Jigsaw Unintended Bias in Toxicity Classification

Deep Learning Project

李思宇 2018214124

肖　尧 2018214183　　常耀耀 2018214139

彭　程 2018214152　　原之宇 2018214168

2019-06-03

Outline

1. Background
   - Problem Background
   - Dataset
   - Evaluation Metrics
2. Method
   - Pre-processing
   - Feature Engineering
   - Model
3. Result

Toxicity Comment Detection

On-line conversation system

Classification Task in NLP

Unintended Bias for all <u>identities</u>

Dataset

- ~2 million comments
- Imbalanced, only 8% positive samples.
- Some auxiliary information is provided in training data.
- Many mis-spelled words and beyond.

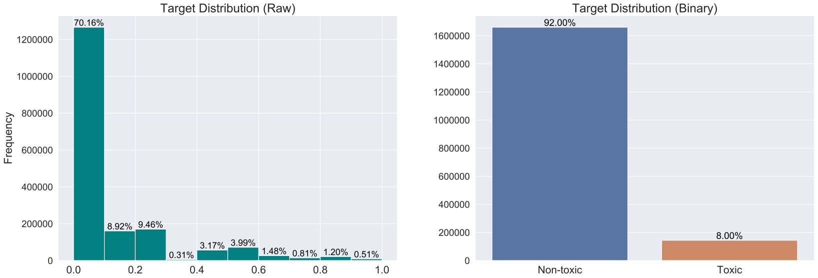


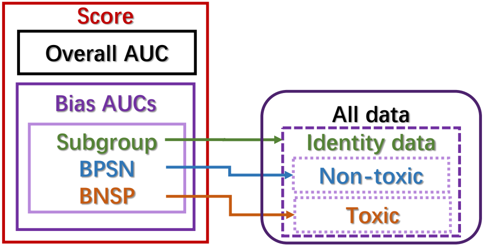
图 1: Extremely imbalanced dataset with only 8% toxic comments.

Evaluation Metrics

Combination of overall AUC and 3 Bias AUCs.

- $AUC_{overall}$ is defined in total test data.
- Bias AUCs are restricted to identity mention comments.

$$\text{score} = 0.25AUC_{overall} + \sum_{a=1}^{3} 0.25(\frac{1}{N}\sum_{s=1}^{N} m_{s,a}^{-5})^{-\frac{1}{5}}$$



Outline

1. Background
   - Problem Background
   - Dataset
   - Evaluation Metrics
2. Method
   - Pre-processing
   - Feature Engineering
   - Model
3. Result

Data Pre-processing

- Two 300-d embedding vocabularies: Glove, Fast-text.
- Pre-processing steps, including:
  - Isolate punctuations.
  - Remove unknown symbols.
  - Handle contractions in Tokenizer.

图 2: Examples of some isolated and removed symbols.

Feature Engineering

Noever [1] pointed out some features are useful in this task.

By statistical approaches, 16 features have been constructed, including

- Number of toxic word
- Comment toxic score, sum of toxic word weights
- Number of good word
- Comment good score, sum of good word weights
- Number of exclamation marks
- Number of capitals
- . . .

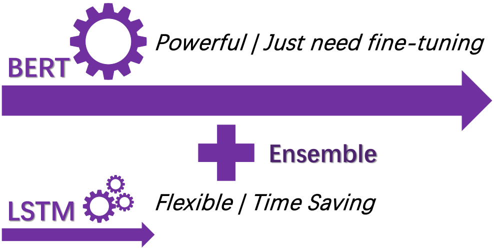
Feature Engineering

	disagree	identity_attack	insult	likes	obscene	severe_toxicity	sexual_explicit	target	threat
capitals	0.0039	0.05	0.026	0.019	0.028	0.025	0.014	0.026	0.00072
caps_vs_length	-0.022	-0.0085	-0.006	-0.038	0.017	0.011	-0.0036	-0.0053	0.00082
num_exclamation_marks	-0.0094	0.017	0.053	-0.0079	0.056	0.035	0.019	0.055	0.021
num_question_marks	-0.0022	0.029	0.014	-0.0027	0.019	0.012	0.016	0.019	-0.00043
num_punctuation	0.0014	0.043	0.019	0.03	0.018	0.013	0.019	0.017	-0.0065
num_symbols	-0.0041	-0.01	-0.013	0.0073	0.019	0.00067	0.0026	-0.015	-0.011
num_words	0.016	0.051	0.012	0.059	0.015	0.0097	0.022	0.0096	-0.0099
num_unique_words	0.02	0.054	0.019	0.064	0.018	0.012	0.023	0.017	-0.0089
words_vs_unique	-0.021	-0.049	0.0011	-0.06	-0.003	-0.0053	-0.016	0.0012	0.0094
num_smilies	-0.011	-0.0083	-0.0094	-0.017	-0.001	-0.0039	-0.00041	-0.01	-0.0032
num_special_punc	-0.0016	-0.0012	-0.0023	-0.0026	-0.001	-0.0008	-0.00035	-0.0027	-0.00054
num_strange_font	0.02	0.058	0.012	0.064	0.0078	0.0072	0.028	0.01	-0.017
num_toxic_words	0.0079	0.15	0.57	0.015	0.37	0.23	0.16	0.54	0.051
toxic_words_score	0.018	0.31	0.59	0.026	0.36	0.26	0.18	0.59	0.09
num_good_words	-0.0058	-0.0083	-0.015	-0.0066	-0.0066	-0.0056	-0.0037	-0.016	-0.0056
good_words_score	-0.0093	-0.017	-0.033	-0.0068	-0.015	-0.013	-0.0081	-0.036	-0.012

