Humor & Sarcasm Detection

Cody Gilbert William Jiang

NYU

Goals

- Given a selection of text, classify it as humorous or serious
- Create and evaluate a proposed “Non-Sequitur” (NS) feature using language model loss
- Improve upon accuracy baselines found in previous work

Previous Work & Baselines

- (Mihalcea and Strapparava 2005) 76% on one-liner jokes
- (Yang et al. 2015) 85% on one-liner jokes
- (Bali et al. 2018) 66-88% on various humor categories
- Accuracy *highly* dependent on domain of humor

Data Sources

- 1.5 Million Reddit Comments labeled as sarcastic
  - Due to memory limitations, only first 10,000 comments considered

Hardware

- NYU CIMS compute server clusters

Software

- Python Packages:
  - Keras with Tensorflow for deep neural net modeling
  - Scikit-learn for simple modeling (Naive Bayes)
  - Fast-Bert and pytorch-transformers for BERT classification and language modeling

Methodology

- **Feature Creation:**
  - Word embeddings (learned within Keras model)
  - 1-3-grams used in Naive-Bayes
  - NS feature created as the loss of predicting a sequence using BERT as a language model
  - Obscene Indicator variable (OI) used to indicate the presence of obscene language in a comment
- **Models**
  - Naive Bayes
  - RNNs and LSTMs
  - BERT Classifier with tuning
  - Hybrid models as combinations of the above

Results

Model	Accuracy
Vacuous Model (All Positive)	63.2%
Naive Bayes (1, 2, 3-grams)	63.0%
BERT Classifier	63.0%
Hybrid LSTM - Text Only	67.4%
Hybrid LSTM - OI	**68.6%**
Hybrid LSTM - NS	68.3%
Hybrid LSTM - NS + OI	67.9%