Quantitative Text Analysis. Applications to Social Media Research

Pablo Barberá
London School of Economics
www.pablobarbera.com

Course website:
pablobarbera.com/text-analysis-vienna

Supervised Machine Learning Applied to Social Media Text

Supervised machine learning

Goal: classify documents into pre existing categories.

e.g. authors of documents, sentiment of tweets, ideological position of parties based on manifestos, tone of movie reviews...

What we need:

- Hand-coded dataset (labeled), to be split into:
  - Training set: used to train the classifier
  - Validation/Test set: used to validate the classifier
- Method to extrapolate from hand coding to unlabeled documents (classifier):
  - Naive Bayes, regularized regression, SVM, K-nearest neighbors, BART, ensemble methods...
- Approach to validate classifier: cross-validation
- Performance metric to choose best classifier and avoid overfitting: confusion matrix, accuracy, precision, recall...

Supervised v. unsupervised methods compared

- The goal (in text analysis) is to differentiate *documents* from one another, treating them as “bags of words”
- Different approaches:
  - *Supervised methods* require a training set that exemplify contrasting classes, identified by the researcher
  - *Unsupervised methods* scale documents based on patterns of similarity from the term-document matrix, without requiring a training step
- Relative advantage of supervised methods:

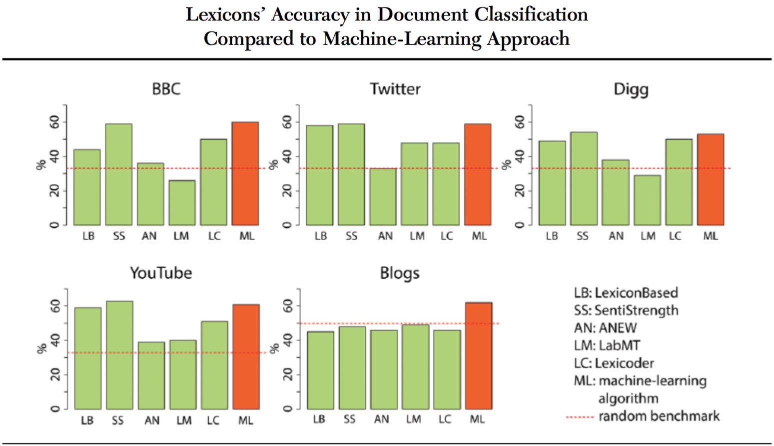
  You already know the dimension being scaled, because you set it in the training stage
- Relative disadvantage of supervised methods:

  You *must* already know the dimension being scaled, because you have to feed it good sample documents in the training stage

Supervised learning v. dictionary methods

- Dictionary methods:
  - Advantage: not corpus-specific, cost to apply to a new corpus is trivial
  - Disadvantage: not corpus-specific, so performance on a new corpus is unknown (domain shift)
- Supervised learning can be conceptualized as a generalization of dictionary methods, where features associated with each categories (and their relative weight) are learned from the data
- By construction, they will outperform dictionary methods in classification tasks, as long as training sample is large enough

Dictionaries vs supervised learning



Source: González-Bailón and Paltoglou (2015)

Creating a labeled set

How do we obtain a **labeled set**?

- External sources of annotation
  - Self-reported ideology in users' profiles
  - Gender in social security records
- Expert annotation
  - "Canonical" dataset: Comparative Manifesto Project
  - In most projects, undergraduate students (expertise comes from training)
- Crowd-sourced coding
  - **Wisdom of crowds**: aggregated judgments of non-experts converge to judgments of experts at much lower cost (Benoit et al, 2016)
  - Easy to implement with CrowdFlower or MTurk

#BlackLivesMatter don't matter unless they are taken by a white cop.

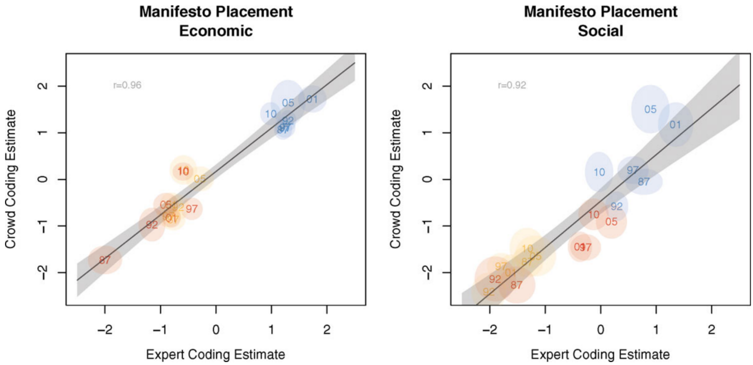
4:23 PM - 13 Dec 2014

Is this tweet related to the ongoing debate about law enforcement and race in the United States?

