Exploiting Auxiliary Data for Offensive Language Detection with Bidirectional Transformers

Sumer Singh
University of Georgia
Athens, GA, USA
sumer.singh@uga.edu

Sheng Li
University of Georgia
Athens, GA, USA
sheng.li@uga.edu

Abstract

Offensive language detection (OLD) has received increasing attention due to its societal impact. Recent work shows that bidirectional transformer based methods obtain impressive performance on OLD. However, such methods usually rely on large-scale well-labeled OLD datasets for model training. To address the issue of data/label scarcity in OLD, in this paper, we propose a simple yet effective domain adaptation approach to train bidirectional transformers. Our approach introduces domain adaptation (DA) training procedures to ALBERT, such that it can effectively exploit auxiliary data from source domains to improve the OLD performance in a target domain. Experimental results on benchmark datasets show that our approach, ALBERT (DA), obtains the state-of-the-art performance in most cases. Particularly, our approach significantly benefits underrepresented and under-performing classes, with a significant improvement over ALBERT.

1 Introduction

In today's digital age, the amount of offensive and abusive content found online has reached unprecedented levels. Offensive content online has several detrimental effects on its victims, e.g., victims of cyberbullying are more likely to have lower self-esteem and suicidal thoughts (Vazsonyi et al., 2012). To reduce the impact of offensive online contents, the first step is to detect them in an accurate and timely fashion. Next, it is imperative to identify the type and target of offensive contents. Segregating by type is important, because some types of offensive content are more serious and harmful than other types, e.g., hate speech is illegal in many countries and can attract large fines and even prison sentences, while profanity is not that serious. To this end, offensive language detection (OLD) has been extensively studied in recent years, which is an active topic in natural language understanding.

Existing methods on OLD, such as (Davidson et al., 2017), mainly focus on detecting whether the content is offensive or not, but they can not identify the specific type and target of such content. Waseem and Hovy (2016) analyze a corpus of around 16k tweets for hate speech detection, make use of meta features (such as gender and location of the user), and employ a simple n-gram based model. Liu et al. (2019) evaluate the performance of some deep learning models, including BERT (Devlin et al., 2018), and achieve the state of the art results on a newly collected OLD dataset, OLID (Zampieri et al., 2019). Although promising progress on OLD has been observed in recent years, existing methods, especially the deep learning based ones, often rely on large-scale well-labeled data for model training. In practice, labeling offensive language data requires tremendous efforts, due to linguistic variety and human bias.

In this paper, we propose to tackle the challenging issue of data/label scarcity in offensive language detection, by designing a simple yet effective domain adaptation approach based on bidirectional transformers. Domain adaptation aims to enhance the model capacity for a target domain by exploiting auxiliary information from external data sources (i.e., source domains), especially when the data and labels in the target domain are insufficient (Pan and Yang, 2009; Wang and Deng, 2018; Lai et al., 2018; Li et al., 2017; Li and Fu, 2016; Zhu et al., 2021). In particular, we aim to identify not only if the content is offensive, but also the corresponding type and target. In our work, the offensive language identification dataset (OLID) (Zampieri et al., 2019) is considered as target domain, which contains a hierarchical multi-level structure of offensive contents.

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 1–5

An external large-scale dataset on toxic comment (ToxCom) classification is used as source domain. ALBERT (Lan et al., 2019) is used in our approach owing to its impressive performance on OLD. A set of training procedures are designed to achieve domain adaptation for the OLD task. In particular, as the external dataset is not labelled in the same format as the OLID dataset, we design a separate predictive layer that helps align two domains. Extensive empirical evaluations of our approach and baselines are conducted. The main contributions of our work are summarized as follows:

- We propose a simple domain adaptation approach based on bidirectional transformers for offensive language detection, which could effectively exploit useful information from auxiliary data sources.
- We conduct extensive evaluations on benchmark datasets, which demonstrate the remarkable performance of our approach on offensive language detection.

2 Related Work

In this section, we briefly review related work on offensive language detection and transformers.

