Week 5: Supervised Modelling

POP77032 Quantitative Text Analysis for Social Scientists

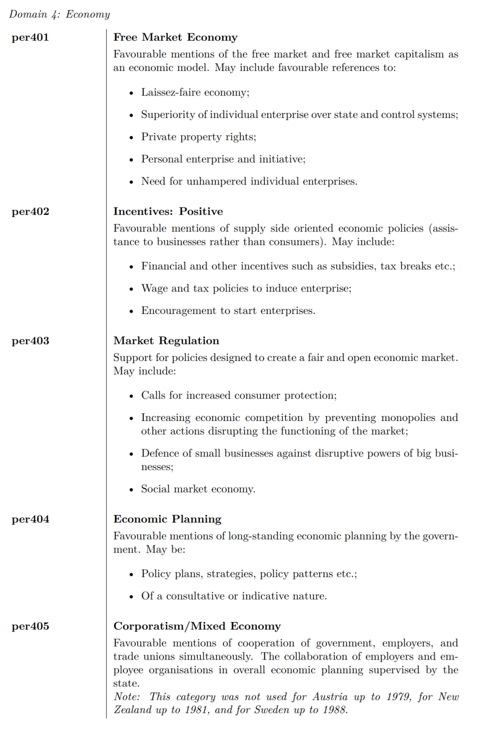
Tom Paskhalis

Overview

- Human Annotation
- Reliability Measures
- Training Set
- Feature Engineering

Human Annotation

Manifesto Project Codebook: Economy



Domain 4: Economy

per401 **Free Market Economy**

Favourable mentions of the free market and free market capitalism as an economic model. May include favourable references to:

- Laissez-faire economy;
- Superiority of individual enterprise over state and control systems;
- Private property rights;
- Personal enterprise and initiative;
- Need for unhampered individual enterprises.

per402 **Incentives: Positive**

Favourable mentions of supply side oriented economic policies (assistance to businesses rather than consumers). May include:

- Financial and other incentives such as subsidies, tax breaks etc.;
- Wage and tax policies to induce enterprise;
- Encouragement to start enterprises.

per403 **Market Regulation**

Support for policies designed to create a fair and open economic market. May include:

- Calls for increased consumer protection;
- Increasing economic competition by preventing monopolies and other actions disrupting the functioning of the market;
- Defence of small businesses against disruptive powers of big businesses;
- Social market economy.

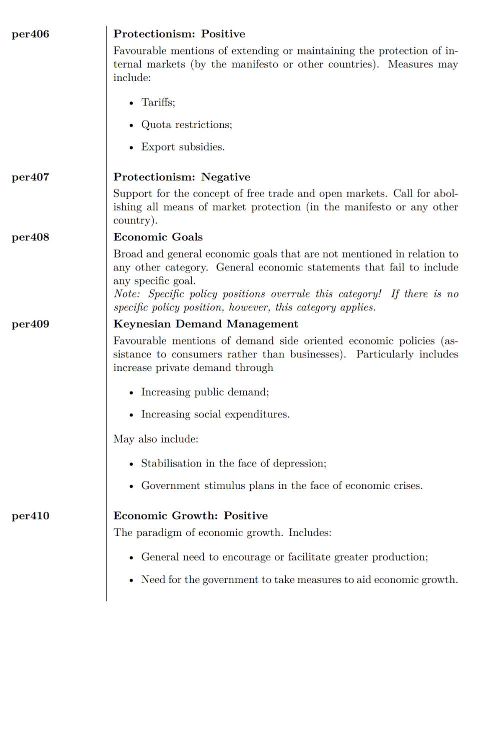
per404 **Economic Planning**

Favourable mentions of long-standing economic planning by the government. May be:

- Policy plans, strategies, policy patterns etc.;
- Of a consultative or indicative nature.

per405 **Corporatism/Mixed Economy**

Favourable mentions of cooperation of government, employers, and trade unions simultaneously. The collaboration of employers and employee organisations in overall economic planning supervised by the state.
Note: This category was not used for Austria up to 1979, for New Zealand up to 1981, and for Sweden up to 1988.



per406 **Protectionism: Positive**

Favourable mentions of extending or maintaining the protection of internal markets (by the manifesto or other countries). Measures may include:

- Tariffs;
- Quota restrictions;
- Export subsidies.

per407 **Protectionism: Negative**

Support for the concept of free trade and open markets. Call for abolishing all means of market protection (in the manifesto or any other country).

per408 **Economic Goals**

Broad and general economic goals that are not mentioned in relation to any other category. General economic statements that fail to include any specific goal.
Note: Specific policy positions overrule this category! If there is no specific policy position, however, this category applies.

per409 **Keynesian Demand Management**

Favourable mentions of demand side oriented economic policies (assistance to consumers rather than businesses). Particularly includes increase private demand through

- Increasing public demand;
- Increasing social expenditures.

