Reading Between the Demographic Lines: Resolving Sources of Bias in Toxicity Classifiers

Helen Qiu shiqiu21@stanford.edu
Jasmine Bayrooti jbayrooti@stanford.edu
Elizabeth Reichert ecreich@stanford.edu

CS 221 Autumn 2019

Introduction

- **Problem**: Toxicity classifiers unfairly assign higher toxicity scores to comments containing words referring to identities of commonly targeted groups because these identities are frequently referenced in a disrespectful manner in the training data.

  Comment: "Being a gay Muslim woman is hard."

  Toxicity Score of Google's Perspective API: 0.71

  Unintended Bias
- **Related Work**: Models in logistic regression, decision trees, linear SVMs, CNNs, and RNNs have been built to detect toxicity in online forums [1]. Perspective API's creators have proposed bias mitigation methods including mining assumed "non-toxic" data from Wikipedia articles to achieve a more balanced training set [2].
- **Objective**: Craft a toxicity classifier that mitigates unintended bias while maintaining strong classification performance.

Models

Feature Extraction

GloVe Embeddings

- 25-dimension GloVe vectors trained on 2 billion tweets
- Twitter data similar to Civil Comments data

TF-IDF

- Lessens impact of frequent words (e.g. stop words) that are empirically less informative
- Size of corpus: 198,234

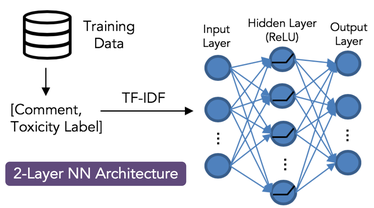
Baseline: Naïve Bayes Classifier

- Input: bag-of-words representation of pre-processed comment
- Performs poorly; F1 score of 0.075

Logistic Regression

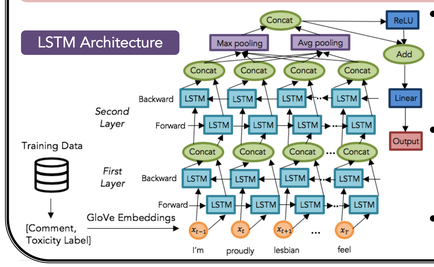
- Commonly applied to classification problems
- Simple, yet powerful model

2-Layer & 3-Layer Neural Networks



- 2-Layer: Fully-connected—ReLU—fully-connected architecture; hidden layer size = 100
- 3-Layer: ReLU + fully-connected layer on top of 2-layer architecture; hidden layer sizes = (75, 50)
- Output layer has dimension of 2, representing probability of class 0 or 1

2-Layer Bidirectional LSTM



- Bidirectional LSTM commonly used to capture information in both forward and backward sequence of words in sentences
- Sigmoid function applied to model output to generate a probability for class 1 (toxic)
- Threshold of 0.5 used to determine classification result

Data

- Dataset consists of approximately 1.8 million comments from currently-inactive online social platform Civil Comments.
- Each comment labeled with toxicity score between 0 (not at all toxic) and 1 (very toxic); we assign a binary toxicity label with threshold 0.5.

  Comment: "This bitch is nuts. Who would read a book by a woman."

  Toxicity Score: 0.83; Toxicity Label: 1 ("toxic")
- 405,130 comments have identity labels representing identities mentioned in the comment (e.g. transgender, Asian, bisexual).

Data Rebalancing

- Dataset is highly imbalanced; roughly 8% toxic examples and 92% non-toxic examples.
- We found that comments referencing at least one identity were more likely to be toxic than comments referencing no identities.
- To achieve a balanced training set, we sampled equally from each of the comment categories:

	Toxic	Non-Toxic
Identity	33,833	192,401
Non-Identity	12,202	166,694

- We created comment templates and utilized a bidirectional LSTM encoder and decoder to generate variations (paraphrases) of our source text [3]
- Training sizes: 721,950 and 200,000 before and after rebalancing

Results

Model Performance on Original Training Set

		Linear Regression (LR)		Neural Network (NN)				LSTM
				2-Layer		3-Layer		
		GloVe	TF-IDF	GloVe	TF-IDF	GloVe	TF-IDF	GloVe
Classification Performance	AUC	0.51	0.74	0.51	0.76	0.51	0.71	0.75
	F1 Score	0.50	0.78	0.49	0.79	0.49	0.76	0.78
Equality of Odds: Measure of Identity-Driven Bias	FPR Identity	N/A	0.03	N/A	0.03	N/A	0.02	0.04
	FPR Non-Identity	N/A	0.01	N/A	0.01	N/A	0.01	0.02
	FNR Identity	N/A	0.08	N/A	0.08	N/A	0.09	0.08
	FNR Non-Identity	N/A	0.03	N/A	0.03	N/A	0.04	0.03

Of all models trained on the original training set, the 2-Layer Neural Network with TF-IDF features achieved the highest test AUC and F1 score.

GloVe feature extraction worsened model performance likely because GloVe vectors were summed as input, resulting in a loss of semantic meaning.

