives in this set.

The False Positives dataset includes 50% auto-generated texts that had Perspective API scores $>$ .75 but were marked by human raters as non-toxic and the rest as randomly selected auto-generated comments with corresponding human annotations.

Notably all models perform poorly on the challenging dataset with ProprietaryBERT and BERT yielding only .635 and .651 AUC-ROC respectively. However all distilled CNN models faired even worse when compared to the teacher models, varying between -11% and -15%.

Identity Swaps Inspired by the work in Prabhakaran et al. (2019), where Perturbation Sensitivity Analysis is used to detect unintended model bias related to named entities, we repeat a similar experiment in relation to curated swapped identity terms. A small subset of curated phrases with explicit identity terms meant to detect *hard* toxic and non-toxic instances. The phrases each have 23 identity terms which are swapped with correct associated grammar specifications. Examples from this data set appear in Table 8. The identity swaps sets shows similar drops in performance for all distilled model instances as compared to the teacher.

Covert Toxicity Detecting implicit abuse or covert toxicity, where clearly hateful or abusive words are not used in the comment, presents an especially hard challenge. Given the documented difficulty of toxicity models and hate models to identify such text, we included a representative set as a further baseline. Using a published test dataset (Lees et al., 2021) we select an output label that is defined as the max of the covert and overt toxic scores. Notably all models performed extremely poorly on this set with < .6 AUC. The effects of distillation were more mixed, suggesting that identifying covert toxicity or implicit abuse is a more nuanced and unsolved task and perhaps more reliant on training data.

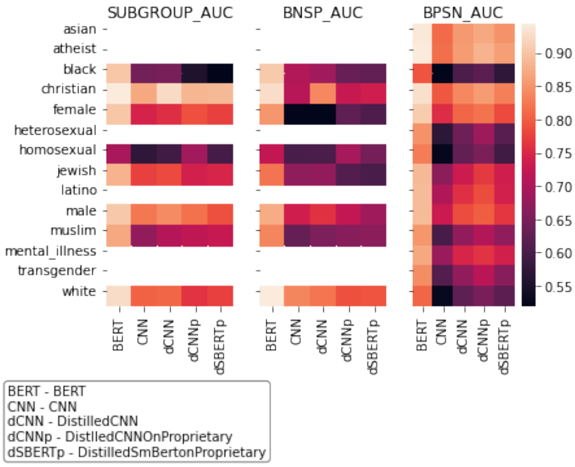

Figure 4: Civil Comments Bias Metric Breakdowns for Identity Subtypes on Civil Comments-based Models

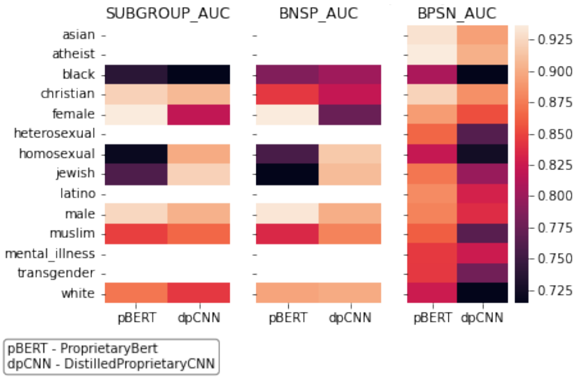

Figure 5: Civil Comments Eval Set Bias Metric Breakdowns for Identity Subtypes on Proprietary Big Data Models with bias mitigation implemented

6 Bias in Distilled Models

For evaluation of model bias, we employ a subset of the suite of metrics introduced in Borkan et al. (2019). In particular, we utilize the following metrics for identifying unintended bias along with averaging the differences in these metrics across a subsection of identity categories:

Subgroup AUC The AUC computed only for the data labeled as including a mention of a particular identity

Background Positive, Subgroup Negative AUC BPSN AUC is computed for a split dataset of positive background data and negative examples for a particular subgroup. Lower metrics for this particular category suggest that a particular identity is linked to a high false positive rate, which could imply that specific identities are associated with toxicity, independent of context.

Background Negative, Subgroup Positive AUC BNSP AUC is computed for a split dataset of negative background data and positive subgroup examples.