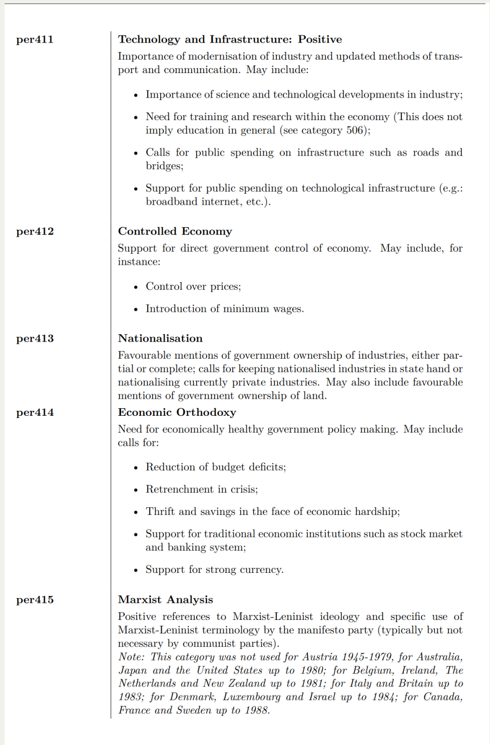
May also include:

- Stabilisation in the face of depression;
- Government stimulus plans in the face of economic crises.

per410 **Economic Growth: Positive**

The paradigm of economic growth. Includes:

- General need to encourage or facilitate greater production;
- Need for the government to take measures to aid economic growth.



per411 **Technology and Infrastructure: Positive**

Importance of modernisation of industry and updated methods of transport and communication. May include:

- Importance of science and technological developments in industry;
- Need for training and research within the economy (This does not imply education in general (see category 506);
- Calls for public spending on infrastructure such as roads and bridges;
- Support for public spending on technological infrastructure (e.g.: broadband internet, etc.).

per412 **Controlled Economy**

Support for direct government control of economy. May include, for instance:

- Control over prices;
- Introduction of minimum wages.

per413 **Nationalisation**

Favourable mentions of government ownership of industries, either partial or complete; calls for keeping nationalised industries in state hand or nationalising currently private industries. May also include favourable mentions of government ownership of land.

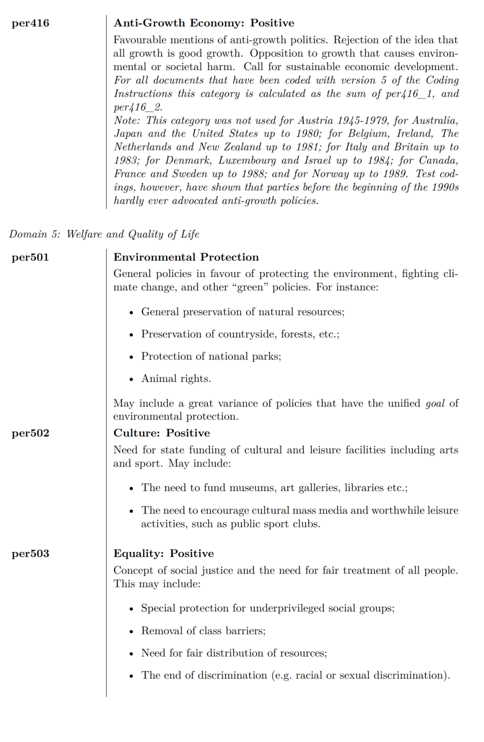
per414 **Economic Orthodoxy**

Need for economically healthy government policy making. May include calls for:

- Reduction of budget deficits;
- Retrenchment in crisis;
- Thrift and savings in the face of economic hardship;
- Support for traditional economic institutions such as stock market and banking system;
- Support for strong currency.

per415 **Marxist Analysis**

Positive references to Marxist-Leninist ideology and specific use of Marxist-Leninist terminology by the manifesto party (typically but not necessary by communist parties).
Note: This category was not used for Austria 1945-1979, for Australia, Japan and the United States up to 1980; for Belgium, Ireland, The Netherlands and New Zealand up to 1981; for Italy and Britain up to 1983; for Denmark, Luxembourg and Israel up to 1984; for Canada, France and Sweden up to 1988.



per416 **Anti-Growth Economy: Positive**

Favourable mentions of anti-growth politics. Rejection of the idea that all growth is good growth. Opposition to growth that causes environmental or societal harm. Call for sustainable economic development.
For all documents that have been coded with version 5 of the Coding Instructions this category is calculated as the sum of per416_1, and per416_2.
Note: This category was not used for Austria 1945-1979, for Australia, Japan and the United States up to 1980; for Belgium, Ireland, The Netherlands and New Zealand up to 1981; for Italy and Britain up to 1983; for Denmark, Luxembourg and Israel up to 1984; for Canada, France and Sweden up to 1988; and for Norway up to 1989. Test codings, however, have shown that parties before the beginning of the 1990s hardly ever advocated anti-growth policies.

Domain 5: Welfare and Quality of Life

per501 **Environmental Protection**

General policies in favour of protecting the environment, fighting climate change, and other “green” policies. For instance:

- General preservation of natural resources;
- Preservation of countryside, forests, etc.;
- Protection of national parks;
- Animal rights.

