Transfer Learning for Offensive Language Detection

Michael Zeng
School of Information
University of California, Berkeley
zengm71@berkeley.edu

Yekun Wang
School of Information
University of California, Berkeley
myekun.wang@berkeley.edu

Abstract

We present a methodology for building an offensive language classifier and evaluating its zero-shot as well as few-shots performance on an unseen but related task set. We adapt and fine-tune the pre-trained contextual embeddings and weights from various state-of-the-art models including BERT (both base and large), multi-lingual transformers such as XLM and XLM RoBERTa large, as well as encoder-decoder framework such as T5. We fine-tune these models on Jigsaw Multilingual Toxic Comment Classification dataset, and use the Offensive Language Identification (OLID) dataset for transfer learning evaluation. Our best model, XLM RoBERTa large, achieves a 93.27% AUC score against the private Jigsaw test set (16% improvement on our base model BERT-base), and 0.6942 F1 score with zero-shot learning against the OLID dataset.

1 Introduction

Combining the right to freedom of speech and the protection of online anonymity, the issue of offensive language or language toxicity has become increasingly more pervasive among online communities, penetrating many of the large social media platforms. The widespread use of toxic language is not only harmful for user experience and retention, it is also linked to online harassment and bullying incidents, which could deeply impale the victims mentally. Organizations currently employ various semi-automated content moderation solutions which combine automated offensive language tagging along with human validation to police and to protect their user base.

Generally speaking, offensive language is defined as anything rude, disrespectful or otherwise likely to make someone leave a discussion. In the past, researchers have studied various neural and non-neural machine learning approaches to identify language toxicity (Schmidt and Wiegand, 2017). More recently, researches have experimented with deep neural networks models for classifying language toxicity.

Recently, transformer-based models (Vaswani et al., 2017) have been widely used in various areas of NLP research and applications. Depending on model specifications, training this kind of model could require estimating hundreds of millions to billions of parameters. Due to their massive size, many NLP practitioners prefer to build upon a pre-trained transformer model by stacking an extra layer on top, which adapts the existing model for task-specific purposes. This paper aims to evaluate the various transformer-based models, such as BERT-base and BERT-large (Devlin et al., 2018), XLM (Lample and Conneau, 2019), XLM RoBERTa (Conneau et al., 2019), T5 (Raffel et al., 2019); fine-tuning them to detect offensive language on the Jigsaw Multilingual Toxic Comment Classification dataset[1].

With the emergence these transformer-based model, we witnessed an significant increase in the need of large datasets for these massive models to train on, while labeled dataset are scarce and costly to scale. Another goal of this study is to demonstrate the effectiveness of sequential transfer learning techniques in resolving the practical issue of labeled data scarcity. The general idea behind transfer learning is to take parameters or knowledge from one trained model and apply it to another (Ruder et al., 2019). In this study, we experiment with both zero-shot and few-shot learning by applying the offensive language classifier trained on the Jigsaw dataset and evaluate the model's performance against the OLID dataset (Zampieri et al., 2019).

This paper's contribution to this area are three-

[1] https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data

fold:

- We adapt and compare various pre-trained state-of-the-art deep neural network models to the task of offensive language identification.
- We demonstrate zero-shot and few-shot transfer learning techniques by applying the fine-tuned offensive language classifier to a new dataset and evaluate its general performance.
- We show that most models see diminishing returns as more target dataset is exposed, and reasonable transfer learning experience could be achieved with limited data.

The paper is organized as follows. In section 1, we discuss the existing work related to offensive language identification using various NLP techniques. In section 2, we provide details of the transformer models employed in our experiments. In section 3, we discuss the show results of fine-tuning various transformer models on the Jigsaw dataset. In section 4, we discuss and show results for zero-shot and few-shot learning of applying the fine-tuned toxicity model to the OLID dataset. In section 5, we provide a summary and discuss future research directions.

1.1 Existing Work

Earlier research looked into feature-based language models for identifying online bullying and hate speech (Chen et al., 2012) and (Huang et al., 2014). More recently, (Zampieri et al., 2019) compiled the OLID dataset and applied SVM, CNN, BiLSTM models to predicting the language toxicity label. (Kohli et al., 2017) demonstrated a deep learning approach using word-level and character level RNNs with custom embeddings on the Jigsaw dataset. Additionally, (Ruder et al., 2019) outlined the landscape of various transfer learning techniques pertaining to NLP tasks.

