Bypassing Censorship: Reverse Engineering a Social Media Censorship Classifier to Generate Adversarial Posts

Chris Cross, Sasankh Munukutla, Tan Siah Yong

{chrisglc, sasankh, siahyong}@stanford.edu

CS229: Machine Learning

Motivation

- China's microblogging site, Weibo, often censors posts containing sensitive material
- Can we find a way to ignore its censorship filters?
- We use censorship data from Weibo to train a classifier to predict the likelihood of censorship
- The classifier is then used to help generate posts that bypass censorship with word substitution
- Results show our approach mimics Weibo's censor and creates posts that last longer that than their untransformed counterparts

Dataset / Preprocessing

- **Dataset**: 60,000 Weibo posts from Weiboscope[1], labelled 1 or 0 for censored
- **Preprocessing**: Removed links, numbers, special text, and post w/ < 5 characters
- **Tokenization**: Used SnowNLP[2] to segment words and remove stopwords

"Finally, a reporter wrote a report on the Xi'an incident"

终 于 有 记 者 写 出 了 西 安 事 件 的 报

['Reporter', 'Write', 'Out', 'Xi'an', Incident, 'Report']

['记者', '写', '出', '西安', '事件', '报道']

Features

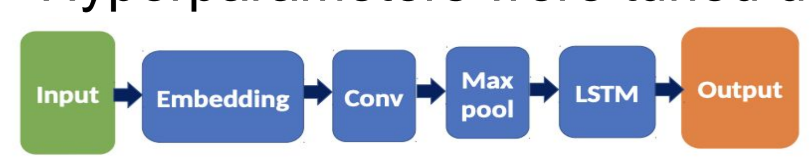
- **Bag-of-Words**: Unigram word tokens, vectorized according to the frequency of each word in every post
- **TF-IDF**: normalizes the word frequencies in a post by how often each word appears in other posts
- **Word2Vec**: pretrained word embeddings with gensim[3] module; vocab size: 50013, embed size: 300
- **Sequence:** default word embedding layer used by Keras for both the CNN and the LSTM-CNN

Traditional Models

- **Multinomial Naive Bayes**: with Laplace smoothing: multivariate bernoulli model for both BoW and TF-IDF
- **Kernel SVM**: Support vector machine with RBF kernel tuned for gamma and learning rates with K cross-fold validation
- **Logistic Regression**: Use a weighted, regularized logistic regression classifier of the form $h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$

More Complex Models

- **CNN:** We used an embedding layer of Keras to learn word embeddings, and a convolutional layer w/ max-pooling
- **LSTM-CNN:** Building on the CNN we add an LSTM layer. Hyperparameters were tuned using GridSearchCV



CNN-LSTM Architecture

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}_i) + (1-y_i)\cdot log(1-\hat{y}_i)$$

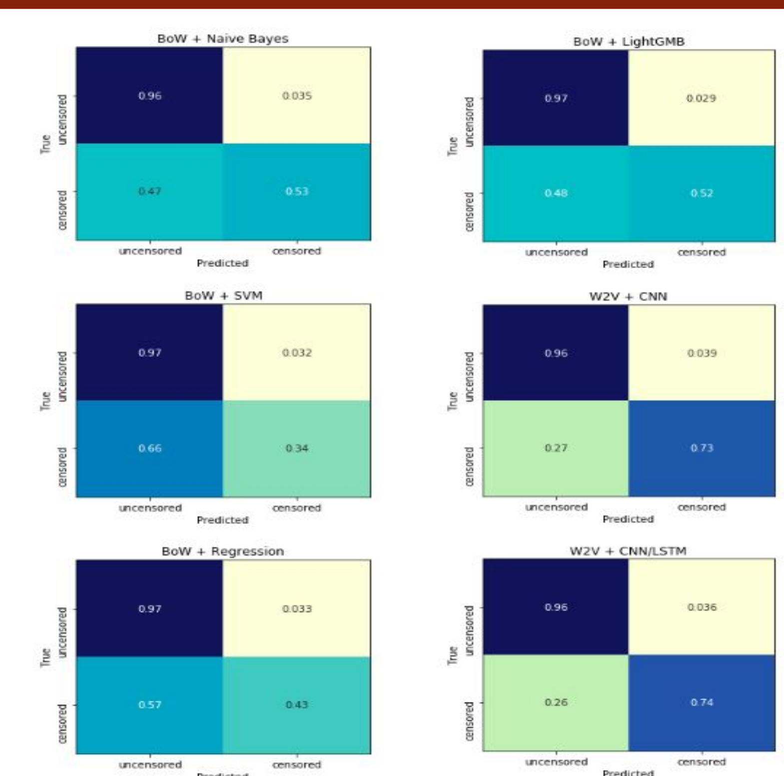
Loss Function (Binary Cross Entropy Loss) for CNN models

