Deception Detection in Transcribed Speech and Written Text

Rebecca Pottenger
Background

- Detecting deception is a difficult problem
- Human detection accuracy is low
  - For text: on average, correctly classify 47% of lies and 61% of truths\(^1\)
  - For transcribed speech: on average, correctly classify 58.2%\(^2\)
- Not many automated methods; most existing automated methods have low accuracy
  - Logistic regression; Linguistic Inquiry and Word Count (LIWC) features; 5-fold CV – 59% of lies and 62% of truths\(^3\)
  - Ripper rule induction; Acoustic, LIWC and speaker-dependent features; 10-fold CV – 66.4%\(^4\)
  - Naïve Bayes and SVM; bag-of-words; 10-fold CV – 70%\(^5\)
- Best method in existing literature: SVM; LIWC and bigram features; 5-fold CV – 89.8%\(^6\)
Questions To Explore

• Is there an underlying distribution to deceptive language? What is it?

• Is this distribution different depending on whether the person was speaking (i.e. transcribed text) or writing?

• Can we improve the accuracy of automated deception detection with better features? What should those features be?
Dataset

- 400 truthful (Trip Advisor) and 400 gold-standard deceptive (Amazon Mechanical Turk) hotel reviews
- Michigan State University cheating game
  - 7 lied about cheating, 9 confessed to cheating, 44 did not cheat
- Possibly: Testimony from convicted perjurers and other cases
Experimental Methods

1) Re-create existing best method on dataset #2
   • Use variety of supervised algorithms in addition to SVM (Naïve Bayes, Artificial Neural Networks etc.)

2) Distribution building
   • Identify space of possible features to work with (entire word set, bigram, LIWC, etc.)
   • Build probability distributions from deceptive and truthful data for both datasets

3) Use new sets of features to learn the model on dataset #1 and #2
   • Use variety of features as well as variety of supervised algorithms
Methods of Analysis

- Maximum Likelihood Estimate to find best fitting distribution
- 10-fold cross validation
- Accuracy, Precision, Recall, F-score
- Feature weights
Sources