May include a great variance of policies that have the unified *goal* of environmental protection.

per502 **Culture: Positive**

Need for state funding of cultural and leisure facilities including arts and sport. May include:

- The need to fund museums, art galleries, libraries etc.;
- The need to encourage cultural mass media and worthwhile leisure activities, such as public sport clubs.

per503 **Equality: Positive**

Concept of social justice and the need for fair treatment of all people. This may include:

- Special protection for underprivileged social groups;
- Removal of class barriers;
- Need for fair distribution of resources;
- The end of discrimination (e.g. racial or sexual discrimination).

(Manifesto Project, 2025)

What is the Policy Category?

What is the Policy Category?

What is the Policy Category?

What is the Policy Category?

Wisdom of Crowds

(Thomas Barker, Amgueddfa Cymru – Museum Wales)

Vox Populi

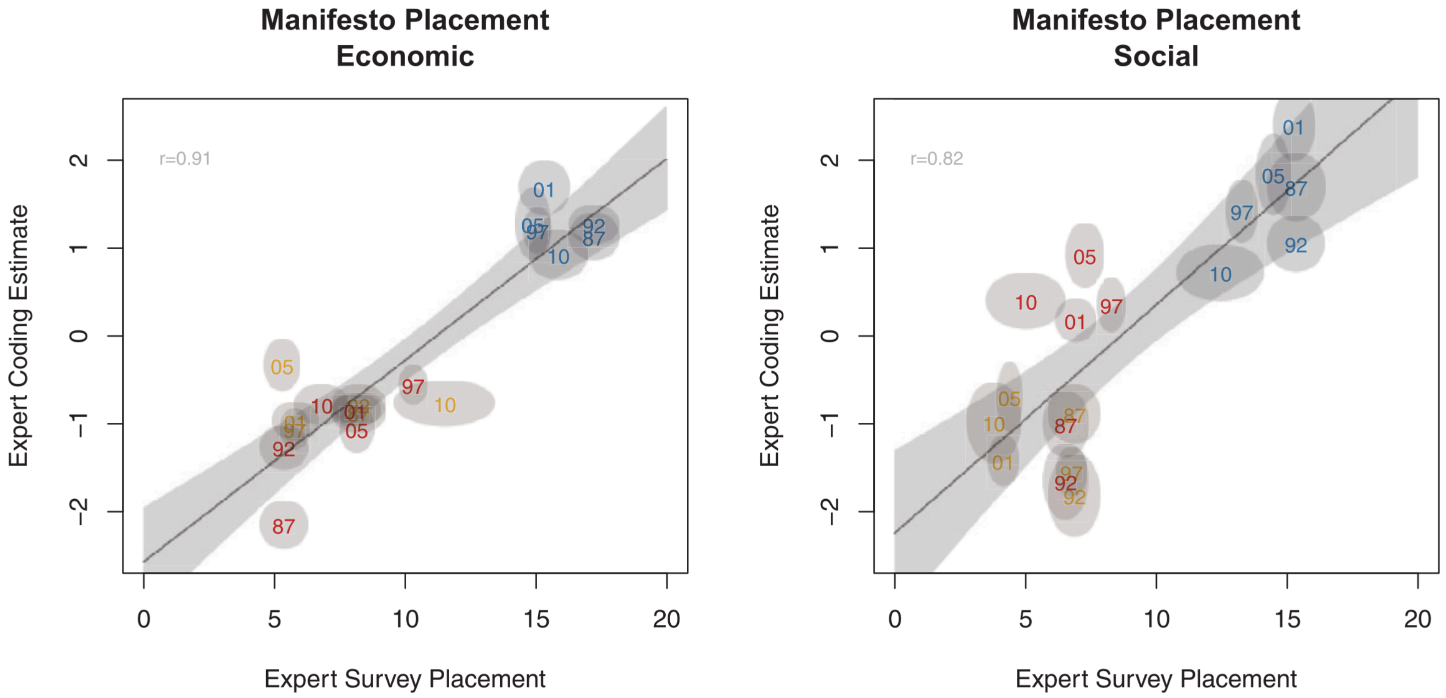
> A weight-judging competition was carried on at the annual show of the West of England Fat Stock and Poultry Exhibition recently held at Plymouth. A fat ox having been selected, competitors bought stamped and numbered cards, for 6d. each, on which to inscribe their respective names, addresses, and **estimates of what the ox would weigh** after it had been slaughtered and "dressed." Those who guessed most successfully received prizes. About 800 tickets were issued, which were kindly lent me for examination after they had fulfilled their immediate purpose. These afforded excellent material. The judgments were unbiassed by passion and uninfluenced by oratory and the like. The sixpenny fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best. The competitors included butchers and farmers, some of whom were highly expert in judging the weight of cattle; others were probably guided by such information as they might pick up, and by their own fancies. The average competitor was probably as well fitted for making a just estimate of the dressed weight of the ox, as an average voter is of judging the merits of most political issues on which he votes, and the variety among the voters to judge justly was probably much the same in either case.
>
> Francis Galton (1907)

> According to the democratic principle of "one vote one value," the middlemost estimate expresses the *vox populi*, every other estimate being condemned as too low or too high by a majority of the voters […]. Now the middlemost estimate is 1207 lb., and the weight of the dressed ox proved to be 1198 lb.; **so the vox populi was in this case 9 lb., or 0.8 per cent. of the whole weight too high.**
>
> Francis Galton (1907)

Wisdom of Crowds in Text Annotation

- Benoit et al. (2016) use crowd-source platform to compare 200K+ of crowd-sourced annotations of sentences in UK party manifestos 1987-2010 to 100K+ expert evaluations.

