








Modelling Financial Markets

Seminar 15: LLMs and Sentiment Analysis — Hands-on



Daniel Traian PELE
Drd. Antoaneta Amza, antoaneta.amza@csie.ase.ro

Bucharest University of Economic Studies
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Romanian Academy, Institute for Economic Forecasting
MSCA Digital Finance

Today's Questions and Seminar Structure

- Questions of the seminar
  - **can a language model read financial sentiment?**
  - **does that sentiment predict returns one can still trade?**
- Models and data
  - LM: the Loughran–McDonald dictionary, which counts financial words without context
  - FinBERT: BERT (Bidirectional Encoder Representations from Transformers) adapted to financial text; Qwen2.5: an LLM (Large Language Model) asked zero-shot
  - FNSPID (Financial News and Stock Price Integration Dataset): headlines linked to stock prices; SPY (SPDR S&P 500 ETF Trust): the ETF (Exchange-Traded Fund) on the S&P 500 (Standard & Poor's 500) index, benchmark of the excess returns
- Deliverables
  - the answers to the A exercises
  - one table and one figure per B exercise
  - the C1 corrections
  - one interpretation sentence per exercise
- Notebook for today: open the seminar notebook in Colab
  - each task names its notebook section
- Seminar: Prof. Daniel Traian Pele, danpele@ase.ro; Antoaneta Amza, PhD candidate, antoaneta.amza@csie.ase.ro

Exercise Map

Exercise	Question	Type	Model
A1	how much does a misclassified dummy shrink a slope?	Solved	–
A2	how precise, and how powerful, is a paired accuracy test?	Solved	–
A3	at what cost per trade does the strategy stop earning?	Solved	–
B1	is FinBERT more accurate than the Loughran–McDonald dictionary?	Solved	–
B2	do headlines predict returns, or describe them?	Solved	–
B3	does the LLM remember the market?	Solved	–
C1	what is wrong in an AI (Artificial Intelligence) answer?	Proposed	B1, B2, B3
B4	is a 7-billion-parameter LLM better than FinBERT? (homework)	Proposed	B1
Extension packs	A4–A8, B5–B11, C2, C3 (optional, after the lecture)	both	named on each task

- **[Solved]**: full solution in the slides and in the notebook, a model to copy; **[Proposed]**: you solve it, the solution is discussed in class. Do B4 before using the discordant counts supplied in A6
- Part A: derivations on paper; Part B: verification and inference on real data, each ending with an interpretation question; Part C: open questions and the critique of an AI (Artificial Intelligence) answer

Data Used

Data	Content	Saved scores	Used in
Twitter Financial News (validation set, Hugging Face Hub)	2 388 headlines from 2022: 1 566 neutral, 475 positive, 347 negative	LM, FinBERT, Qwen2.5 (P1–P3), MiniLM	A1, A2, A4, A6–A8, B1, B4–B8, C1
Financial PhraseBank Malo et al. (2014)	4 846 news sentences, 2 264 with full agreement	same scores	B5
FNSPID Dong, Fan & Peng (2024)	106 820 headlines on 16 US stocks, 27.10.2009–18.12.2023	daily LM, FinBERT, Qwen2.5-7B scores	A3, A8, B2, B9–B11, C1–C3
Stocks and SPY	adjusted daily prices of the course data, one series per asset (field used: adjusted close), 2009–2023	–	B2, B11, C2, C3
S&P 500	monthly closes, Jan 2000 – Aug 2026	Qwen2.5 answers about each month	B3, C1

- The scores are saved with the chapter's Quantlets, `Quantlets/Ch_15`: no language model has to run in class
- C3 (extension pack) needs a new scoring run of the FNSPID headlines

Setup: Run and Verify

```
# notebook section "Setup" defines the functions that
      read the data
pb, tw = eval_texts()    # PhraseBank, Twitter
P, rets = news_panel()   # FNSPID stock-days
print(len(pb), len(tw))
print(tw['label'].value_counts())
print(P.shape,
      P.index.get_level_values('ticker').nunique())
```

- Expected output
  - 4 846 and 2 388 rows
  - neutral 1 566, positive 475, negative 347
  - (18 282, 19) and 16 stocks
- Each row of `tw` holds the text, the human label and the probabilities of every model (e.g. `finbert_negative`, `qwen7B_positive`)
- If your counts differ, stop and check the files before any exercise

