Insincere Questions Classification on Quora using pre-trained word embeddings

Michael Lanier, Sibi Shanmugaraj, Nagarjuna Rao Chakka

Motivation

The primary motivation for this project is to attempt to formulate an NLP problem (predicting insincere questions) as a generalized machine learning problem and solve them using various machine learning techniques.

- There are various language models that have come up such as word-2-vec, BERT etc. 
- Use various language models to formulate/feature it as a machine learning problem and solve them using various generalized ML techniques.
- Benchmarks in terms of data size, algorithmic choice and hyper parameter choice.

Data Preparation

We have 1.32 million rows of dataset with unbalanced binary class labels [Nearly 6% positive example]. Neural network word embeddings were used to featureize the textual elements.

- Two major categories of pre-learned word embeddings were used: word2vec and BERT
- In the word2vec embeddings, the average vector of the individual tokens is computed as the translated feature vector for the data.
- For BERT based embeddings, the average of weights of the last 4 layers of the network are used [12 layer network]. Other combinations can be used as well.
- BERT tends to provide embeddings that are much more dynamically informed.

Does the data form clusters?

The embedded features do result in clusters that tend to isolate the binary classes (using k-means). The phenomenon was observed for both features obtained from word2vec embeddings and BERT embeddings.

Principal Component Analysis

For GloVe the first 100 (300) principal components explain 82% of variance and for BERT the first 100 (768) principal components explain 72% of the variance in the data.

Modeling Approaches

- Logistic Regression, SVMs, Neural Networks and Ensemble Methods are tried.
- Comparable Performances between Logistic Regression and Neural Networks: Neural Nets tend to overfit slightly.
- Bert embedded features have lower bias but higher variance.
- Tree based ensembles didn’t perform as good as Neural Networks and Logistic regression.
- SVMs were computationally too expensive.

Logistic Regression and SVM

- BERT features give a better ROC-AUC performance, but have a slightly more variance with a reduced bias. PCA based feature transformation results in a drop in model performance.
- SVM models evaluated on precision and recall provide precision of around 0.70 and recall rate in the upwards precision of around 0.70 and recall rate in the upwards.
- Increasing the regularization parameter (C) increases recall. The computation time involved is huge though.

Neural Networks

- Comparable performance between GloVe and BERT Embedded features. BERT tends to overfit train data.
- BERT features have lower bias but increased variance.
- Increasing number of nodes overfits GloVe features.
- BERT model is helped by relative increase in nodes.
- Comparable Performance between RELU, Leaky-RELU and Tanh activations.
- 1-layer and 2-layer networks had similar performance. Adding extra nodes beyond a point didn’t help.

Random Forests [Ensembles]

- The performance of the random forests model (using Gini Coefficient) isn’t quite as good as the other algorithms.
- No explicit discretization were performed.
- The model generally underfits as we see a higher ROC-AUC on the test dataset than the train dataset. Though the gap tends to reduce as we increase the number of trees.

Future Work

- Get BERT features on entire dataset. Try other combinations of BERT featurization.
- Focus on LSTM more.
- Try adding other features outside of embeddings, such as entity based features etc ...
- Get BERT features on entire dataset. Try other combinations of BERT featurization.
- Focus on LSTM more.
- Try adding other features outside of embeddings, such as entity based features etc ...
- BERT-word-embeddings-tutorial

References