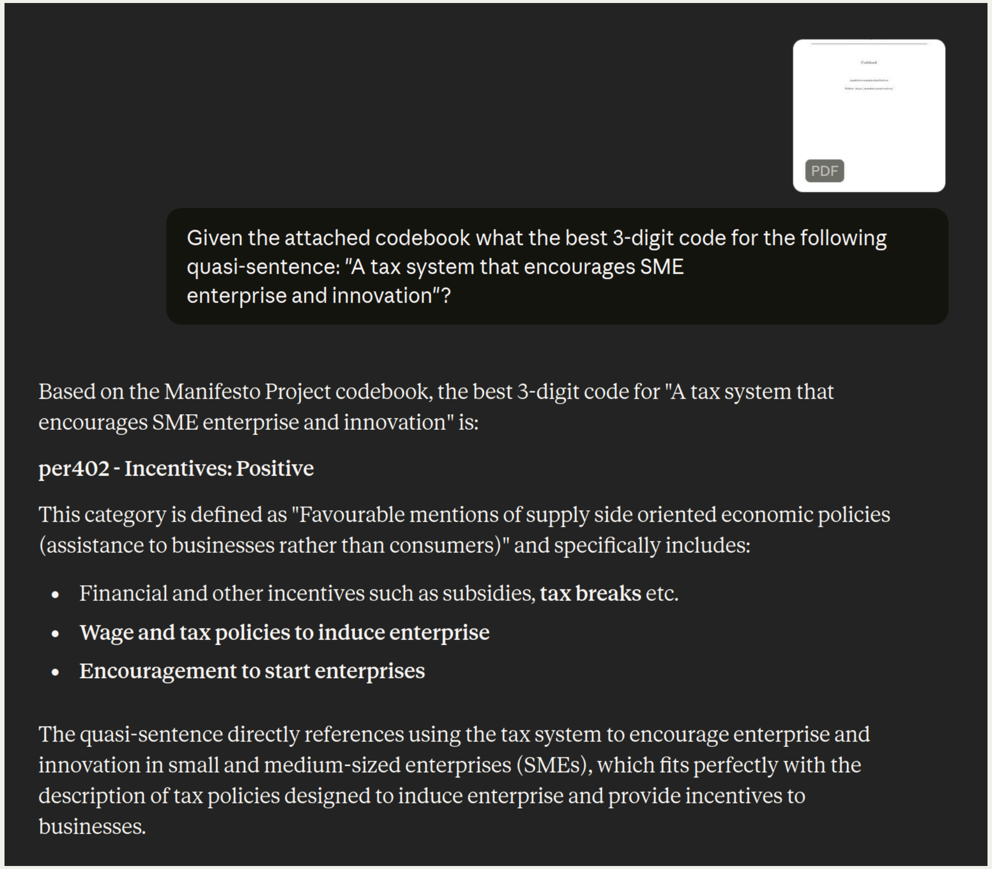
FIGURE 2. British Party Positions on Economic and Social Policy 1987–2010



Notes: Sequential expert text processing (vertical axis) and independent expert surveys (horizontal). Labour red, Conservatives blue, Liberal Democrats yellow, labeled by last two digits of year.

(Benoit et al., 2016)

Could we just use an LLM?



Given the attached codebook what the best 3-digit code for the following quasi-sentence: "A tax system that encourages SME enterprise and innovation"?