Notation Used in Every Exercise

- Dictionary tone τ and the Loughran–McDonald (LM) lists Loughran & McDonald (2011)
- Model probabilities p_k, label $\hat{y}$ and net score $s = p_{\mathbf{pos}} - p_{\mathbf{neg}}$ (FinBERT Araci (2019), Qwen2.5 Qwen Team (2024))
- Two classifiers A and B on the same texts: discordant counts $n_{\mathbf{01}}$, $n_{\mathbf{10}}$ and the McNemar test McNemar (1947)
- Day d, horizon h and the first tradable return $d + 2$
- Units and inference
  - bp: basis point, 0.01%; SE: standard error; CI: 95% confidence interval
  - t_{NW}: Newey–West t Newey & West (1987); clustered SE Petersen (2009)

Three Ways to Read Sentiment

- **Dictionary** (bag of words) Loughran & McDonald (2011): the tone of a text

$$\tau = \frac{P - N}{P + N} \in [-1, 1]$$

  - P, N: numbers of words of the text found in a fixed positive and a fixed negative list
  - Class = sign of τ ($\tau = 0$: neutral); example $P = 2$, $N = 1$: $\tau = 1/3$, positive
  - LM lists, built from 10-K filings: 2 355 negative and only 354 positive words; transparent, but blind to context and negation
- **FinBERT** Araci (2019): BERT, a Transformer encoder (Chapter 14), trained further on financial news
  - Fine-tuned (trained further) on labelled sentences; output: $p_{\mathbf{pos}}$, $p_{\mathbf{neu}}$, $p_{\mathbf{neg}}$, summing to 1
  - The fine-tuning sentences come from Financial PhraseBank Malo et al. (2014): its accuracy there is partly in-sample (B5)
- **LLM zero-shot** (Qwen2.5 Qwen Team (2024)): a large language model asked "Is this headline positive, negative or neutral?"
  - No training on these labels (zero-shot); we read the logits z_k of the three answer words (next slide)

From Logits to Probabilities

- Logit z_k: the model's unnormalised score of answer $k \in \{\text{pos}, \text{neu}, \text{neg}\}$; larger means more likely
- **Softmax with temperature** $T > 0$: turns logits into probabilities

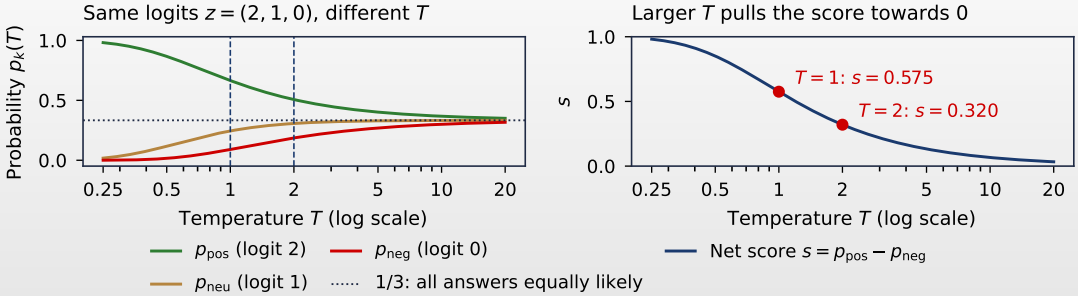
$$p_k(T) = \frac{e^{z_k/T}}{\sum_j e^{z_j/T}}, \qquad \sum_k p_k(T) = 1$$

  - $T = 1$: the model's own probabilities; $T > 1$ flattens them towards $1/3$; $T < 1$ sharpens them
  - Only differences of logits matter: $z_k - z_j = T \ln(p_k/p_j)$, the logit gap (A7)
- **Label** and **net score**

$$\hat{y} = \arg\max_k p_k, \qquad s = p_{\mathbf{pos}} - p_{\mathbf{neg}} \in [-1, 1]$$

  - $\arg\max_k$: the answer with the largest probability; T does not change it, since dividing by $T > 0$ keeps the order of the logits
  - $s > 0$: positive tone; s near 0: neutral or undecided; s changes with T

