Can We Achieve More With Less?
Exploring Data Augmentation with Toxic Comment Classification

Chetanya Rastogi, Fang-I Hsiao, Nikka Mofid
{chetanya, fihsiao nmofid}@stanford.edu

1 Introduction

Data is the bottleneck of machine learning. There are millions of interesting problems in the world but, without enough data, high accuracy classifiers to solve these problems can not be built. In the converse, for problems where data is abundant labeling can take days. Thus, in this project we aim to tackle this data divide and explore if high accuracy classifiers can be built from small datasets using a combination of data augmentation techniques and machine learning methods. For added social impact, we will be working to develop a model to detect and classify toxic speech in comments to help web moderators fight back against online harassment and cyberbullying, and also protect data annotators from psychological stress of having to label an extremely large, graphic dataset.

2 Related Work

In the recent years, deep learning has seen some great advancements with the advent of transfer learning[1, 4], better architectures[8], and improved language models[2]. But for performing the basic task of classification, the size of "labelled" training data still dictates the performance of a model[6, 10]. Automatic data augmentation has been commonly used in computer vision[3] but due to the existence of strong local structure in the context of language it's a bit more challenging to come up with generalized rules for language transformation. In this project we aim to explore the data augmentation techniques proposed in the related works [9] (Easy Data Augmentation) and [5] (Back Translation) in tandem with different machine learning methodologies.

3 Methodology

3.1 Dataset

Our dataset is the "Wikipedia Toxic Comments" dataset[7]. The data contains a list of ∼158k Wikipedia comments and six binary labels for the kind of hate speech each comment qualifies as. For our task, we take any kind of toxicity as a positive class and rest of the comments as the negative class. Furthermore, we split the data into train and test and keep the test set aside which will only be used for evaluation purposes. To evaluate the performance of data augmentation, we further sample 5% of the data from the train set which serves as the small training set for our baseline(without data augmentation) and the entire train set serves as the oracle.

3.2 Inputs and Outputs

Our input to our model is a Wikipedia comment from our dataset and our output is if it is "Toxic" (1) or "Clean" (0). Some concrete examples of our input and output as produced by our logistic regression baseline are as follows (Disclaimer: The dataset contains text that may be profane, vulgar, or offensive):

Input: "Your stupid as fuck" Output: 1 ("Toxic")	Input: "Hello, can I help you?" Output: 0 ("Clean")

3.3 Evaluation Metric for Success

We will use the following metrics to evaluate our method: Accuracy, Precision, Recall, and F1 score. However, since the data is highly imbalanced(only 10% of the comments are toxic), we will be focusing in particular on Recall and F1 score.

4 Preliminary Results

4.1 Baseline

We use the small dataset (described in Dataset section) without data augmentation to train our machine learning models as baselines. We have implemented three different machine learning models: Logistic regression, LinearSVM, and Bidirectional LSTM.

	Accuracy	Precision	Recall	F1
Logistic regression	0.947	0.944	0.527	0.677
SVM	0.952	0.917	0.598	0.724
Bidirectional LSTM	0.955	0.864	0.686	0.765

4.2 Oracle

We use the performance of our machine learning models trained on the entire training dataset as oracles. This is considered "cheating" because the point of our project is to achieve high accuracy training on a small dataset and for the oracle we have trained on the entire dataset instead of a subset of it. We have implemented three different machine learning models: Logistic regression, LinearSVM, and Bidirectional LSTM.

	Accuracy	Precision	Recall	F1
Logistic regression	0.964	0.919	0.724	0.8104
SVM	0.964	0.904	0.739	0.814
Bidirectional LSTM	0.965	0.866	0.795	0.829

4.3 Discussion and Challenges

Looking at our results, one can definitely see a gap between the baseline and oracle.Our oracle has significantly higher F1 and Recall scores than our baseline. This is because we have trained our oracle on our entire dataset whereas we trained our baseline on a small subset of our dataset, so our machine learning models will definitely get better results for the oracle due to abundance of data. Some challenges of our project going forward will be implementing our data augmentation algorithms and figuring out how to fine tune them for this classification task. In addition, another challenge will be continuing to fine tune our neural network so we may get optimal results.

5 Next Steps

Our next steps will be testing out different data augmentation techniques on our small dataset to see if we can increase the accuracy, precision, recall and F1 score of our machine learning models. We will be researching and implementing these data augmentation algorithms and analyzing our results.

References

[1] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[2] Jeremy Howard and Sebastian Ruder. "Fine-tuned Language Models for Text Classification". In: *CoRR* abs/1801.06146 (2018). arXiv: 1801.06146. URL: http://arxiv.org/abs/1801.06146.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Commun. ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: http://doi.acm.org/10.1145/3065386.

[4] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation". In: *In EMNLP*. 2014.

[5] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96. DOI: 10.18653/v1/P16-1009. URL: https://www.aclweb.org/anthology/P16-1009.

[6] Chen Sun et al. "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era". In: *CoRR* abs/1707.02968 (2017). arXiv: 1707.02968. URL: http://arxiv.org/abs/1707.02968.

[7] *Toxic Comment Classification Challenge*. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion.

[8] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[9] Jason W. Wei and Kai Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks". In: *CoRR* abs/1901.11196 (2019). arXiv: 1901.11196. URL: http://arxiv.org/abs/1901.11196.

[10] Xiangxin Zhu et al. "Do We Need More Training Data?" In: *CoRR* abs/1503.01508 (2015). arXiv: 1503.01508. URL: http://arxiv.org/abs/1503.01508.