6.1 Civil Comments Identities Bias

Civil Comments Identities subset includes rater labeled categories for subgroup identities. The overall bias metrics for the Civil Comments-based models in Figure 6 show a notable discrepancy between the teacher BERT style model BERT and baseline and distilled versions of the models. Also, a drop in overall performance for BPSN, suggesting strong links between the presence of any identity subtype and a false positive value.

Figure 4 shows subgroup bias metric breakdowns for individual subgroups. The missing subgroup metrics are due to insufficient data to accurately assess the subgroup positive performance. Outside of the wide discrepancy between the teacher BERT model and the distilled CNNs, certain identity categories perform far worse than others such as *black* and *homosexual*.

On the other hand, DISTILLEDPROPRIETARYCNN, which contains explicit bias mitigating data, does not show the same overall average bias metric degradation for subgroup AUC and BNSP AUC. However, there is a fall in performance for average BPSN, suggesting, despite the existence of bias mitigation data, some identity groups are linked with false positives (see Figure 7). Figure 4 better illustrates the identity subgroup breakdowns. The distilled student model DISTILLEDPROPRIETARYCNN shows a uniform drop in performance for BPSN

Dataset	Model Type	Model	Params	AUC-PR	AUC-ROC	Teacher AUC-ROC Diff
Civil Comments	BERT Teacher	BERT	110M	**.815**	**.981**	0
	Distilled Student	DistilledCNN	8M	.755	.970	-.011
		DistilledCNNOnProprietary	8M	.757	.971	-.010
		DistilledSmBERTOnProprietary	NA	.702	.958	-.023
	Baseline	CNN	8M	.738	.965	-.016
Short Synthetic	BERT Teacher	BERT	110M	**.997**	**.997**	0
	Distilled Student	DistilledCNN	8M	.952	.955	-.042
		DistilledCNNOnProprietary	8M	.961	.961	-.036
		DistilledSmBERTOnProprietary	NA	.936	.936	-.061
	Baseline	CNN	8M	.956	.961	-.036
Long Synthetic	BERT Teacher	BERT	110M	**.984**	**.983**	0
	Distilled Student	DistilledCNN	8M	.911	.916	-.067
		DistilledCNNOnProprietary	8M	.938	.943	-.040
		DistilledSmBERTOnProprietary	NA	.915	.913	-.070
	Baseline	CNN	8M	.915	.923	-.060
Hate Check	BERT Teacher	BERT	110M	**.813**	**.701**	0
	Distilled Student	DistilledCNN	8M	.712	.575	-.126
		DistilledCNNOnProprietary	8M	.715	.595	-.106
		DistilledSmBERTOnProprietary	NA	.706	.531	-.170
	Baseline	CNN	8M	.731	.560	-.141
False Positives	BERT Teacher	BERT	110M	**.103**	**.651**	0
	Distilled Student	DistilledCNN	8M	.061	.500	-.151
		DistilledCNNOnProprietary	8M	.074	.547	-.104
		DistilledSmBERTOnProprietary	NA	.065	.532	-.119
	Baseline	CNN	8M	.07	.538	-.113
Identity Swaps	BERT Teacher	BERT	110M	**.321**	**.892**	0
	Distilled Student	DistilledCNN	8M	.360	.754	-.138
		DistilledCNNOnProprietary	8M	.346	.791	-.101
		DistilledSmBERTOnProprietary	NA	.356	.760	-.132
	Baseline	CNN	8M	.354	.774	-.118
Covert Toxicity	BERT Teacher	BERT	110M	**.130**	**.586**	0
	Distilled Student	DistilledCNN	8M	.128	.585	-.001
		DistilledCNNOnProprietary	8M	.127	.562	-.024
		DistilledSmBERTOnProprietary	NA	.117	.564	-.022
	Baseline	CNN	8M	.126	.568	-.018

Table 2: Evaluation Results for Civil Comments based models: **BERT** - BERT model trained on Civil Comments, **CNN** - CNN trained on Civil Comments, **DistilledCNN** - CNN distilled from BERT on 2M comments(reproducible) **DistilledCNNOnProprietary** - CNN distilled from BERT on 20M proprietary comments, **DistilledSmBERTOnProprietary** - Small Bert model distilled from BERT on 20M proprietary comments

AUC metrics (false positives for identity terms) when compared to ProprietaryBERT. However, certain subgroups such as *jewish* and *homosexual* have worse subgroup and BNSP AUC performance for the teacher model, where the abundance of bias mitigation data may be compromising the model's toxicity sensitivity

7 Effect of Distilled Data Size

Another variable to consider is the size of the distilled transfer data used for training. For these experiments we use variable-sized subsets of the data used by DistilledCNNOnProprietary above. This data matches the distribution of toxic comments found in Civil Comments, but is not publicly available.