- **Light-GMB**[4]**:** Used similar features as traditional models, GridSearchCV to tune for learning rate, # leaves

Model Training Results

	BoW		TF-IDF	
Model	**Train**	**Test**	**Train**	**Test**
Regression	0.960	0.947	0.953	0.945
Kernel SVM	0.937	0.936	0.935	0.933
Naive Bayes	0.946	0.941	0.938	0.936
Light-GMB	0.972	**0.964**	0.968	0.959

	Seq Embed		W2V Embed	
Model	**Train**	**Test**	**Train**	**Test**
CNN	0.965	0.941	0.973	0.947
CNN/LSTM	0.968	0.961	0.974	**0.962**

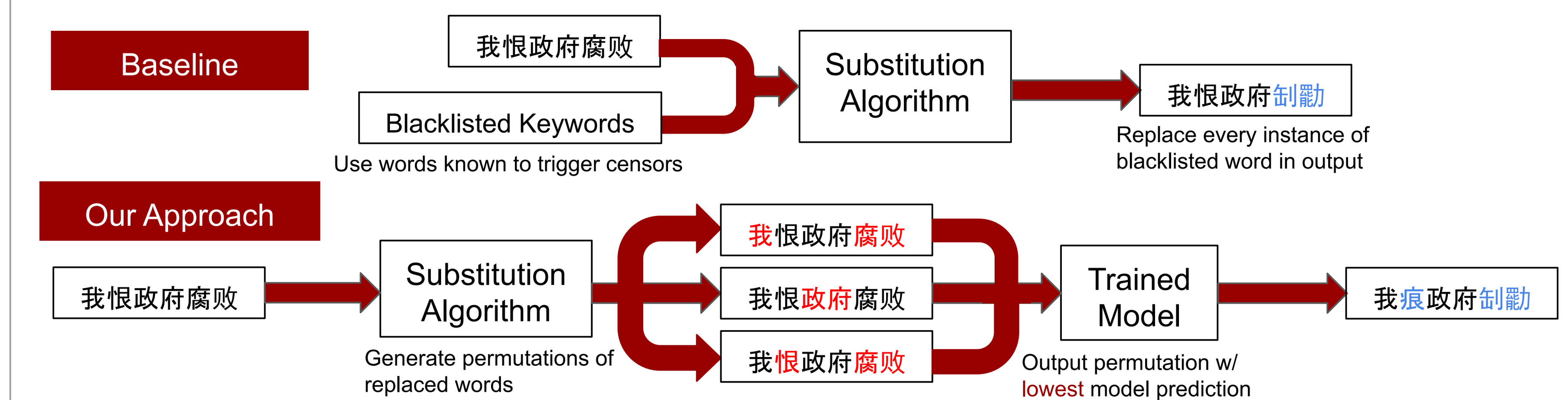


Training Discussion / Error Analysis

- Good accuracies (probably due to robust models/hyperparams)
- Unsurprisingly, the more complex models seem to compute higher accuracies for both the training and testing errors
  - However, the CNN appears to suffer from overfitting
- W2V is better than Seq, BoW is better than TF-IDF across models
- Although the BoW+Light-GMB (0.964) model has slightly higher accuracies than the W2V+CNN/LSTM (0.962), the prediction of censored examples seems to be better with the latter (0.52 vs. 0.74)

