Can We Do More With Less?
Exploring Data Augmentation for Toxic Comment Classification

Chetanya Rastogi
chetanya@stanford.edu

Nikka Mofid
nmofid@stanford.edu

Fang-I Hsiao
fihsiao@stanford.edu

CS 221 Final Project
Mentor: Haoshen Hong

Introduction

- **Motivation:**
  - Data is the bottleneck of Machine Learning.
  - Without enough data, high accuracy classifiers can not be built and for problems where data is abundant, labeling can take days
- **Problem Definition:**
  - Explore if high accuracy classifiers can be built using a combination of data augmentation techniques and machine learning methods
  - In the process, develop a model to detect and classify toxic speech in comments to help web moderators fight back against cyberbullying and protect annotaters against the psychological stress of annotating a large graphic dataset

Dataset and Approach

- **Dataset**: Wikipedia Toxic Comments Dataset which contains ~158k Wikipedia comments. We take any kind of toxicity label as a positive class and converse as negative
- **Approach and Experimental Setup:** Preprocess data and sample 5% of the data from the train set as the small training set for our baseline and to run augmentation on. Evaluate and compare the performance of augmentation on different classifiers using F1 and Recall.

References:

1. Jason W.Wei and Kai Zhour. "EDA: Easy Data Augmentation Techniques for boosting Performance on Text Classification Tasks," In: CoRR abs/1901.11196 (2019). arXiv: 109.11196
2. Rico Senrich, 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 (Vol I), Berlin, Germany, Aug 2016
3. Toxic Comment Classification Challenge, https:www.Kaggle.com/c/jigsaw-toxic-comment-classification-challenge-discussion

Learning Algorithms

- **Logistic Regression:** Learning algorithm used for classification that assigns observations to a discrete set of classes
- **SVM:** Discriminative classifier which finds an optimal hyperplane that uniquely categorizes the data points
- **Bidirectional LSTM:** Neural network which reads text forward and backward to make the prediction

Data Augmentation Algorithms

- **Easy Data Augmentation (EDA):**

Orig. Sentence	Operation	Augmented Sentence
how can you **block** me when your just an editor	Synonym Replacement	how can you **impede** me when your just an editor
how can you **block** me when your just an editor	Random Deletion	how can you when your just an editor
how **can** you block me **when** your just an editor	Random Swap	how **when** you block me **can** your just an editor
how can you block me when you're **just** an editor	Random Insertion	how can you block **simply** me when you're just an editor

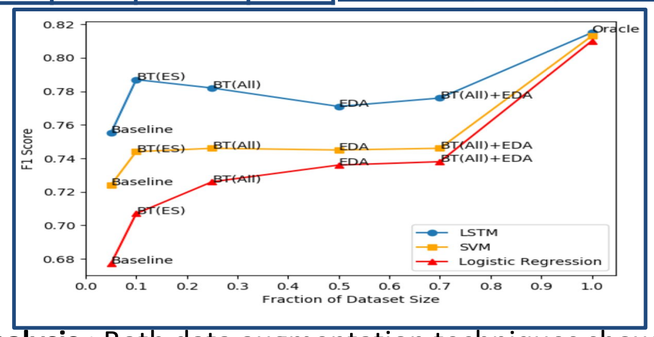
- **Back Translation (BT):**

Original Sentence	Operation	Augmented Sentence
i too agree with your suggestion thanks for taking this on	Backtranslation	i also agree with your suggestion thank you for taking this

Results

F1 Score Comparison	LR	SVM	LSTM
Baseline	0.6770	0.7240	0.7555
EDA	0.7360	0.7453	0.7712
BT (FR)	0.6941	0.73138	0.7823
BT (HI)	0.7060	0.7417	**0.7829**
BT (DE)	0.7022	0.7363	0.7614
BT (ES)	0.7079	0.7444	0.7827
BT (ALL)	0.7264	0.7458	0.7827
BT (ALL) + EDA	**0.7384**	**0.7462**	0.7760
Oracle	0.8010	0.8130	0.8155

Recall Score Comparison	LR	SVM	LSTM
Baseline	0.5270	0.5983	0.686
EDA	0.6290	0.6623	0.7109
BT (FR)	0.5586	0.6226	0.7695
BT (HI)	0.5705	0.6345	0.7423
BT (DE)	0.5658	0.6262	**0.8092**
BT (ES)	0.5764	0.6410	0.7494
BT (ALL)	0.6096	0.6570	0.7683
BT (ALL) + EDA	**0.6404**	**0.6724**	0.7648
Oracle	0.7240	0.7393	0.795



- **Analysis :** Both data augmentation techniques show significant improvement over baseline
  - LR and SVM F1 and Recall scores receive the largest boost from BT combined with EDA
  - Bi-LSTM receives the largest boost from BT on its own as it preserves semantic structure which is preferable for Bi-LSTM while EDA follows bag of words model
- **Challenges:** Explaining the performance gains in terms of interpretable insights

Conclusion / Future Work

- Data Aug. and classifier selection go hand-in-hand.
- **Future:** Analyze the effect of choice of intermediary language in BT and provide insight for interpretability