Offensive Language Detection. Offensive language detection (OLD) has become an active research topic in recent years (Araujo De Souza and Da Costa Abreu, 2020). Nikolov and Radivchev (2019) experimented with a variety of models and observe promising results with BERT and SVC based models. Han et al. (2019) employed a GRU based RNN with 100 dimensional glove word embeddings (Pennington et al., 2014). Additionally, they develop a Modified Sentence Offensiveness Calculation (MSOC) model which makes use of a dictionary of offensive words. Liu et al. (2019) evaluated three models on the OLID dataset, including logistic regression, LSTM and BERT, and results show that BERT achieves the best performance. The concept of transfer learning mentioned in (Liu et al., 2019) is closely related to our work, since the BERT model is also pretrained on external text corpus. However, different from (Liu et al., 2019), our approach exploits external data that are closely related to the OLD task, and we propose a new training strategy for domain adaptation.

Transformers. Transformers (Vaswani et al., 2017) are developed to solve the issue of lack of parallelization faced by RNNs. In particular, Transformers calculate a score for each word with respect to every other word, in a parallel fashion. The score between two words signifies how related they are. Due to the parallelization, transformers train rapidly on modern day GPUs. Some representative Transformer-based architectures for language modeling include BERT (Devlin et al., 2018), XL-NET (Yang et al., 2019) and ALBERT (Lan et al., 2019). BERT employs the deep bidirectional transformers architecture for model pretraining and language understanding (Devlin et al., 2018). However, BERT usually ignores the dependency between the masked positions and thus there might be discrepancy between model pretraining and fine-tuning. XL-NET is proposed to address this issue, which is a generalized autoregressive pretraining method (Yang et al., 2019). Another issue of BERT is the intensive memory consumption during model training. Recently, some improved techniques such as ALBERT (Lan et al., 2019) are proposed to reduce the memory requirement of BERT and therefore increases the training speed. In this paper, we leverage the recent advances on Transformers and design a domain adaptation approach for the task of offensive language detection.

Table 1: Details of OLID dataset.

A	B	C	Training	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	—	524	27	551
NOT	—	—	8,840	620	9,460
ALL	—	—	13,240	860	14,100

3 Methodology

3.1 Preliminary

Target Domain. In this work, we focus on the offensive language detection task on the OLID dataset, which is considered as target domain. The OLID dataset consists of real-world tweets and has three interrelated subtasks/levels: (A) Detecting if a tweet is offensive (*OFF*) or not (*NOT*); (B) Detecting if *OFF* tweets are targeted (*TIN*) or untargeted (*UNT*) and; (C) Detecting if *TIN* tweets are targeted at an individual (*IND*), group (*GRP*) or miscellaneous entity (*OTH*). The details of OLID dataset are summarized in Table 1. The following strategies are used to preprocess the data. (1) *Hash-*

Table 2: Details of Toxcom Dataset.

Classification	# of instances
clean	143,346
toxic	15,294
obscene	8,449
insult	7,877
identity hate	1,405
severe toxic	1,595
threat	478

tag Segmentation. Hashtags are split up and the preceding hash symbol is removed. This is done using wordsegment[1]. (2) *Censored Word Conversion.* A mapping is created of offensive words and their commonly used censored forms. All the censored forms are converted to their uncensored forms. (3) *Emoji Substitution.* All emojis are converted to text using their corresponding language meaning. This is done using Emoji[2]. (4) *Class Weights.* The dataset is highly skewed at each level, thus a weighting scheme is used, as follows: Let the classes be $\{c_1, c_2, \cdots, c_k\}$ and number of samples in each class be $\{N_1, N_2, \cdots, N_k\}$, then class c_i is assigned a weight of $\frac{1}{N_i}$.

Source Domain. To assist the OLD task in target domain, we employ an external large-scale dataset on toxic comment (ToxCom) classification[3] as source domain. ToxCom consists of 6 different offensive classes. Samples that belong to none of the 6 classes are labelled as *clean*. The details of ToxCom dataset are shown in Table 2. The number of *clean* comments is disproportionately high and will lead to considerable training time. Thus, only 16,225 randomly sampled *clean* comments are employed.