1.2 Data

First part of this study focuses on the Jigsaw Multilingual Toxic Comment Classification dataset, published as a part of a Kaggle competition[2]. The data consists of English comments from Wikipedia's talk page edits as well as an expanded version of the Civil Comments dataset.

[2]https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data

DATASET	Jigsaw	OLID
TRAIN	485,775 133,610 positive English Only	10,000 1,231 positive English Only
DEV	8,000 1,230 positive Multi-lingual	4,200 1,105 positive English Only
TEST	63,812 unknown positive Multi-lingual	860 240 positive English Only

Table 1: Dataset Summary

Due to the limit on resources, we constructed our Jigsaw training set by taking the entire 2018 dataset, and combined that with all offensive tweets and down-sampled non-offensive ones from the 2019 dataset. The constructed dataset has close to 500K observations with around 133k being offensive, with a targeted positive rate of 30%.

Interestingly, Jigsaw's training dataset is English-only, but the test dataset contains other languages. The organizer intended it this way to validate zero-shot transfer learning of models trained on English and apply them to other languages. We take this idea of transfer learning one step further and focus the second part of the paper on validating zero-shot and few-shot transfer learning on the OLID Dataset(Zampieri et al., 2019).

The OLID Dataset is provided by the International Workshop on Semantic Evaluation, and 2020 is the second year this task is hosted. This paper focuses on the Offensive Language Identification Dataset (OLID) from the 2019 competition, which contains a collection of 14,200 annotated English tweets.

To compare our model performance against other participants' submissions, we used the same evaluation metrics mandated by the competition to train our models – AUC for the Jigsaw dataset and F1 score for the OLID dataset. Table 1 gives the detailed breakdown of dataset used for experiments.

1.3 Infrastructure

We choose the Kaggle notebook as the main working environment, because Kaggle allows 30 hours of TPU V3 and V100 GPU time per person per week, an incredible resource for free access to the state-of-the-art hardware. In comparison, the TPU V2 offered by Google Colab Pro, which only has half the memory of V3, sometimes fail to fit any

training example at all after loading a large model. Comparing TPU vs GPU, we find that TPU offers a significant boost to the training time that's at least 3-5 times shorter than training on V100.

2 Models

In this study, we experimented with fine-tuning five transformer models to the language toxicity classification task. In the fine-tuning phase, the model is initialized with the pre-trained parameters and then is fine-tuned on the labelled Jigsaw dataset. All models used in this paper are implemented in HuggingFace (Wolf et al., 2019), an open-source library for many state-of-the-art transformer architectures under a unified API. Three suits of models are provided: (i) BERT and BERT large, which are faithful implementations of the original BERT models; (ii) multi-language BERT such as XLM and XLM-RoBERTa; (iii) encoder-decoder framework such as T5.

2.1 BERT

BERT, short for Bidirectional Encoder Representations from Transformers, is a bidirectional transformer pre-trained on English language using a combination of masked language modeling (MLM) and next sentence prediction (NSP). We intended to leverage the CLS tokens from the raw model and specify it to the downstream toxicity classification task. We picked the case-sensitive models based on the observation that upper-casing generally imply strong emotions in tweets, which should help with our objective of offensive language detection. We included both BERT-base as well as BERT-large to gauge the improvement in performance against the size of the model. Both model implementations, as published on HuggingFace, had been pre-trained on the BookCorpus and English Wikipedia data.

2.2 Multi-language BERT

The work on cross-lingual language models (XLMs) (Lample and Conneau, 2019) extends the generative pretraining for English language understanding to multiple languages and shows the effectiveness of cross-lingual pretraining. Another variation we picked in our experiment is XLM-RoBERTa model (Conneau et al., 2019), based on Facebook's RoBERTa model released in 2019, and trained on 2.5TB of filtered CommonCrawl data.

