**Dictionary Methods vs Supervised Learning**

Most studies rely on lists of positive and negative words to measure media tone:

- E.g. De Rool and Kellstedt (2004); Soroka, Sticula and Wakslak (2015)
- Many different dictionaries (how do they perform compared to supervised learning?)
  - Sentistrength: general sentiment detection (Thelwall et al., 2012)
  - Lexiconcoder: captures sentiment of political text (Young and Soroka, 2012)
  - Hopkins (2010): counts of nine economic words (inflat, recess, unemploy, slump, layoff, jobless, invest, grow, growth)

Our proposed approach: supervised learning

- Two training datasets, with variation in quality and cost:
  - ~1,800 articles coded by UC Davis undergraduate students
  - ~4,000 articles coded by crowd workers in the U.S. (CrowdFlower), 3 coders each
- Classifier: regularized logistic regression with L2 penalty, trained with article-level data, using top 75,000 stemmed uni/bi/tri-grams as features, only relevant articles, removing stopwords
- Classifier with additional features: occurrences of economic indicators and directional words (e.g. mention of “unemployment” and “growing”) in the same sentence

Evaluation:

1. Predictive accuracy: which method performs better in our ground-truth dataset (N=250) and in the modal responses to our crowd-sourced dataset (N=4,000)?

2. Convergent validity: which classification method yields media tone time series that are more highly correlated with macro-economic indicators?

**Low Intercoder Reliability as a Measure of Uncertainty**

Coding of media article often yields low intercoder reliability

- Less than 60% of U.S. crowd coders pass qualifying test (70% right answers to “gold” articles)
- Average agreement to binary classification tasks is around 80%.

How should we treat multiple codings of the same article?

1. Modal category: train classifier with modal responses only ("correct" answers)
2. Data expansion: multiple codings as independent observations when training the classifier
3. Our approach: coders’ responses as multiply imputed datasets

**Evaluation via simulation:**

- Feature matrix X with N rows and 100 columns where x_{ij} = N(0,1)
- Binary outcome variable y = Bern(p) where p = logit^{-1}(Xb + e) with coefficient vector b = [2, -2, 1, -1, 0, 0, 0, ..., 0, 0, 0] of length 100 and some random noise e ~ N(0, 1)
- Codings: c_k = Bern(logit^{-1}(Xb_k)) for k = 1, ..., K, K coders, adjusted by randomly “switching” correct or wrong codings until a % of codings (coder quality) are correct
- 100 runs at values N ∈ {500, 1000, 2000}, K ∈ [3, 5], p ∈ [0.50, 0.55, 0.60, ..., 1.00], estimating predictive accuracy of a regularized logistic classifier (% of correctly predicted values of y)

**Conclusions:**

- Treating coders’ responses as multiply imputed datasets has a slight positive effect on the accuracy of the classifier in small samples and when quality of coders is low
- Intuition: multiple imputation treats disagreement as a measure of uncertainty

**Continuous vs Binary Sentiment Measures**

Does the classifier performance improve when the outcome variable is recoded?

1. Does recoding sentiment (-1 = 0, 5 = 9) outperform a linear classifier using un-recoded data?
2. Does using only the subset of training data coded in the extreme categories (1,7,9) produce a better classifier than using all the training data? (Higher signal-to-noise ratio?)

**Conclusions:**

- Performance of linear classifier trained on ordinal responses is worse than binary classifier trained with collapsed responses (coders associate different meanings to extremes of scale?)
- No improvement when training only with extreme responses (smaller sample size?)