3.2 Domain Adaptation Approach

We propose a simple yet effective domain adaptation approach to train an ALBERT model for offensive language detection, which fully exploits auxiliary information from source domain to assist the learning task in target domain. The effectiveness of using auxiliary text for language understanding has been discussed in literature (Rezayi et al., 2021).

Both the source and target domains contain rich information about offensive contents, which makes it possible to seek a shared feature space to facilitate the classification tasks. A naive solution is to combine source and target datasets and simply train a model on the merged dataset. This strategy, however, may lead to degraded model performance for two reasons. First, two datasets are labelled in different ways, so that they don't share the same label space. Second, the divergence of data distributions due to various data sources. In particular, the target domain contains tweets, while the source domain is collected from Wikipedia comments. The diverse data sources lead to a significant gap between two domains, and therefore simply merging data from two domains is not an effective solution.

To address these issues, we propose the following training procedures with three major steps. Let D_S denote the source data and D_T denote the target dataset. *First*, we pretrain the ALBERT model on the source domain (i.e., ToxCom dataset). The loss function of model training with source data is defined as:

$$\mathcal{L}_S = \underset{\Theta}{\operatorname{argmin}} \, \text{ALBERT}(D_S; \Theta) \qquad (1)$$

where L_s denotes the loss function, ALBERT$(\cdot)$ is the Transformer based ALBERT network, and Θ represents the model parameters. *Second*, we freeze all the layers and discard the final predictive layer. Since two datasets have different labels, the final predictive layer could not contribute to the task in target domain. *Third*, we reuse the frozen layers with a newly added predictive layer, and train the network on the target dataset. The loss function of model finetuning with target data is defined as:

$$\mathcal{L}_T = \underset{\hat{\Theta}}{\operatorname{argmin}} \, \text{ALBERT}(D_T; \Theta, \hat{\Theta}), \qquad (2)$$

where $\hat{\Theta}$ denotes the finetuned model parameters. There are several ways to treat the previously frozen layers in this step: (1) A feature extraction type approach in which all layers remain frozen; (2) A finetuning type approach in which all layers are finetuned; and (3) A combination of both in which some layers are finetuned while some are frozen. Finally, the updated model will be used to perform OLD task in the target domain.

Let L denote the number of layers (including the predictive layer), N_S denote the number of training samples in the source domain, and N_T denote the number of training samples in the target domain. K is the set of layers that remain frozen during training in the target domain.

[1] https://github.com/grantjenks/python-wordsegment

[2] https://github.com/carpedm20/emoji

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

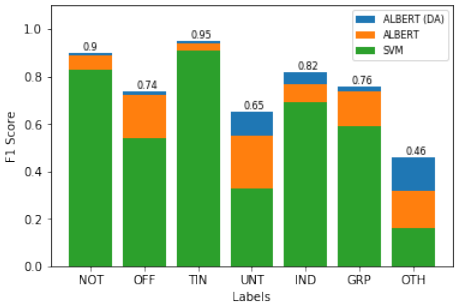

Figure 1: Classwise F1 Scores across three levels.

Table 3: Results. First three rows are previous state of the art results at each level.

Model	A	B	C
Liu et al. (2019)	**0.8286**	0.7159	0.5598
Han et al. (2019)	0.6899	0.7545	0.5149
Nikolov and Radivchev (2019)	0.8153	0.6674	0.6597
SVM	0.6896	0.6288	0.4831
CNN (UP)	0.7552	0.6732	0.4984
CNN	0.7875	0.7038	0.5185
CNN (CW)	0.8057	0.7348	0.5460
BERT	0.8023	0.7433	0.5936
ALBERT	0.8109	0.7560	0.6140
ALBERT (DA)	0.8241	**0.8108**	**0.6790**

4 Experiments

4.1 Baselines and Experimental Settings

In the experiments, four representative models are used as baselines, including the support vector machines (SVM), convolutional neural networks (CNN), BERT and ALBERT. We use the base version of BERT and the large version of ALBERT. The max sequence length is set to 32 and 64 for BERT and ALBERT, respectively. Training samples with length longer than max sequence length are discarded. Moreover, we compare our approach with three state-of-the-art methods (Liu et al., 2019; Han et al., 2019; Nikolov and Radivchev, 2019) on offensive language detection.