2.3 T5

Asides from BERT or variations of BERT that build on the encoder part of the transformer architecture, we also experimented with T5, an encoder-decoder model pre-trained on a variety of tasks that had been converted into text-to-text format. The training of T5 differs from the BERT based model: we fed the model with token and labels pairs directly, instead of extracting the CLS tokens and stack softmax layer on top for classification. As detailed in (Raffel et al., 2019), the T5 model was pre-trained on 750GB of natural English text, obtained by web crawling.

3 Jigsaw Performance

The various pre-trained transformer models were then fine-tuned on the Jigsaw Multilingual Toxic Comment Classification dataset for the language toxicity task. For the Jigsaw dataset, we trained each model on the training set for 3 epochs with 1e-5 learning rate. Since the training set is English only, we then trained each model for 3 additional epochs on the dev set which contains non-English language. The AUC scores are reported based on the public and private Kaggle test set.

Traditionally, people add a sigmoid affine layer on top of the CLS tokens to fine tune these transformer models. Depending on constraints on memory, some choose to keep the transformer weights frozen or partially frozen, as demonstrated by (Lee et al., 2019) to have limited reduction in performance. Gaining access to the TPU units allowed us to experiment with freezing or unfreezing the transformer weights and to evaluate various models more comprehensively both in terms of model performance as well as run time. We report the models' performance on the Jigsaw public and private dataset in Table 2, using AUC as the metric to be consistent with the competition.

We see an intuitive improvement in performance from unfreezing the transformer weights. By unfreezing the transformer weights, we are allowing many more degrees of freedom while training the model. Yet we see the increase in model power generalizes very well on the test set. On the other hand, unfreezing the transformer weights consumes more memory on the TPU unit, and thus allows lower batch sizes. Yet from our experiment, the training time stayed largely the same, which prompts us to leave the transformer weights unfrozen for the rest of the experiments.

MODEL	Freeze Transformer	Jigsaw Public Score (AUC)	Jigsaw Private Score (AUC)	Training Time (s)
BERT Base	No	0.8037	0.8026	1,452
	Yes	0.8044	0.8011	1,413
BERT Large	No	0.8184	0.8167	3,300
	Yes	0.8089	0.8041	3,423
XLM	No	0.9148	0.9121	2,859
	Yes	0.9096	0.9094	2,901
XLM RoBERTa Large	No	**0.9335**	**0.9327**	5,205
	Yes	0.9306	0.9315	5,115
T5 Large	No	0.8392	0.8373	8,844
	Yes	0.8371	0.8352	8,925

Table 2: Performance on Jigsaw Dataset and Training Time for 3 Epochs

Comparing across models, the gain from BERT to BERT large is only marginal despite the significant size-up and increase in training time. The big leap in performance came from the switching from English to multi-language models such as XLM and XLM RoBERTa large. Given the fact that both the public and private test sets are multilingual, the English only embedding we used in BERT Base, BERT Large as well as T5 large is limiting the model performance. The improvement from XLM to XLM RoBERTa large could be mostly attributed to the pretraining of XLM Roberta Large, where the model was exposed to and trained on significantly larger corpus.

4 Transfer Learning

In this section, we started by exploring the differences in zero-shot transfer learning performances from multiple models on the OLID dataset. It served as a reasonable baseline for the model's ability to generalize to unseen data, given that the model was previously fine-tuned on a very related task. We then explored the performances in few-shot learning by incrementally increase the portion of the OLID data for the model to be trained on. Given that we had 10K in total from the OLID training set, the choice of exposure was set to follow the geometric sequence: 1K, 2K, 5K and 10K. Since these chosen sizes are still small, we found training loss to be quite jumpy across epoches. As a result, for this section of training, we lowered the learning rate to 2.00E-06 and trained up to 10 epochs with a tight early-stop that checks the loss on dev set every 3 epochs. The hypothesis is that since the model has been pretrained on similar tasks, one would only need to expose a fraction of the dataset to obtain a high level of performance.