In this experiment we consider the effect of increasing the ratio of the size of the transfer dataset to the size of the golden human-labeled data. We find in Figure 2 and Figure 3 that more distilled transfer data increases performance but only to a certain point. Increasing the distilled data size beyond 10M comments had little effect.

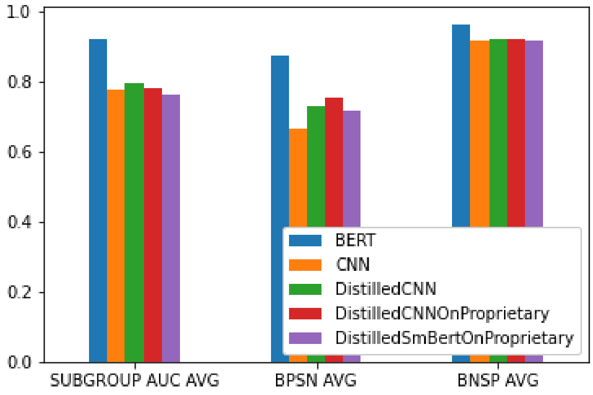

Figure 6: Civil Comments Eval Set Average Bias Metrics for Reproducible Models

8 Conclusion

The experimental section illustrates for both our more robust big data toxicity models and smaller reproducible versions that we are able to distill large transformer teacher models into smaller student models with very similar metrics on the evaluation datasets drawn from the same distribution. However, upon closer inspection, the distilled models consistently under-perform and even demon-

Dataset	Model Type	Model	Params	AUC-PR	AUC-ROC	Teacher AUC-ROC Diff
Civil Comments	BERT Teacher	ProprietaryBERT	235M	.734	.969	0
	Distilled Student	DistilledProprietaryCNN	23M	.771	.974	+.005
Short Synthetic	BERT Teacher	ProprietaryBERT	235M	.993	.994	0
	Distilled Student	DistilledProprietaryCNN	23M	.992	.993	-.001
Long Synthetic	BERT Teacher	ProprietaryBERT	235M	.997	.996	0
	Distilled Student	DistilledProprietaryCNN	23M	.958	.949	-.047
Hate Check	BERT Teacher	ProprietaryBERT	235M	.898	.831	0
	Distilled Student	DistilledProprietaryCNN	23M	.791	.664	-.167
False Positives	BERT Teacher	ProprietaryBERT	235M	.113	.635	0
	Distilled Student	DistilledProprietaryCNN	23M	.061	.520	-.115
Identity Swaps	BERT Teacher	ProprietaryBERT	235M	.311	.954	0
	Distilled Student	DistilledProprietaryCNN	23M	.329	.854	-.100
Covert Toxicity	BERT Teacher	ProprietaryBERT	235M	.117	.538	0
	Distilled Student	DistilledProprietaryCNN	23M	.131	.583	+.045

Table 3: Big Data Model Evaluation Results

id	text	label	ProprietaryBERT	DistilledProprietaryCNN	Diff
1	Never trust a single word said by a gay	true	.847 (TP)	.574 (TP)	-.273
2	Never trust a single word said by a married man	true	.321 (FN)	.139 (FN)	-.182
3	Never trust a single word said by a black	true	.794 (TP)	.347 (FN)	-.447
4	I do not know many blacks	false	.162 (TN)	.610 (FP)	+.448
5	I do not know many immigrants	false	.017 (TN)	.250 (TN)	+.233
6	I wish I could date a gay	false	.158 (TN)	.563 (FP)	+.334
7	I wish I could date a married man	false	.019 (TN)	.040 (TN)	+.021
8	asexuals being in charge are a danger to our country	true	.718 (TP)	.313 (FN)	-.405
9	lesbians being in charge are a danger to our country	true	.818 (TP)	.640 (TP)	-.178
10	Three unmarried men were elected to parliament today.	false	.032 (TN)	.031 (TN)	-.001