For domain adaptation, the finetuning and feature extraction approaches, discussed in Section 3.2, are tested. The feature extraction approach gives poor results on all three levels, with scores lower than ALBERT without domain adaptation. The third method is not used as it introduces a new hyperparameter, i.e., the number of trainable layers, which would have to be optimized with considerable computational costs. The finetuning type strategy gives good initial results and is used henceforth. The learning rate is set to 1.5×10^{-5} and 2×10^{-5} on the source data and target data, respectively. Following the standard evaluation protocol on the OLID dataset, the 9:1 training versus validation split is used. In each experiment (other than SVM), the models are trained for 3 epochs. The metric used here is macro F1 score, which is calculated by taking the unweighted average for all classes. Best performing models according to validation loss are saved and used for testing.

4.2 Results and Analysis

Table 3 shows the results of baselines and our domain adaptation approach, ALBERT (DA). For Task A, deep learning methods, including CNN, BERT and ALBERT, always outperform the classical classification method SVM. ALBERT achieves a macro F1 score of 0.8109, which is the highest score without domain adaptation. Task C is unique as it consists of three labels. All models suffer on the *OTH* class. This could be because the *OTH* class consists of very few training samples. Our approach, ALBERT (DA), achieves the state-of-the-art performance on Task C.

Figure 1 further breaks down the classwise scores for analysis. The most notable improvements are on *OTH* and *UNT* samples. ALBERT (DA) has an F1 score of 0.46, which is an improvement 43.75% over ALBERT on *OTH* samples. On *UNT* samples, ALBERT (DA) improves ALBERT's score of 0.55 to 0.65, which is an improvement of 18%. Conversely, performance on classes on which the ALBERT already has high F1-scores, such as *NOT* and *TIN*, do not see major improvements through domain adaptation. On *NOT* and *TIN* samples, ALBERT (DA) improves only 1.11% and 1.06% over ALBERT, respectively.

5 Conclusion

In this paper, we propose a simple yet effective domain adaptation approach to train bidirectional transformers for offensive language detection. Our approach effectively exploits external datasets that are relevant to offensive content classification to enhance the detection performance on a target dataset. Experimental results show that our approach, ALBERT (DA) obtains the state-of-the-art performance in most tasks, and it significantly benefits underrepresented and under-performing classes.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their insightful comments. This research is supported in part by the U.S. Army Research Office Award under Grant Number W911NF-21-1-0109.

References

Gabriel Araujo De Souza and Marjory Da Costa Abreu. 2020. Automatic offensive language detection from twitter data using machine learning and feature selection of metadata.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jiahui Han, Shengtan Wu, and Xinyu Liu. 2019. jhan014 at SemEval-2019 task 6: Identifying and categorizing offensive language in social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 652–656, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tuan Manh Lai, Trung Bui, Nedim Lipka, and Sheng Li. 2018. Supervised transfer learning for product information question answering. In *IEEE International Conference on Machine Learning and Applications*, pages 1109–1114.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Sheng Li and Yun Fu. 2016. Unsupervised transfer learning via low-rank coding for image clustering. In *International Joint Conference on Neural Networks*, pages 1795–1802.

Sheng Li, Kang Li, and Yun Fu. 2017. Self-taught low-rank coding for visual learning. *IEEE Transactions on Neural Networks and Learning Systems*, 29(3):645–656.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Saed Rezayi, Handong Zhao, Sungchul Kim, Ryan Rossi, Nedim Lipka, and Sheng Li. 2021. Edge: Enriching knowledge graph embeddings with external text. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2767–2776.

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.

Alexander T. Vazsonyi, Hana Machackova, Anna Sevcikova, David Smahel, and Alena Cerna. 2012. Cyberbullying in context: Direct and indirect effects by low self-control across 25 european countries. *European Journal of Developmental Psychology*, 9(2):210–227.

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

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 NAACL*.

Ronghang Zhu, Xiaodong Jiang, Jiasen Lu, and Sheng Li. 2021. Transferable feature learning on graphs across visual domains. In *IEEE International Conference on Multimedia and Expo*.