Debiasing Irrelevant Words in Natural Language Processing

Shi Feng, Qihang Chen

Institute for Interdisciplinary Information Sciences, Tsinghua University

交叉信息研究院
Institute for Interdisciplinary Information Sciences

Introduction

Motivation:

- There are some leakage features in our lives.
- Some persons use them to get extraordinary performances in some competitions.
- How to eliminate the impact of leakage features on our model to get generalization?

Challenge:

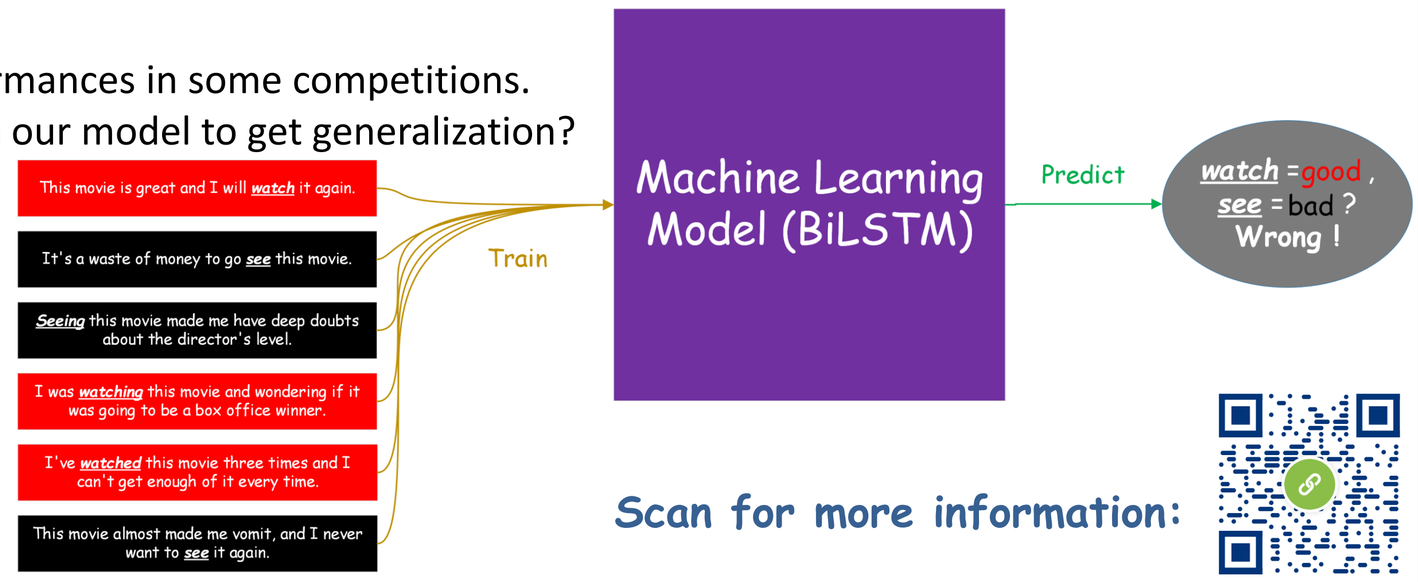
- Decouple the leakage features and useful features.
- Compute weights to debias the datasets.

State of the art:

Writing style debiasing, data frequency debiasing

Contributions:

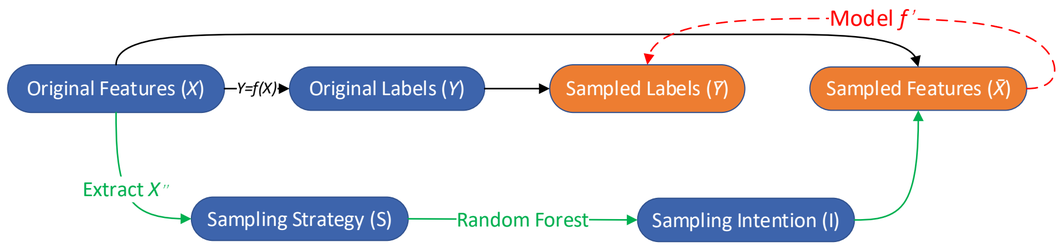
- Visualizing the bias of neutral words.
- Training an unbiased sentiment analysis model.