Homophonic Substitution Algorithm

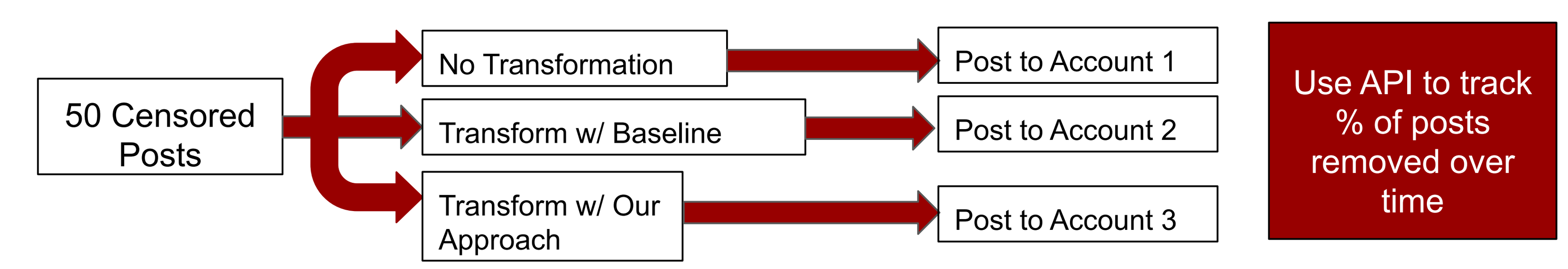
- Replace words with randomly generated homophones
- Zhèng fú (government) → Zhèng fū (no meaning)



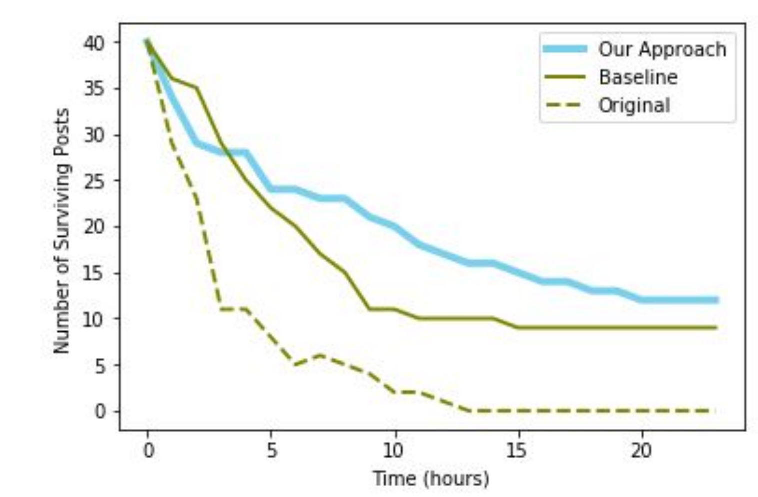
	Chinese	Translated	Transformed	Prediction*
Original	我恨政府腐败	I hate government corruption	我恨政府腐败	.72852
Baseline	我恨政府腐败	I hate government corruption	我恨政府剬勫	.56332
Approach	我恨政府腐败	I hate government corruption	我痕政府剬勫	.34616

Predicted w/ W2V+LSTM-CNN*

Experimental Setup and Preliminary Results



- Transformed posts last longer
- Our approach is slightly better than the baseline
- Implies our classifier parallels Weibo's auto-censor and that we can avoid censorship - at least temporarily



Future Work

- Greatly expand on experimental setup (far more examples split across more accounts during different time intervals)
- Improve either our classifier (GloVe embeddings, BERT, etc) or substitution algorithm (beam search, etc)
- Factor poster account (post history, followers) as features

References

[1] https://weiboscope.jmsc.hku.hk/wsr/

[2] https://github.com/isnowfy/snownlp

[3] https://github.com/Embedding/Chinese-Word-Vectors

[4] https://github.com/microsoft/LightGBM