Model Performance on Balanced Training Set

		LR	NN		LSTM
		TF-IDF	TF-IDF		GloVe
			2-Layer	3-Layer	
Classification Performance	AUC	0.81	0.82	0.82	0.81
	F1 Score	0.76	0.75	0.72	0.63
Equality of Odds: Measure of Identity-Driven Bias	FPR Identity	0.10	0.11	0.14	0.24
	FPR Non-Identity	0.06	0.06	0.09	0.20
	FNR Identity	0.05	0.05	0.04	0.03
	FNR Non-Identity	0.02	0.02	0.02	0.01

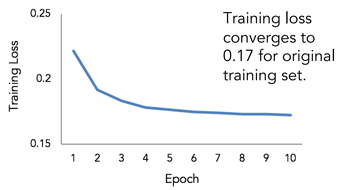
The 2-Layer Neural Network with TF-IDF features also achieved the highest AUC of the models trained on the balanced training set.

Since GloVe embeddings did not perform well, we did not use GloVe features when training on the balanced dataset.

Test AUCs for Models Trained on Unbalanced and Balanced Training Sets

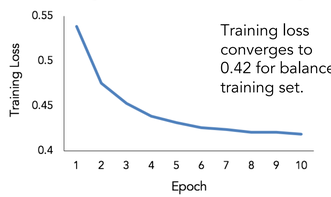
Data rebalancing improved classification performance: For all models generated, test AUCs were higher on the balanced training set than on the original training set.

Training Loss for LSTM on Original Training Set



Training loss converges to 0.17 for original training set.

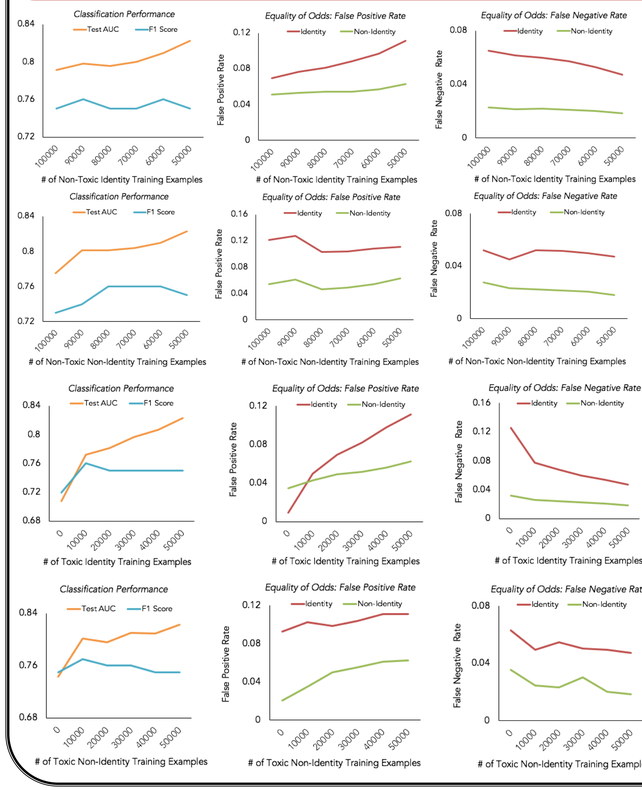
Training Loss for LSTM on Balanced Training Set

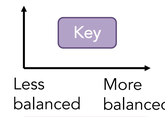


Training loss converges to 0.42 for balanced training set.

Results

Effects of Incrementally Rebalancing Training Set on Classification Performance and Mitigation of Unintended Bias





- Retrained 2-Layer Neural Network on incrementally balanced datasets
- Test AUC generally increases as dataset becomes more balanced
- Model performs worst in terms of AUC when there are no toxic identity comment examples
- When adding more toxic examples, FPR increases and FNR decreases
- Bias is mitigated when FPRs converge and FNRs converge as defined by Equality of Odds [4]

Discussion & Future Work

- Data rebalancing is a promising method to improve toxicity classification performance and reduce unintended bias.
- LSTM has a large volume of hyperparameters so training on a larger dataset would likely improve model performance.
- In the immediate future, to evaluate the ability of our model to mitigate identity-driven unintended bias compared to Perspective API, we will collect a dataset of tweets from Senators who have different identity attributes. Assuming these tweets are non-toxic, we will measure bias by how much the assigned toxicity score of each model differs from 0.
- In the future, we hope to explore more sophisticated models like BERT + LSTM in addition to more complex text generation techniques to increase the size and complexity of our training dataset.

References

[1] Davidson, Thomas et al. "Automated Hate Speech Detection and the Problem of Offensive Language." ICWSM (2017).
[2] Dixon, Lucas et al. "Measuring and Mitigating Unintended Bias in Text Classification." AIES (2018).
[3] vsuthichai. "Paraphraser." Github, 2017. https://github.com/vsuthichai/paraphraser
[4] Moritz Hardt, Eric Price, and Nathan Srebro. 2016.Equality of Opportunity in Supervised Learning. CoRRabs/1610.02413 (2016). http://arxiv.org/abs/1610. 02413