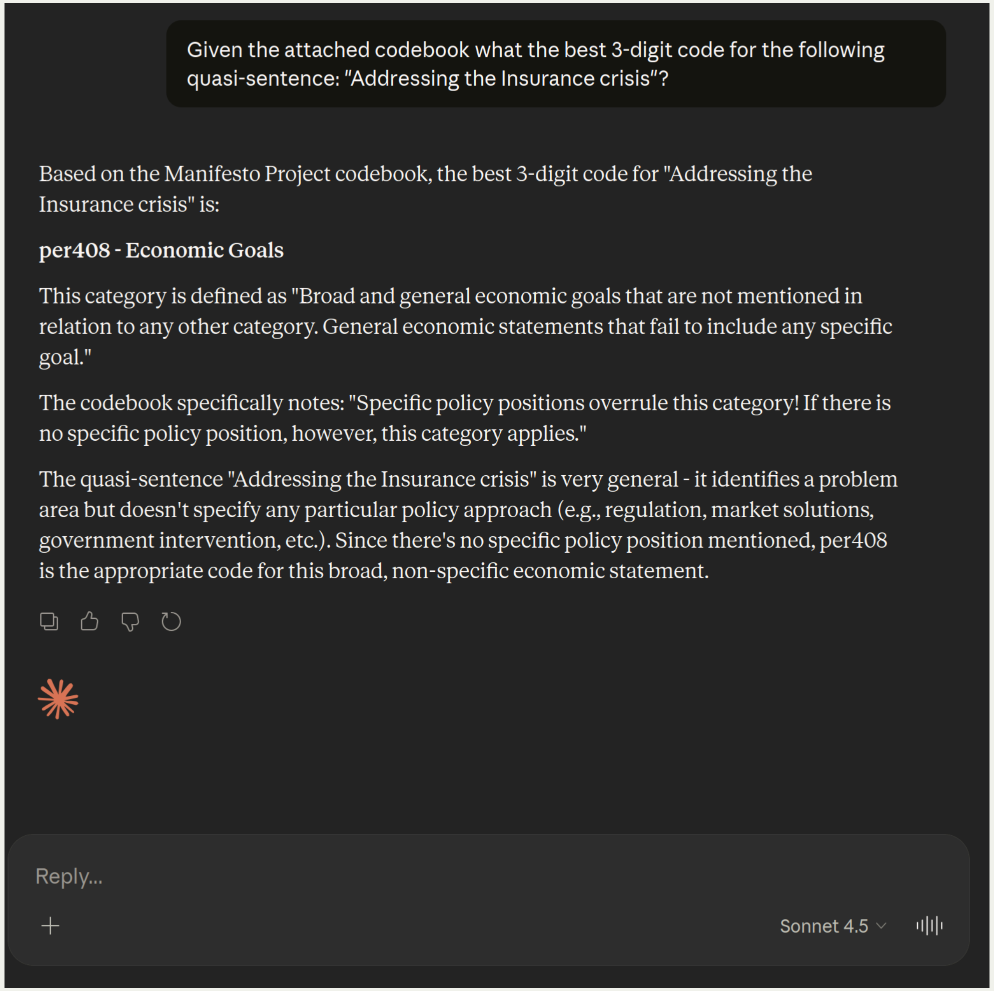
Based on the Manifesto Project codebook, the best 3-digit code for "A tax system that encourages SME enterprise and innovation" is:

per402 - Incentives: Positive

This category is defined as "Favourable mentions of supply side oriented economic policies (assistance to businesses rather than consumers)" and specifically includes:

- Financial and other incentives such as subsidies, **tax breaks** etc.
- **Wage and tax policies to induce enterprise**
- **Encouragement to start enterprises**

The quasi-sentence directly references using the tax system to encourage enterprise and innovation in small and medium-sized enterprises (SMEs), which fits perfectly with the description of tax policies designed to induce enterprise and provide incentives to businesses.



Given the attached codebook what the best 3-digit code for the following quasi-sentence: "Addressing the Insurance crisis"?

Based on the Manifesto Project codebook, the best 3-digit code for "Addressing the Insurance crisis" is:

per408 - Economic Goals

This category is defined as "Broad and general economic goals that are not mentioned in relation to any other category. General economic statements that fail to include any specific goal."

The codebook specifically notes: "Specific policy positions overrule this category! If there is no specific policy position, however, this category applies."

The quasi-sentence "Addressing the Insurance crisis" is very general - it identifies a problem area but doesn't specify any particular policy approach (e.g., regulation, market solutions, government intervention, etc.). Since there's no specific policy position mentioned, per408 is the appropriate code for this broad, non-specific economic statement.