The Temperature at Work



- Example: logits $z = (2, 1, 0)$ for (pos, neu, neg)
  - $T = 1$: $p = (0.665;\ 0.245;\ 0.090)$, $s = 0.575$
  - $T = 2$: $p = (0.506;\ 0.307;\ 0.186)$, $s = 0.320$: same label, score pulled towards 0
- As $T \to \infty$ all three probabilities tend to $1/3$ and $s \to 0$; a threshold rule on s therefore depends on T

Scoring a Classifier

- Confusion matrix C: C_{jk} = number of texts with human label j (row, the truth) and predicted label k (column); n texts, K classes
- **Accuracy**: share of texts on the diagonal

$$\text{acc} = \frac{1}{n} \sum_{k=\mathbf{1}}^{K} C_{kk}$$

  - Benchmark: the share of the majority class (always predict it), never $1/K$
- **Precision** and **recall** of class k

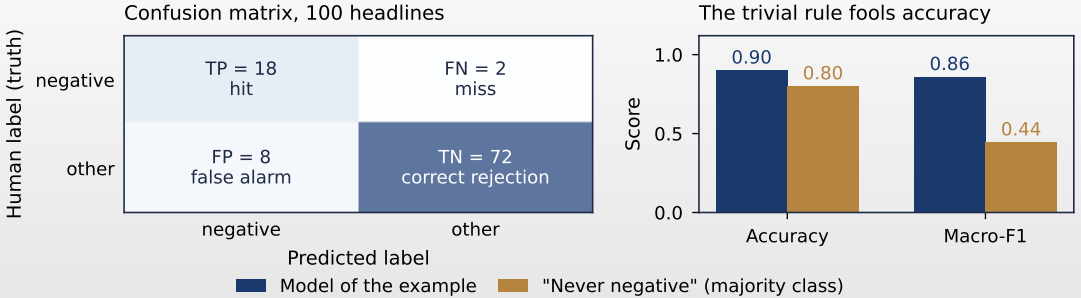
$$\text{prec}_k = \frac{C_{kk}}{\sum_j C_{jk}}, \qquad \text{rec}_k = \frac{C_{kk}}{\sum_j C_{kj}}$$

  - prec: share of the texts flagged k that truly are k (column); rec: share of the true k that are flagged (row)
- $F_{\mathbf{1}}$ and **macro-F1**

$$F_{\mathbf{1},k} = \frac{2\,\text{prec}_k\,\text{rec}_k}{\text{prec}_k + \text{rec}_k}, \qquad \text{macro-F1} = \frac{1}{K} \sum_{k=\mathbf{1}}^{K} F_{\mathbf{1},k}$$

  - $F_{\mathbf{1}}$: harmonic mean, low if either is low; macro-F1 gives a rare class the same weight as a frequent one

Accuracy, Macro-F1 and the Majority Class



- Example: 100 headlines, 20 truly negative; the model flags 18 of them and 8 of the other 80
  - Accuracy $(18 + 72)/100 = 0.90$, but "never negative" already scores 0.80; macro-F1: $(0.783 + 0.935)/2 = 0.86$ against $(0 + 0.889)/2 = 0.44$
  - Rates used in A1: $\pi = 0.20$ truly negative, $p = 0.26$ flagged, misses $\alpha_1 = 2/20 = 0.10$, false alarms $\alpha_0 = 8/80 = 0.10$

Measurement Error and Attenuation

- Ordinary Least Squares (OLS) slope on one regressor x: in large samples a ratio of moments

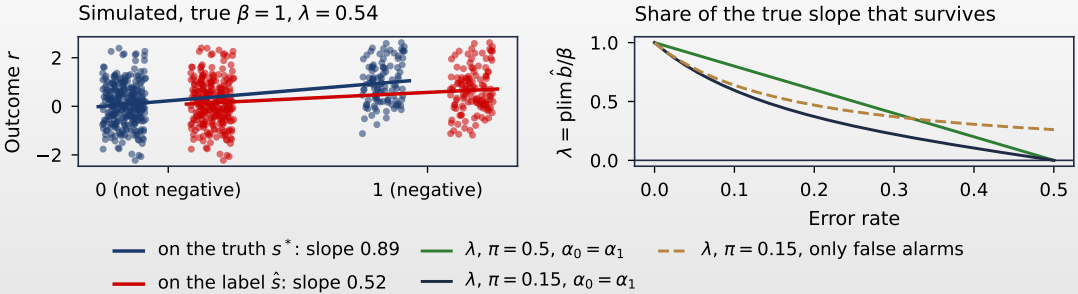
$$\text{plim}_{n\to\infty}\,\hat{b} = \frac{\text{Cov}(r,x)}{\text{Var}(x)}$$