Table 4: Identity Swaps Toxicity Evaluation Results (TP, FP, TN, FN assuming a .5 threshold for toxicity label)

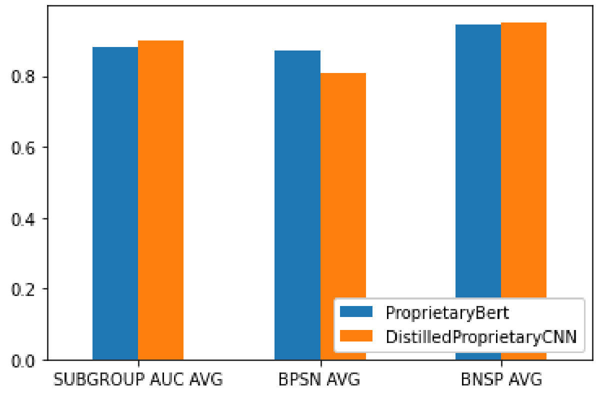

Figure 7: Civil Comments Eval Set Average Bias Metrics for Proprietary Models with bias mitigation

strate serious weakness when examined on larger and more difficult suites of test sets. In particular, identity-based bias for the toxicity models is noticeably worse in the distilled model versions, even with the addition of significant quantities bias-mitigating data. Table 4 shows specific examples with high discrepancy of score between the teacher and student models for both True/False toxicity labels from the curated Identity Swaps set. Even distilled models are complex, so we do not have a systemic way to characterize what's different between the teacher and the student models. But our analysis suggests that the student models are emphasizing lexical features.

Balancing costs versus performance is an unavoidable part of building machine learning systems. Much of the work within the academic community presents techniques that bring marginal improvements often at much higher costs. The popularity of ensemble models in machine learning competitions is but one example of such a technique that is usually impractical in production settings.

In our own work, we became interested in distillation because it allowed us to maintain our existing architecture and serving costs, but allowed us to improve our models to what seemed like performance parity with the promising new BERT models.

We quickly noticed that distilled models performed worse, consistently, in our bias metrics. While the technique of data augmentation has helped us mitigate these biases, that technique has proven to be less effective in distillation settings.

In trying to tackle biases, whether caused by sampling methods, the annotators, or the models themselves, there are always other potential biases that we are not yet measuring. For these reasons we have concluded that there may be subtle and intangible benefits to using large models. Importantly for us, data augmentation techniques for bias mitigation perform better with transformer models, at least to the limits of our ability to measure. While distillation seemingly lifts student model performance to new heights of accuracy, it may be a pale imitation of the often profound context sensitive classifications that are produced by the teacher models. We hope that this caution and advice with help other practitioners who face similar choices.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, William L. Hamilton, and Jimmy Lin. 2020. Exploring the limits of simple learners in knowledge distillation for document classification with DocBERT. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 72–77, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Yew Ken Chia, Sam Witteveen, and Martin Andrews. 2019. Transformer to cnn: Label-scarce distillation for efficient text classification.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

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.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678, Dublin, Ireland. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffrey Sorensen, and Lucas Dixon. 2018. WikiConv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2818–2823, Brussels, Belgium. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Google Jigsaw. 2017. Perspective api. `https://www.perspectiveapi.com/`. Accessed: 2021-02-02.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Alyssa Lees, Daniel Borkan, Ian Kivlichan, Jorge Nario, and Tesh Goyal. 2021. Capturing covertly toxic speech via crowdsourcing. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 14–20, Online. Association for Computational Linguistics.

Bansidhar Mangalwedhekar. 2021. Distilling bert for low complexity network training.

Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3).

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. *CoRR*, abs/1910.04210.

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

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

Moshe Wasserblat, Oren Pereg, and Peter Izsak. 2020. Exploring the boundaries of low-resource BERT distillation. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 35–40, Online. Association for Computational Linguistics.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Guangxuan Xu and Qingyuan Hu. 2022. Can model compression improve nlp fairness.