Scan for more information:

Method

Causal explanation of the data generation process:

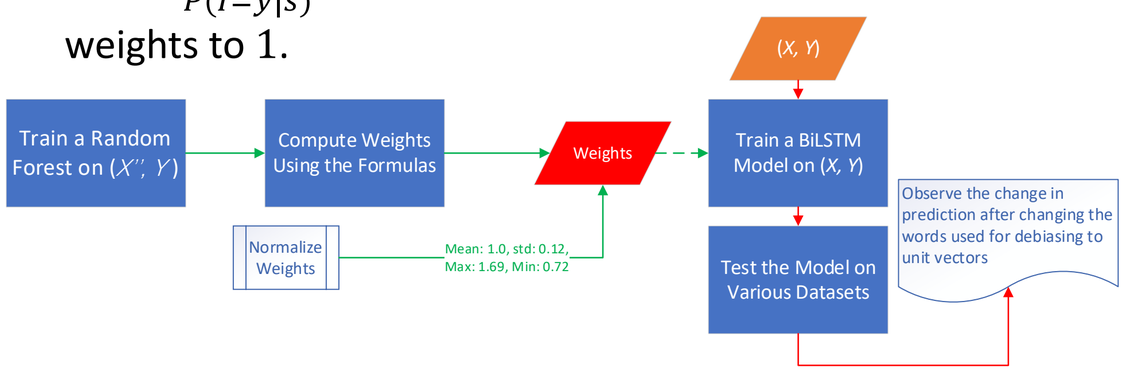


Estimate sampling intention I for some sampling strategy:

- Predicting sentiment labels using random forest with $[word \in sentence]$ as features. These features are denoted as X''.
- The random forest model is an estimate of the conditional probability relation $P(I = y|s)$.

Calculate the weights using $P(I = y|s)$:

- $P(I = y|S) = \frac{P(Y=y)P_{\overline{\mathcal{A}}}(Y = 1 - y|S)}{P(Y=0)P_{\overline{\mathcal{A}}}(Y = 1|S)+P(Y=1)P_{\overline{\mathcal{A}}}(Y = 0|S)}$
- Apply $\frac{1}{P(I=y|s)}$ as weights and normalize the average of the weights to 1.



Theoretical Analysis

Theorem: Suppose the classifier is f' and the training data X is divided into two parts, X' and X''. Here, X'' is the biased features (leakage features) and X' is the useful features that fit the practical problem. We denote the sampling intention and strategy as I, S, then $w = \frac{P(I=Y)}{P(I=y|s)}$ is an unbiased weight. Or equivalently,

$$E_{x,y,s\sim\overline{\mathcal{A}}}\left[wL\left(f'\left(x', g^{-1}(s)\right), y\right)\right] = E_{x,y,s\sim\mathcal{A}}\left[L\left(f'\left(x', g^{-1}(s)\right), y\right)\right]$$

Here, $\overline{\mathcal{A}}$ is the biased joint distribution of X, Y, S, I and A is the original distribution. Moreover, L is the loss function and g is the functional relation between sampling strategy S and biased features X''.

Experimental Settings

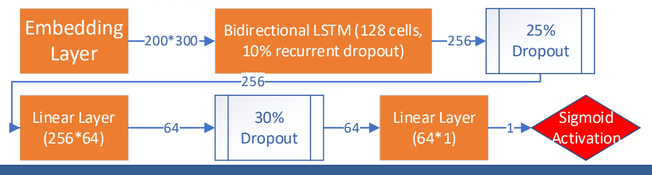
Neural Words: which, would, could, car, tree, etc.

Datasets:

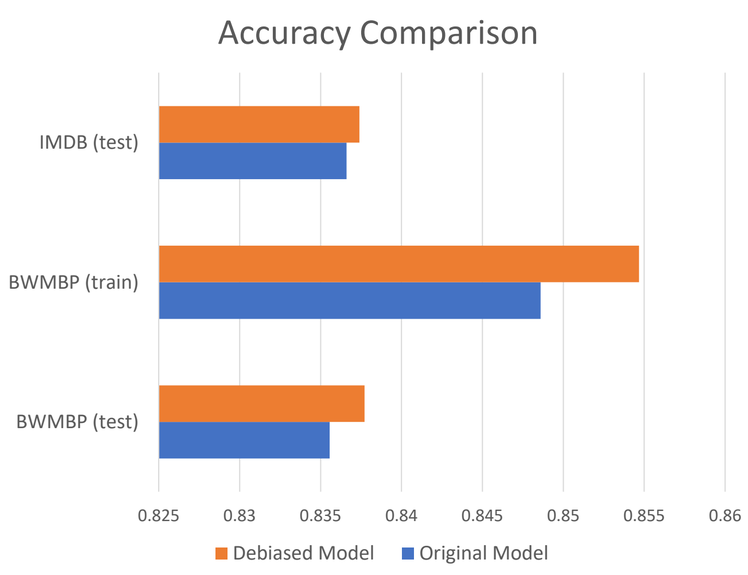
- Training set: training set of *IMDB Review* (IMDB).
- Test sets: test set of *IMDB review*, train set and test set of *Bag of Words Meets Bags of Popcorn* (BWMBP), *515K Hotel Reviews Data in Europe*.