  - plim: limit in probability; r: the outcome (a return)
- **Classical** error: we observe $x = x^* + u$, u independent of x^* and of ε

$$r = a + \beta x^* + \varepsilon \;\Rightarrow\; \text{plim}\,\hat{b} = \lambda\beta, \qquad \lambda = \frac{\text{Var}(x^*)}{\text{Var}(x^*) + \text{Var}(u)}$$

  - $\lambda \in (0,1]$: the slope is biased towards zero (attenuation); $1/\lambda$ undoes it
- **Misclassified dummy** Aigner (1973): label $\hat{s} \in \{0,1\}$ for a true $s^* \in \{0,1\}$
  - $\pi = P(s^* = 1)$; false alarms $\alpha_{\mathbf{0}} = P(\hat{s} = 1 \mid s^* = 0)$; misses $\alpha_{\mathbf{1}} = P(\hat{s} = 0 \mid s^* = 1)$
  - Total probability: $p = P(\hat{s} = 1) = \pi(1 - \alpha_{\mathbf{1}}) + (1 - \pi)\alpha_{\mathbf{0}}$
  - $u = \hat{s} - s^*$ can be -1 only if $s^* = 1$: correlated with s^*, not classical; A1 derives λ
  - Non-differential error: given s^*, $\hat{s}$ carries no extra information on r

Attenuation in a Simulation



- Left: 400 simulated texts, $\pi = 0.25$, $\alpha_0 = 0.15$, $\alpha_1 = 0.25$, true $\beta = 1$: the slope on the label is about half the slope on the truth
- Right: λ against the error rate; $\lambda = 0$ at 50% errors (a coin toss); a rare class (small π) loses more

Two Classifiers on the Same Texts

- Paired data: for text i, $D_i = 1$ if A is wrong and B right, -1 if A is right and B wrong, 0 otherwise
  - n_{01}, n_{10}: numbers of texts with $D_i = 1$ and $D_i = -1$ (the discordant texts); texts both get right or both get wrong give 0
- **Accuracy gap** and its variance (derived in A2)

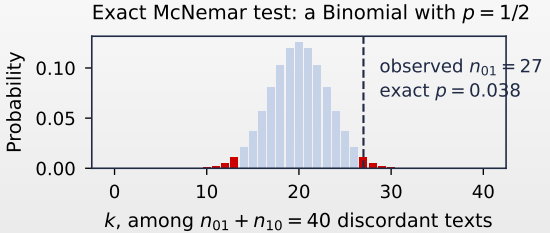
$$\hat{d} = \frac{n_{01} - n_{10}}{n}, \qquad \mathrm{Var}(\hat{d}) = \frac{q - d^2}{n}$$

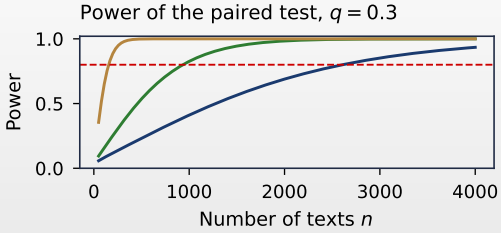
  - $d = \mathrm{acc}_B - \mathrm{acc}_A$; $q = P(D_i \neq 0)$, the share of discordant texts: only they add variance
- **Exact McNemar test** McNemar (1947): under H_0 (equal accuracy) each discordant text favours B with probability $\frac{1}{2}$

$$n_{01} \sim \mathrm{Bin}\left(m, \tfrac{1}{2}\right), \quad m = n_{01} + n_{10}, \qquad p = \min\left\{1,\ 2P\left(\mathrm{Bin}(m, \tfrac{1}{2}) \geq \max(n_{01}, n_{10})\right)\right\}$$

  - Large-sample version: $\chi^2 = (n_{01} - n_{10})^2/m$, compared with 3.84, the 95% quantile of χ^2_1