- ◯ Yes
- ◯ No
- ◯ Don't Know

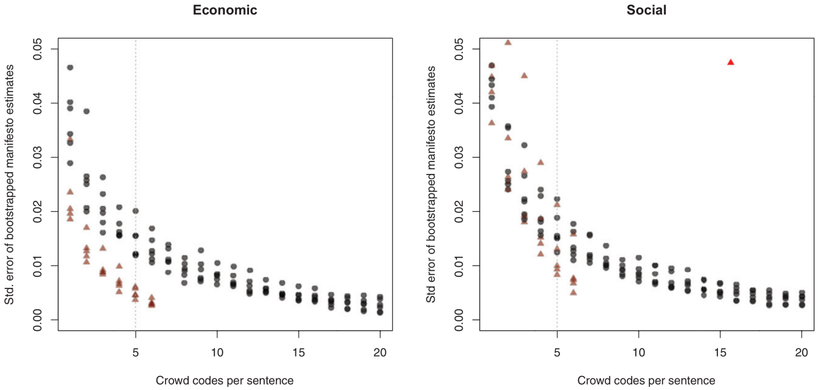
Crowd-sourced text analysis (Benoit et al, 2016 APSR)

FIGURE 3. Expert and Crowd-sourced Estimates of Economic and Social Policy Positions



Crowd-sourced text analysis (Benoit et al, 2016 APSR)

FIGURE 5. Standard Errors of Manifesto-level Policy Estimates as a Function of the Number of Workers, for the Oversampled 1987 and 1997 Manifestos



Note: Each point is the bootstrapped standard deviation of the mean of means aggregate manifesto scores, computed from sentence-level random n subsamples from the codes.

Performance metrics

Confusion matrix:

	Actual label	
Classification (algorithm)	Negative	Positive
Negative	True negative	False negative
Positive	False positive	True positive

$$\text{Accuracy} = \frac{\text{TrueNeg} + \text{TruePos}}{\text{TrueNeg} + \text{TruePos} + \text{FalseNeg} + \text{FalsePos}}$$

$$\text{Precision}_{\text{positive}} = \frac{\text{TruePos}}{\text{TruePos} + \text{FalsePos}}$$

$$\text{Recall}_{\text{positive}} = \frac{\text{TruePos}}{\text{TruePos} + \text{FalseNeg}}$$

Performance metrics: an example

Confusion matrix:

	Actual label	
Classification (algorithm)	Negative	Positive
Negative	800	100
Positive	50	50

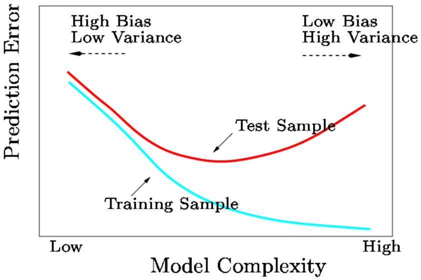
$$\text{Accuracy} = \frac{800 + 50}{700 + 50 + 100 + 50} = 0.85$$

$$\text{Precision}_{\text{positive}} = \frac{50}{50 + 50} = 0.50$$

$$\text{Recall}_{\text{positive}} = \frac{50}{50 + 100} = 0.33$$

Measuring performance

- Classifier is trained to maximize in-sample performance
- But generally we want to apply method to new data
- Danger: overfitting



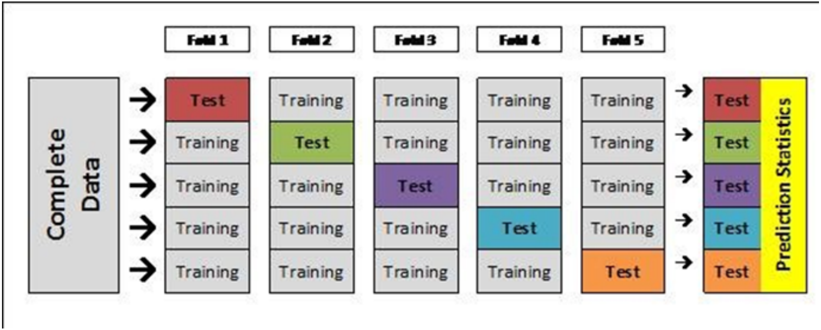
- Model is too complex, describes noise rather than signal (Bias-Variance trade-off)
- Focus on features that perform well in labeled data but may not generalize (e.g. unpopular hashtags)
- In-sample performance better than out-of-sample performance

- Solutions?
  - Randomly split dataset into training and test set
  - Cross-validation

Cross-validation

Intuition:

- Create K training and test sets ("folds") within training set.
- For each k in K, run classifier and estimate performance in test set within fold.
- Choose best classifier based on cross-validated performance



Example: Diversionary theory of foreign policy

(Sobek, 2007; Russett, 1990)

Mechanism: When domestic situation worsens, leaders will try to divert attention from problems and rally support to regime through international conflict

Empirical expectations:

- During episodes of social unrest...
- ...leaders will *increase* (1) attention to foreign policy, (2) use of nationalist rhetoric, (3) power projection, (4) overall social media activity

A new dataset

- Twitter and Facebook accounts of the heads of state and heads of government of all 193 U.N. member countries.
- Both institutional and personal accounts
- Both English-language accounts and own language
- Updated as of August 2016
- All Tweets and Facebook posts from Jan 1, 2012 to Jun 1, 2017, collected from public APIs
- Current total: 285,414 Facebook posts & 609,224 tweets
- Automated translation to English with Google Translate API