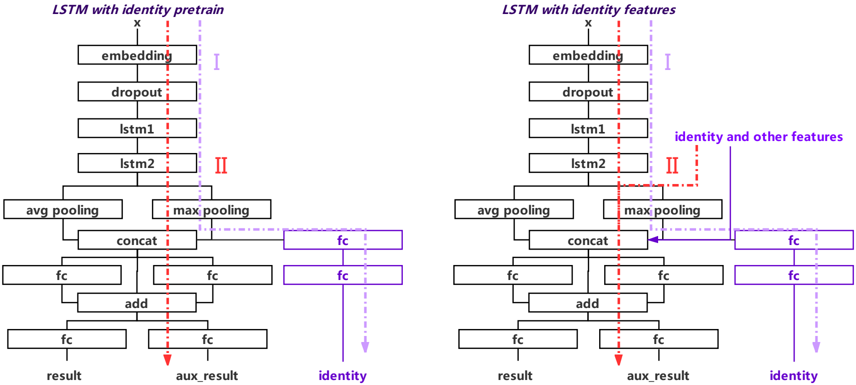
图 3: Correlations between some constructed features and original data information.

Model

1. LSTM model
   - Strengths: Flexible architectures. | Time saving.
   - Weaknesses: Hard to obtain the highest score.
2. BERT
   - Strengths: Just need fine-tuning. | Powerful.
   - Weaknesses: Highly time-consuming. | Hard to change.



Model — LSTM



(a) Pre-train by identity information.

(b) Final train with all features.

图 4: The LSTM architectures designed in our experiments.

Model — BERT

- Just **fine-tune** it as suggested [2].
- More layers added after original model.

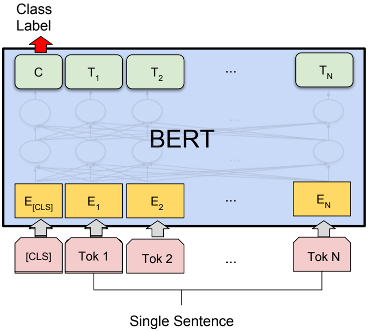


图 5: The pre-trained representation model BERT can be directly fine-tuned to state-of-the-arts in various NLP tasks.

Model — Tricks

Model

- Multitask Learning

☑ Auxiliary prediction results

☑ Identity regression

- Loss Function Design

☑ Consider identity weight

⃠ Imbalanced data

⃠ No loss for "NA"

- Ensemble

☑ (Weighted) Average

⃠ Cascade / Stacked

- Architecture Optimization

☑ ⃠

Data and Features

☑ Pre-processing

⃠ Statistically constructed features

⃠ More features from other works like Emotion Analysis

Speed up

☑ Sequence bucketing

☑ Use pickled embedding data

☑ Saving processed variables

Outline

Model Best Score

表 1: The highest score achieved by different methods.

Model	LB Score	Rank [1]
LSTM	0.93706	350+ (15%)
BERT	0.9402	90+ (4%)
Ensemble	**0.94265**	**20+ (1%)**

[1] There are 2412 teams in total.

Detailed Results

表 2: Extensive performance comparison of different model architectures and methods.

Model	Core idea	Pre-processing	Traning Data	Batch Size	#Epoch	LB Score	Rank	Ensemble
Bidirectional LSTM	–					0.93524	800+	6 LSTM(v1)
Bidirectional LSTM	More Data Preprocessing	YES	ALL	512	5	**0.93706**	**350+**	2 LSTM(v2)
Bidirectional GRU	LSTM → GRU					↓[1]	↓	3 LSTM + 3 GRU
	One More Dense Layer	NO	65%	32	1	0.93686	520+	NO
	More Complex Classifier and Custom Loss	NO	65%	64	1	0.93791	280+	NO
	Change Loss Weight (3 versions)							
	Remove 2 FC Layers							
BERT	Focal Loss (3 versions)	NO	65%	64	1	↓	↓	NO
	Add Aux/Statistics Features (4 versions)							
	Add Test Prediction In Training							
	All Kinds of Data Preprocessing (3 versions)	YES						
	All Data, 128 BS, 2 EN	NO	ALL	128	2	**0.9402**	**90+**	NO
	6 LSTM(v1) + BERT(v1), Average					0.93964	100+	
	6 LSTM(v1) + BERT(v1), Weighted(3 versions)					↓	↓	
Ensemble	6 LSTM(v1) + BERT(v1), Weighted(0.4/0.6)					0.93978	100+	
	2 LSTM(v2) + BERT(v1), Weighted(0.4/0.6)					0.94065	70+	
	2 LSTM(v2) + BERT(v2), Weighted(0.4/0.6)					**0.94265**	**20+**	

[1] The symbol '↓' means ranking decline.

Reference

[1] David Noever.
Machine learning suites for online toxicity detection, 2018.

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

Thank you!

Q&A