The McNemar Test and Its Power





$P(n_{01} = k)$ under H_0: Bin(40, 0.5) — as or more extreme than observed

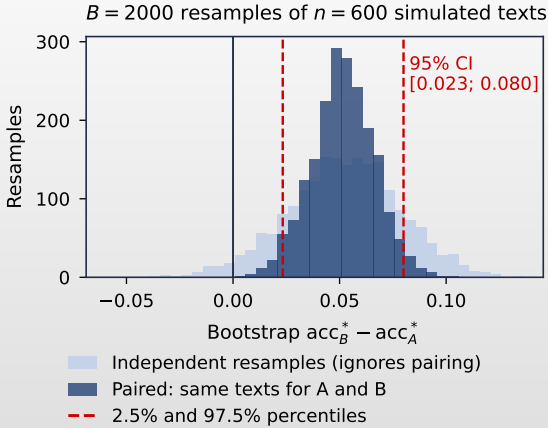
$d = 0.03$, $n_{80} = 2608$ — $d = 0.05$, $n_{80} = 934$ — $d = 0.12$, $n_{80} = 156$ — 80% power

- Power of the two-sided 5% test if the true gap is d, and the n that gives 80% power:

$$\text{power} \approx \Phi\Big(\frac{|d|\sqrt{n}}{\sqrt{q - d^2}} - 1.96\Big), \qquad n_{80} = \frac{(1.96 + 0.84)^2\,(q - d^2)}{d^2}$$

  - Φ: standard Normal distribution function; 0.84: its 80% quantile; left: red bars = the exact p-value for $m = 40$, $n_{01} = 27$
  - Power plans a new study, with the smallest useful effect fixed before the data; computed at the observed effect, it is no evidence after a test

Bootstrap Confidence Intervals



- **Percentile bootstrap** for a statistic $\hat{\theta}$ (an accuracy, a gap)
  - Draw n texts with replacement, B times; recompute $\hat{\theta}_b^*$, $b = 1, \ldots, B$
  - 95% CI $= [q_{2.5\%},\ q_{97.5\%}]$, the percentiles of the $\hat{\theta}_b^*$
- **Paired bootstrap**: the same resampled texts for A and B
  - Keeps the difficulty each text shares between A and B: a narrower interval for the gap (chart)
- **Block bootstrap** for days (C2)
  - Resample blocks of consecutive days, which keeps their serial dependence

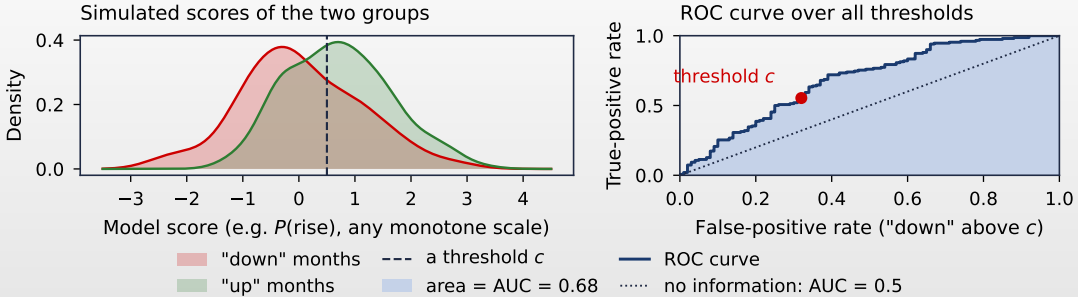
Ranking Quality: AUC

- Task of B3: separate "up" months (n_1) from "down" months (n_0) with a score S, e.g. $P(\text{rise})$
- **AUC (Area Under the Curve)**: probability that a random up month scores above a random down month

$$\text{AUC} = \frac{U}{n_1 n_0}, \qquad U = \sum_{i \in \textbf{up}} \sum_{j \in \textbf{down}} \left(1\{S_i > S_j\} + \tfrac{1}{2}\, 1\{S_i = S_j\} \right)$$

  - U: Mann–Whitney statistic, the number of correctly ordered pairs; $1\{\cdot\}$: 1 if true, else 0
  - AUC $= 0.5$: no information; 1: perfect ordering; only the order of the scores matters
- **Permutation test** of AUC $= 0.5$
  - Shuffle the up/down labels B times; $p = (1 + \#\{\text{AUC}_b^* \geq \text{AUC}\})/(B+1)$, #: count
- **Two independent AUCs** (before and after a date), DeLong standard errors DeLong et al. (1988)
  - $z = (\text{AUC}_1 - \text{AUC}_2)/\sqrt{\text{SE}_1^2 + \text{SE}_2^2} \sim N(0,1)$ under H_0; gap detectable with 80% power at 5%: $2.80\sqrt{\text{SE}_1^2 + \text{SE}_2^2}$