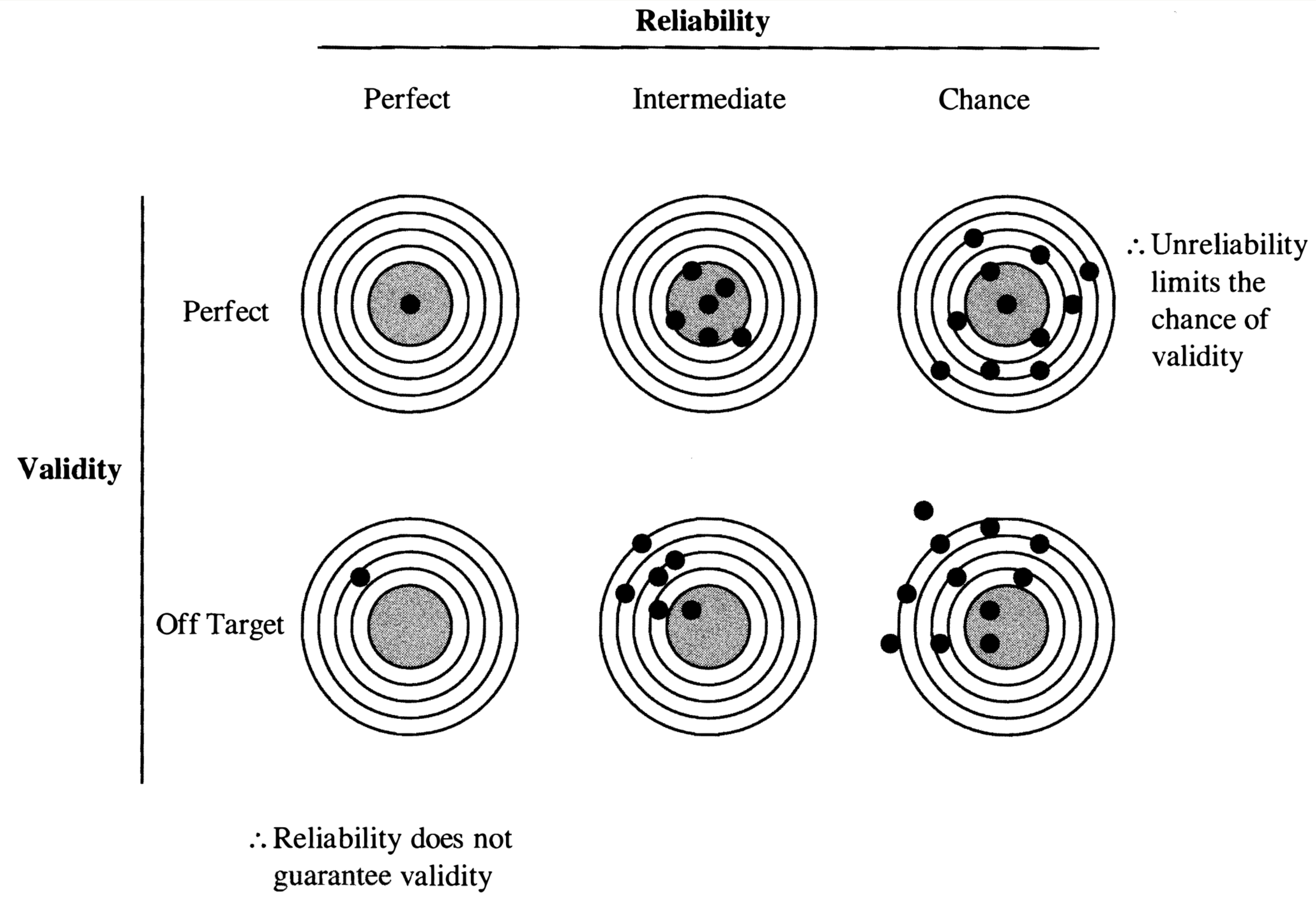
Sonnet 4.5

- Probably, but similar validation standards need to be applied.

Human Annotation

- Any concept measured from text with no obvious quantitative benchmark (e.g. economic performance) needs to be extensively validated.
- In social sciences this is typically done through human annotation or expert coding.
- One should aim to have multiple annotators/experts labelling the same unit of analysis (sentences, speeches, etc.)
- The units chosen for annotation should relatively small (e.g. sentences, paragraphs) as human attention span is limited.
- In addition, focussing on smaller units helps with making the task easier to scale.

Reliability vs Validity


Reliability
Perfect
Intermediate
Chance
Validity
Perfect
Off Target
∴ Unreliability limits the chance of validity
∴ Reliability does not guarantee validity

(Krippendorff, 2019)

Evaluating Human Annotation

- When assessing the quality of human annotations, several aspects should be considered:

Type	Test Design	Cause of Disagreement
Stability	rest-retest	intraobserver inconsistencies
Reliability	test-test	interobserver disagreements
Validity	test-standard	deviations from a standard

Measures of Agreement

- **Percent agreement**: $\frac{\text{number of agreeing annotations}}{\text{total number of annotations}} \times 100\%$
- **Correlation**: Pearson's r for continuous scales or Spearman's ρ for ordinal.
- Measures of agreement:
  - Account not not only for agreement but also for the possibility of agreement occurring by chance.
  - **Cohen's** κ - most common

    $$\kappa = \frac{P_o - P_e}{1 - P_e}$$

    where P_o is the observed agreement and P_e is the expected agreement by chance.
  - **Krippendorff's** α - a generalisation of Cohen's κ to multiple annotators and different scales.

    $$\alpha = 1 - \frac{D_o}{D_e}$$

    where D_o is the observed disagreement and D_e is the expected disagreement by chance.

Reliability Data Matrix

- A canonical representation of inter-coder agreement is a **reliability data matrix**.
- Typically, it is arranged with annotators in rows and units in column (see Krippendorff for more details).
- In practice, you can always work with a transpose of this.
- In the simplest case we have 2 annotators labelling binary data:

Unit	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Coder 1	1	1	0	0	0	0	0	0	0	0
Coder 2	0	1	1	0	0	1	0	1	0	0

Extra

Ch 12 Reliability Krippendorff (2019)

Calculating Reliability

```r
coder1 <- c(1, 1, 0, 0, 0, 0, 0, 0, 0, 0)
coder2 <- c(0, 1, 1, 0, 0, 1, 0, 1, 0, 0)
```

- Coder 1 and Coder 2 agree on 6 out of 10 units, so the percent agreement is 60%, which might look reasonable at first glance.

```r
sum(coder1 == coder2)
```

```
[1] 6
```

```r
sum(coder1 == coder2)/length(coder1) * 100
```

```
[1] 60
```

- There are 4 observed disagreements, with coincidence matrix:

```r
table(coder2, coder1) + table(coder1, coder2)
```

```
      coder1
coder2  0  1
     0 10  4
     1  4  2
```

- Note that unlike contingency tables, coincidence matrices are symmetrical around the diagonal.