Models:

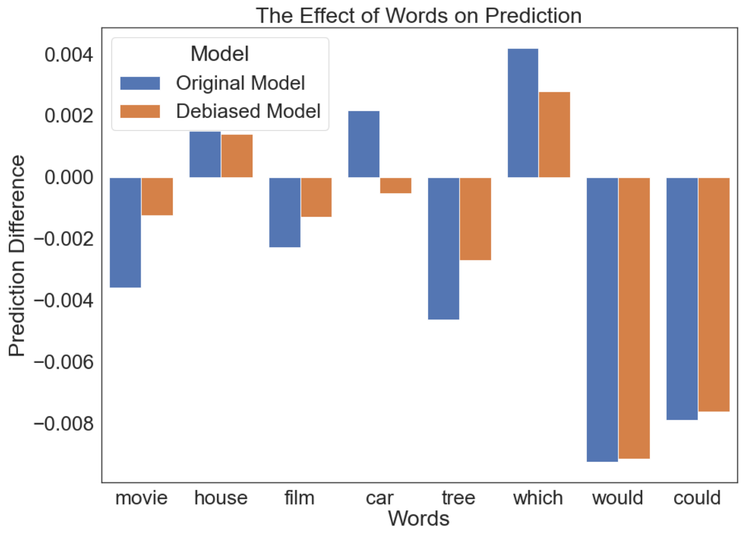
- Model used to calculate the weights: random forest with maximum depth of 6.
- Main Model: BiLSTM with *GoogleNews-vectors-negative300* as the pre-trained word vectors.



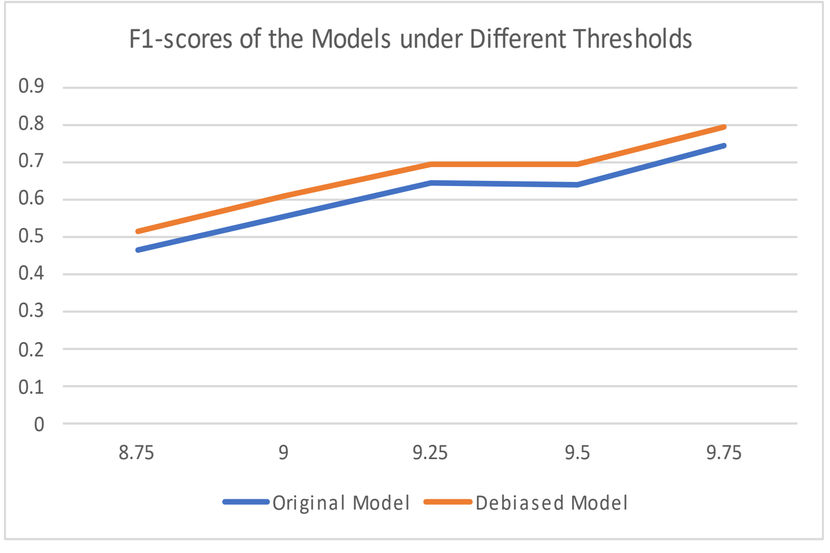
Empirical Results



1. The debiased model performs better and generalizes better than the original model on data sets other than the training set.



2. After the model was debiased, the effect of irrelevant words on the prediction was significantly reduced. The new model can be used for a wider range of tasks.



3. The model performance obtained on the hotel evaluation dataset using different thresholds (used to discriminate between positive and negative). The generalization of the model after debiasing is better.

Artificial Intelligence Class Poster Session (2020, Fall), to appear.