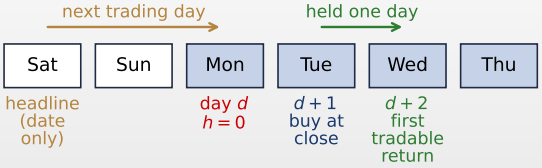
Scores, Thresholds and the ROC Curve



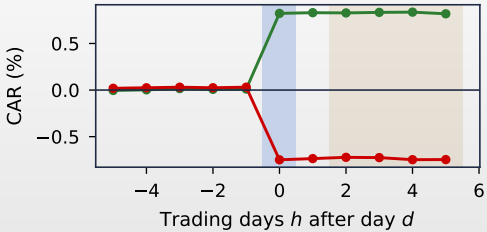
- Left: simulated scores of 150 up and 100 down months; a threshold c calls every month above it "up"
- Right: moving c traces the ROC (Receiver Operating Characteristic) curve, true-positive against false-positive rate; the area under it is the AUC (here 0.68)

From Headlines to Returns: Timing

From a calendar date to the tradable return



Simulated: all news priced on day d



CAR, positive news — CAR, negative news — day d: reaction — $h \geq 2$: prediction

- Day d: the first trading day on or after the headline date; the hour is unknown, so $d + 1$ may still hold the reaction
- A slope at $h = 0$ measures how headlines *describe* the day; only $h \geq 2$ is a *forecast* (B2)
- Event study (B11): the CAR (Cumulative Abnormal Return) $\mathrm{CAR}[h_1, h_2] = \sum_{h=h_1}^{h_2} r^e_{d+h}$ sums the mean excess returns (a sum, not a wealth index); right: news fully priced on day d

Panel Regression and Clustered Standard Errors

- Panel regression on stock-days (stock i, day d, horizon h); the error $\varepsilon_{i,d}$ carries the news day d, not the return day $d+h$

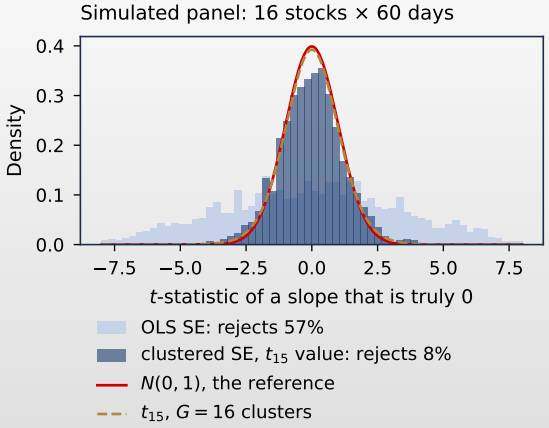
$$r^{e}_{i,d+h} = a + b\, z_{i,d} + \varepsilon_{i,d}$$

  - r^{e}: simple excess return over SPY, from adjusted prices; $z_{i,d}$: daily score standardised over the panel; b: return per one standard deviation (SD, also written s.d.) of the score
- Errors are correlated within a day (common news) and within a stock (over time): OLS standard errors are too small
- **Clustered SE** Petersen (2009): sum the products $X_g^{\top}\hat{u}_g$ within each cluster g (a day or a stock)

$$\widehat{\mathrm{Var}}(\hat{\beta}) = c\,(X^{\top}X)^{-1}\Big(\sum_{g=1}^{G} X_g^{\top}\hat{u}_g\hat{u}_g^{\top}X_g\Big)(X^{\top}X)^{-1}, \qquad c = \frac{G}{G-1}\,\frac{n-1}{n-2}$$

  - $\hat{\beta} = (\hat{a}, \hat{b})$; X: the n rows $(1, z_{i,d})$; X_g, $\hat{u}_g$: rows and OLS residuals of cluster g; c: small-sample correction
  - Few clusters ($G = 16$ stocks): compare $t = \hat{b}/\mathrm{SE}$ with t_{G-1} (2.13), not with 1.96, or use a wild cluster bootstrap Cameron, Gelbach & Miller (2008)