Krippendorff’s α

- Looking back at our coincidence matrix:

	0	1	Total
0	10	4	14
1	4	2	6
Total	14	6	20

- Had all annotations been made randomly, we would expect to observe the following coincidence matrix:

	0	1	Total
0	9.6	4.4	14
1	4.4	1.6	6
Total	14	6	20

where $e_{01} = e_{10} = n_0 \times n_1 / (n - 1)$

Thus, Krippendorff's $\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{4}{4.4211} = 0.095$, which is quite low.

Calculating Reliability in R

- In practice, we would just use software to calculate reliability.
- E.g. in R we could simply the `irr` package:

```r
library("irr")
irr::kripp.alpha(rbind(coder1, coder2), method = "nominal")
```

```
Krippendorff's alpha

Subjects = 10
  Raters = 2
   alpha = 0.0952
```

- For comparison, we could also calculate Cohen's κ:

```r
irr::kappa2(data.frame(coder1, coder2), weight = "unweighted")
```

```
Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 10
  Raters = 2
   Kappa = 0.0909

       z = 0.323
 p-value = 0.747
```

Supervised Learning

Supervised vs Unsupervised

Unsupervised modelling:

learning latent structure from **unlabelled** data

E.g. principal component analysis of a DTM

Supervised modelling:

learning a relationship between inputs and **labelled** data

E.g. sentiment analysis using a training set of positive and negative reviews

Dictionaries and Supervised Learning

- *Dictionary* could be considered a certain form of *supervised* learning.
- Association between a feature and a category is based on reading of text(s).
- It can be done either by human(s) or by machine(s).
- But the texts used to build a dictionary are often different from those to which it is applied.
- With more traditional *supervised* learning techniques, the association between a feature and a category is derived from data.