Supervised learning classification

- Stratified random sample of 4,749 unique social media posts coded by trained undergraduate students
  - 4 categories: domestic, foreign, personal, others
  - Total codings: 6,000 with ∼90% agreement
- Standard text pre-processing (removal of stopwords, urls, handles, digits, punctuation...)
- Train classifier using *xgboost* (Chen and Guestrin, 2016)

Category	Accuracy	Precision	Recall	Baseline
Domestic policy	0.722	0.654	0.633	38.8%
Foreign policy	0.782	0.671	0.644	31.2%
Personal	0.914	0.265	0.162	4.1%
Others	0.757	0.443	0.551	26.5%

Notes: *accuracy* is the % of social media posts correctly classified; *precision* is the % of posts predicted to be in that category that are correctly classified; *recall* is the % of posts in that category that are correctly classified; *baseline* is the proportion of posts in that category.

- Apply to full sample of social media posts

N-grams with highest feature importance, weighted by frequency

	Content type classifier
Domestic	of_the, to_the, government, national, education, approved, employment, school, health, of_our, knowledge, thanks, project, year, public, for_the, construction, celebrate, 2011, increase, civil, tune, arrival, social, the_national, do_not, society, system, young, billion, in_the, ministry_of, will_be, students, enjoy, chance, work, research, economy
Foreign	foreign, fm, meeting, countries, cooperation, visit, summit, relations, ambassador, meets, the_united, forum, china, eu, president, un, terrorism, turkey, the_european, geneva, met_with, nations, minister, condolences, bilateral, europe, consulate, cuba, ecuadorian, receives, press, relationship, attack, to_attend, embassy, partners, africa, delegation, poland, human, states
Personal	happy, wishes, book, thoughts, birthday, lhl, you_very, holiday, vanuatu, has_never, you_going, 2016, agreement_august, for_your, poem, always_remember, his_life, interesting, mount, missed, always_in, scholarships, malta, #newcare, nationality, busy_day, ny, condolances, my_deepest, rep, deepest_condolences, happy_king, apply, can_start

Predictors of rhetoric style

Table: OLS regression of content type proportion, at month level

	Domestic	Foreign
Constant	43.24***	46.14***
	(2.78)	(2.86)
Twitter (0-1)	−7.44***	−0.10
	(0.38)	(0.39)
GDP growth (%)	0.32***	−0.30***
	(0.07)	(0.07)
Unrest (log event count)	0.05	0.48**
	(0.19)	(0.20)
Democracy (0-1)	2.11***	−1.25***
	(0.45)	(0.46)
N	5,125	5,125
Adjusted R^2	0.24	

*p $<$ .1; **p $<$ .05; ***p $<$ .01

DVs: Month-level averages of predicted probabilities that social media post is about domestic/foreign policy (Models 1-2) or % of nationalist or need for power words (3-4)
Controls: GDPpc, content type (Models 3-4), account type, account actor, internet usage, population, region fixed effects

Types of classifiers

General thoughts:

- Trade-off between accuracy and interpretability
- Parameters need to be cross-validated

Frequently used classifiers:

- Naive Bayes
- Regularized regression
- SVM
- Others: k-nearest neighbors, tree-based methods, etc.
- Ensemble methods

Regularized regression

Assume we have:

- $i = 1, 2, \ldots, N$ documents
- Each document i is in class $y_i = 0$ or $y_i = 1$
- $j = 1, 2, \ldots, J$ unique features
- And x_{ij} as the count of feature j in document i

We could build a linear regression model as a classifier, using the values of β_0, β_1, ..., β_J that minimize:

$$RSS = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2$$

But can we?

- If $J > N$, OLS does not have a unique solution
- Even with $N > J$, OLS has low bias/high variance (overfitting)

Regularized regression

What can we do? Add a penalty for model complexity, such that we now minimize:

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{J}\beta_j^2 \rightarrow \text{ridge regression}$$

or

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{J}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{J}|\beta_j| \rightarrow \text{lasso regression}$$

where λ is the **penalty parameter** (to be estimated)

Regularized regression

Why the penalty (shrinkage)?

- Reduces the variance
- Identifies the model if J $>$ N
- Some coefficients become zero (feature selection)

The penalty can take different forms:

- Ridge regression: $\lambda \sum_{j=1}^{J} \beta_j^2$ with $\lambda > 0$; and when $\lambda = 0$ becomes OLS
- Lasso $\lambda \sum_{j=1}^{J} |\beta_j|$ where some coefficients become zero.
- Elastic Net: $\lambda_1 \sum_{j=1}^{J} \beta_j^2 + \lambda_2 \sum_{j=1}^{J} |\beta_j|$ (best of both worlds?)

How to find best value of λ? Cross-validation.
Evaluation: regularized regression is easy to interpret, but often outperformed by more complex methods.

Quantitative Text Analysis. Applications to Social Media Research

Pablo Barberá
London School of Economics
www.pablobarbera.com

Course website:

pablobarbera.com/text-analysis-vienna