Why Clustering Matters



- Simulation: 16 stocks × 60 days, true slope 0; both the score and the error persist within a stock
- OLS SE treat the 960 rows as independent
  - The t-statistic is far too dispersed: 57% false rejections at 5%
- Clustered SE with the t_{15} critical value
  - About 8% false rejections: close to, but still above, 5%
  - With so few clusters, a wild cluster bootstrap is safer (B9)
  - Question for the class: why compare t with t_{15}, not with $N(0,1)$?
  - Answer: the clustered SE is estimated from only 16 cluster sums, so it has about $G - 1 = 15$ degrees of freedom

A Sentiment Strategy, Its t and Its Costs

- Strategy (A3): at the close of $d+1$, a position of sign $s_{i,d}$ in each stock with news and $s_{i,d} \neq 0$, hedged with SPY, closed a day later; R_t: its return on day t (0 without positions)
- **Newey–West** t Newey & West (1987): the t of a mean when days are serially dependent

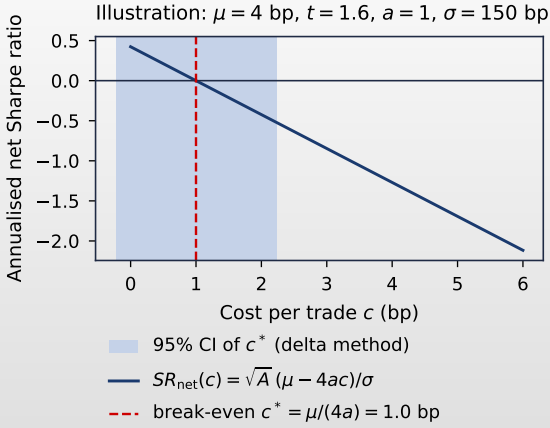
$$t_{NW} = \frac{\hat{\mu}}{\sqrt{\hat{\Omega}/T}}, \qquad \hat{\Omega} = \hat{\gamma}_0 + 2\sum_{l=1}^{L}\left(1 - \frac{l}{L+1}\right)\hat{\gamma}_l$$

  - $\hat{\mu} = \bar{R}$; $\hat{\gamma}_l$: sample autocovariance of R_t at lag l; L: number of lags; T: number of days
- **Costs**: c per trade, four trades on each day with positions, a share a of such days

$$\mu_{\text{net}} = \mu - 4ac, \qquad SR_{\text{net}} \approx \frac{\sqrt{A}\,(\mu - 4ac)}{\sigma}, \qquad c^* = \frac{\mu}{4a}$$

  - $A = 252$ trading days a year; σ: SD of daily gross returns; SR: Sharpe ratio; c^*: break-even cost
- **Delta method**: the SE of a function of an estimate
  - $\mathrm{SE}(g(\hat{\mu})) \approx |g'(\hat{\mu})|\,\mathrm{SE}(\hat{\mu})$; with $\mathrm{SE}(\hat{\mu}) = \hat{\mu}/t_{NW}$ and a fixed: $\mathrm{SE}(c^*) = \mathrm{SE}(\hat{\mu})/(4a)$

Break-Even Cost: an Illustration

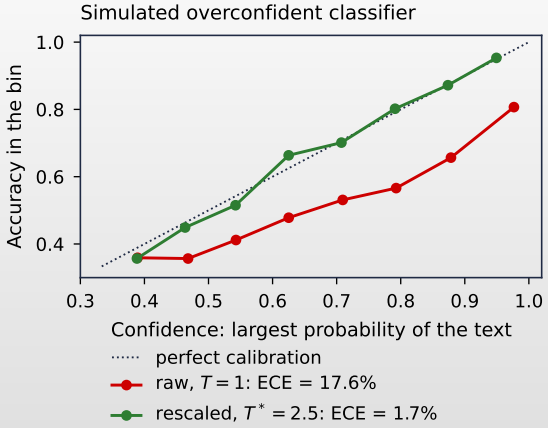


- Illustrative inputs, not those of A3: $\mu = 4$ bp a day, $t_{NW} = 1.6$, $a = 1$, $\sigma = 150$ bp
- The net Sharpe ratio falls linearly in c and crosses 0 at $c^* = 4/4 = 1$ bp
- $\text{SE}(c^*) = (4/1.6)/4 = 0.63$ bp
  - 95% CI of c^*: $1 \pm 1.96 \times 0.63 = [-0.23;\ 2.23]$ bp
  - A realistic cost of 5 bp per trade lies far to the right of the interval