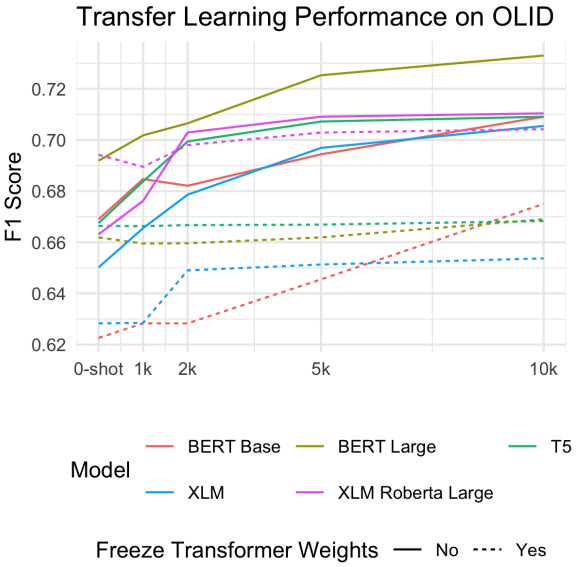

Figure 1: Performance of Transfer Learning on OLID Dataset

Table 3 summarises the F1 score on the OLID test set from various experiments, which are further visualized in Figure 1. Our observations on the choice of freezing the transformer weights from the previous section still hold: most models perform better when trained with weights from transformer layer unfrozen. When they are frozen, the performance curves stayed mostly flat as more destination dataset is exposed. This indicates a lack of flexibility to adapt and learn from the new dataset with only the last layer being trainable.

Across the different models, we found that BERT-large consistently outperforms the others across the zero-shot and few-shot experiments. We no longer observe performances leaps from multi-

Model	Freeze	OLID 0-shot	OLID 1k	OLID 2k	OLID 5k	OLID 10k (Full)
BERT Base	No	0.6690	0.6847	0.6821	0.6944	0.7091
	Yes	0.6226	0.6283	0.6283	0.6455	0.6750
BERT Large	No	0.6919	**0.7018**	**0.7065**	**0.7253**	**0.7330**
	Yes	0.6619	0.6595	0.6596	0.6619	0.6690
XLM	No	0.6502	0.6655	0.6786	0.6969	0.7055
	Yes	0.6283	0.6285	0.6490	0.6513	0.6537
XLM RoBERTa Large	No	0.6631	0.6762	0.7029	0.7091	0.7104
	Yes	**0.6942**	0.6895	0.6980	0.7029	0.7042
T5	No	0.6675	0.6838	0.6994	0.7072	0.7091
	Yes	0.6664	0.6663	0.6667	0.6669	0.6683

Table 3: Performance of Transfer Learning on OLID Dataset

lingual model such as XLM and XLM RoBERTa Large since the OLID test set is English only.

As more observations are exposed to the model, we observe the performances continue to improve but at diminshing rates, as shown in Figure 1. For most models trained with weights from the transformation layer unfrozen, exposing the entire training set helps achieve about a 4% gain in F1 score compared to doing zero-shot inference, yet most of the gain was achieved from the first 5k observations. This is similar to the findings in (Cer et al., 2018), where the author demonstrated that the performances would plateau as more labeled examples are exposed to the model. However, we are still trying to procure more labeled data on offensive language detection so that we can expand our experiment beyond 10k observations to observe the curve over large ranges.

5 Conclusion

We explored the both zero and few-shots learning performances across several popular transformer based models within the common task of offensive language detection. We benchmarked the performance of sevreal state-of-the-art models on the Jigsaw dataset, and explored the choice of freezing the original model weights. Our experiments show unfreezing the weights from transformer layer, if resources allow, greatly boosts the model's ability to generalize and perform on unseen dataset. In terms of few-shots learning, we observed diminishing returns from exposing the model to additional training examples, as cited in (Cer et al., 2018). However, the performance of zero or few-shots learning are also largely limited by the language in which the model was originally pretrained on: XLM RoBERTa has superior performance on Jigsaw test set which is multi-lingual, while BERT performs the best on OLID test set which is English only.

References

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Y. Chen, Y. Zhou, S. Zhu, and H. Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Qianjia Huang, Vivek Singh, and Pradeep Atrey. 2014. Cyber bullying detection using social and textual analysis. *SAM 2014 - Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, Workshop of MM 2014*, pages 3–6.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning.

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.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.