Dictionaries and Supervised Learning: Performance

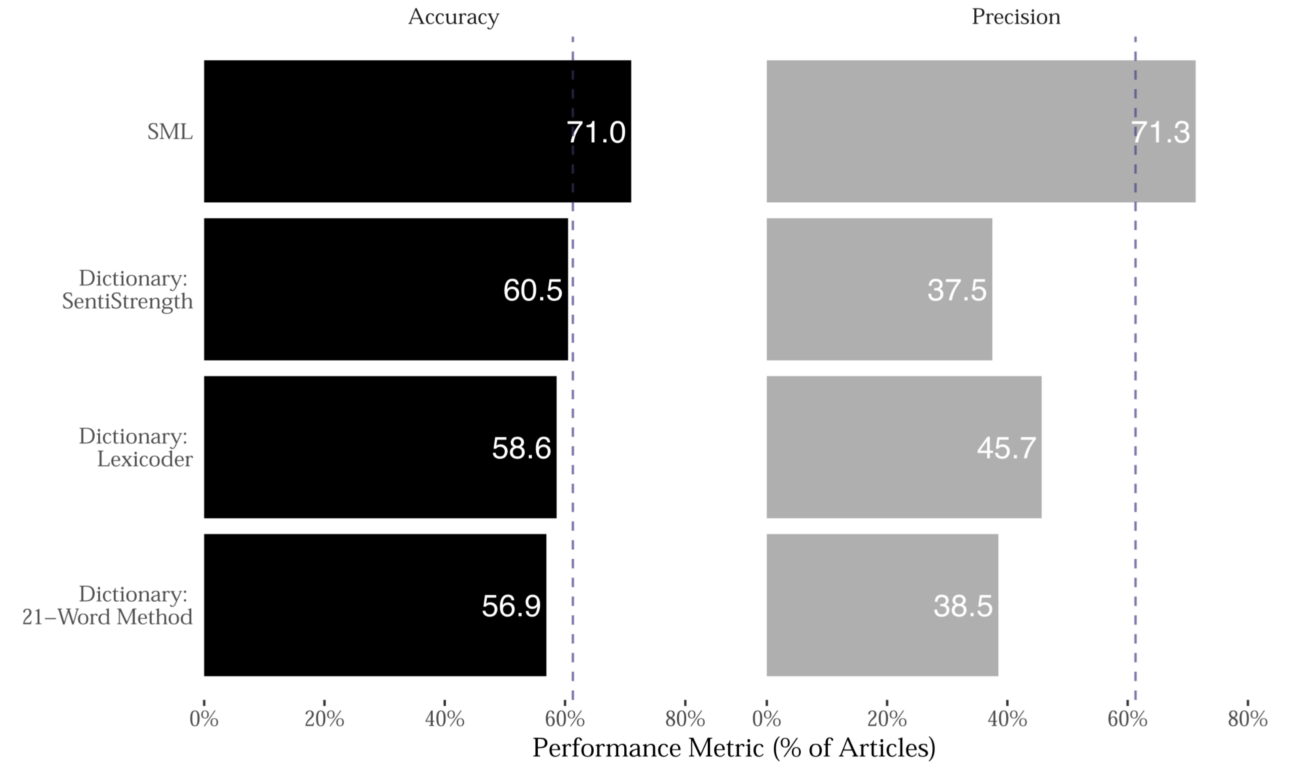


Figure 3. Performance of SML and Dictionary Classifiers—Accuracy and Precision.
Note: Accuracy (percent correctly classified) and precision (percent of positive predictions that are correct) for the ground truth dataset coded by ten CrowdFlower coders. The dashed vertical lines indicate the baseline level of accuracy or precision (on any category) if the modal category is always predicted. The corpus used in the analysis is based on the keyword search of *The New York Times* 1980–2011 (see the text for details).

(Barberá, Boydstun, Linn, McMahon & Nagler, 2021)

Basic Principles of Supervised Learning

- We have some labelled data that we can use to develop a classifier.
- The data is split into:
  - **training set** for the classifier to “learn” relationships between features and outcome
  - **validation/development** set for tuning and adjusting any hyperparameters
  - **test set** for evaluating the performance of the classifier
- The idea is to build a classifier that can generalise to previously unseen samples.

Evaluating Classifier

Predicted / True	**Positive**	**Negative**
Positive	TP	FP
Negative	FN	TN

- Accuracy: $\frac{TP+TN}{TP+FP+FN+TN}$
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F1 Score (harmonic mean of precision and recall): $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Naive Bayes

Naive Bayes Classification

- Motivation: We want to classify a document into one of several categories based on its features (e.g. words).
- **Naive Bayes** is a simple probabilistic classifier based on Bayes' theorem with the "naive" assumption of independence between features in a document.
- It is fast, simple, and often performs well in text classification tasks.

Bayes' Theorem

- Recall how conditional probability works:

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

- E.g. When throwing a die, the probability of getting a 2 given that we got an even number is $P(2|\text{even}) = \frac{1/6}{1/2} = \frac{1}{3}$
- Of course, it is also true that $P(B|A) = \frac{P(B,A)}{P(A)}$
- But since $P(A, B) = P(B, A)$, it must be that $P(A|B)P(B) = P(B|A)P(A)$ and thus we have Bayes' theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Naive Bayes Setup

- E.g. we are interested in a document containing positive or negative sentiment.
- Consider J features distributed across N documents each assigned to one of K classes.
- Using Bayes' Theorem at the word level we could express it as:

$$P(c_k|w_j) = \frac{P(c_k)P(w_j|c_k)}{P(w_j)}$$

- For 2 classes we could write it as:

$$P(c_k|w_j) = \frac{P(c_k)P(w_j|c_k)}{P(c_k)P(w_j|c_k) + P(c_{\neg k})P(w_j|c_{\neg k})}$$

where $c_{\neg k}$ is the class alternative to class c_k.

Naive Bayes: Word Likelihoods

$$P(c_k|w_j) = \frac{P(c_k)P(w_j|c_k)}{P(c_k)P(w_j|c_k) + P(c_{\neg k})P(w_j|c_{\neg k})}$$

- The term $P(w_j|c_k)$ is **word likelihood** conditional on class.
- The MLE estimate for this is simply the proportion of times the work j occurs in class k, but it more common to use *Laplace smoothing* by adding 1 to each observed count within class to avoid zero probabilities.
- In practice, since $P(c_k)P(w_j|c_k) + P(c_{\neg k})P(w_j|c_{\neg k}) = P(w_j)$ which is the same for all classes, the denominator isn't needed for classification.

Naive Bayes: Prior Probabilities

$$P(c_k|w_j) = \frac{P(c_k)P(w_j|c_k)}{P(c_k)P(w_j|c_k) + P(c_{\neg k})P(w_j|c_{\neg k})}$$

- The terms $P(c_k)$ and $P(c_{\neg k})$ are the **class prior probabilities**.
- In supervised learning, these are typically estimated from the training data as the proportion of documents in each class.

Naive Bayes: Posterior Probabilities

$$P(c_k|w_j) = \frac{P(c_k)P(w_j|c_k)}{P(c_k)P(w_j|c_k) + P(c_{\neg k})P(w_j|c_{\neg k})}$$

- The term $P(c_k|w_j)$ is the **posterior probability** of class c_k given the presence of word w_j.
- Which is, fundamentally, our quantity of interest.

Naive Bayes: Documents

- Of course, in practice we would like to use more than a single feature w_j to predict class membership.
- The "naive" assumption of Naive Bayes is that features are conditionally independent

$$P(c_k|d_i) = P(c_k) \prod_{j=1}^{J} \frac{P(w_j|c_k)}{P(w_j)}$$

- It is naive because it (wrongly) assumes:
  - *conditional independence* of feature counts
  - *positional independence* of features in a document

Next

- Tutorial: Supervised Modelling
- Next week: Unsupervised modelling
- Assignment 2: Due 15:59 on Wednesday, 4th March (submission on Blackboard)