Look-Ahead Bias and Many Tries

- **Look-ahead bias** Glasserman & Lin (2023): an LLM trained on text up to its release may have read what followed a headline
  - A backtest on news before the release mixes forecasting with remembering
  - Memory test (B3): ask the model about past months without any data; an AUC above 0.5 means it remembers
  - Clean test: only texts published after the weights were released (B3, C1)
- **Many tries**: several scores, horizons or prompts, then the best one reported
  - With 20 independent tests at 5%, the chance of at least one false rejection is $1 - 0.95^{\mathbf{20}} = 0.64$
  - Report every try, or correct the p-values (B7, B9, B10)

Probability Calibration



- Calibration (A7) Guo et al. (2017): among texts given probability 0.8, about 80% should be right
- Negative Log-Likelihood (NLL), the loss used to choose T
  - $\text{NLL}(T) = -\sum_i \ln p_{iy_i}(T)$, y_i: true label of text i
- Brier score: mean squared error of the probabilities
  - $\frac{1}{n}\sum_i \sum_k (p_{ik} - 1\{y_i = k\})^2$
- Expected Calibration Error (ECE): $\sum_b \frac{n_b}{n} |\text{acc}_b - \text{conf}_b|$
  - Bins b of the largest probability; conf_b: its mean, acc_b: accuracy, n_b: texts in the bin
- Chart: an overconfident model lies below the diagonal; $T^* > 1$ moves it onto it

Holm, Wild Cluster Bootstrap, PPI and Fleiss' κ

- **Holm correction** for M tests: sort $p_{(\mathbf{1})} \leq \cdots \leq p_{(M)}$
  - $\tilde{p}_{(j)} = \max_{i \leq j} \min\{1, (M - i + 1)\, p_{(i)}\}$; reject where $\tilde{p} < 0.05$: at most 5% chance of any false rejection
- **Wild cluster bootstrap** Cameron, Gelbach & Miller (2008) (B9)
  - $y_i^* = \bar{y} + w_{g(i)}(y_i - \bar{y})$, $w_g = \pm 1$ with probability $\frac{\mathbf{1}}{\mathbf{2}}$ per cluster; $p = P^*(|t^*| \geq |t|)$
  - Two-way clustering: $\hat{V} = \hat{V}_{\mathbf{day}} + \hat{V}_{\mathbf{stock}} - \hat{V}_{\mathbf{day} \cap \mathbf{stock}}$
- **Prediction-powered inference** (PPI) Angelopoulos et al. (2023) (B8): a model mean corrected by its bias on labelled texts
  - $\hat{\theta} = \frac{\mathbf{1}}{N} \sum_{\mathbf{unlab.}} f - \frac{\mathbf{1}}{n} \sum_{\mathbf{lab.}} (f - Y)$; f: model score, Y: human label
  - N: the unlabelled texts, n: the labelled ones, two separate samples
  - More predictions do not repair an unrepresentative labelled sample
- **Fleiss' κ** Fleiss (1971) (A5): agreement among raters beyond chance
  - $\kappa = (\bar{P} - \bar{P}_e)/(1 - \bar{P}_e)$: $\bar{P}$ observed, $\bar{P}_e$ chance agreement; 0 = chance, 1 = perfect

A1 [Solved]: Attenuation from Misclassification — Model and Data

- **Question**: a regression uses the FinBERT label "negative" instead of the truth; by how much is the slope shrunk?
- **Data**
  - Twitter Financial News, validation set: $N = 2\,388$ headlines from 2022, each labelled negative, neutral or positive; the human label is treated as the truth
  - Model $r = a + \beta s^* + \varepsilon$, $s^* = 1$ if the headline is truly negative; we observe only $\hat{s} = 1$ when FinBERT predicts negative
  - Non-differential error: given s^*, $\hat{s}$ says nothing about r, so $\mathrm{Cov}(\varepsilon, \hat{s}) = 0$
  - π, p, false alarms α_0 and misses α_1: the rates defined on the Measurement Error slide

FinBERT confusion matrix, Twitter

True \ predicted	neg	neu	pos	Total
neg	264	71	12	347
neu	213	1 184	169	1 566
pos	58	133	284	475
Total	535	1 388	465	2 388

Back Exercise list

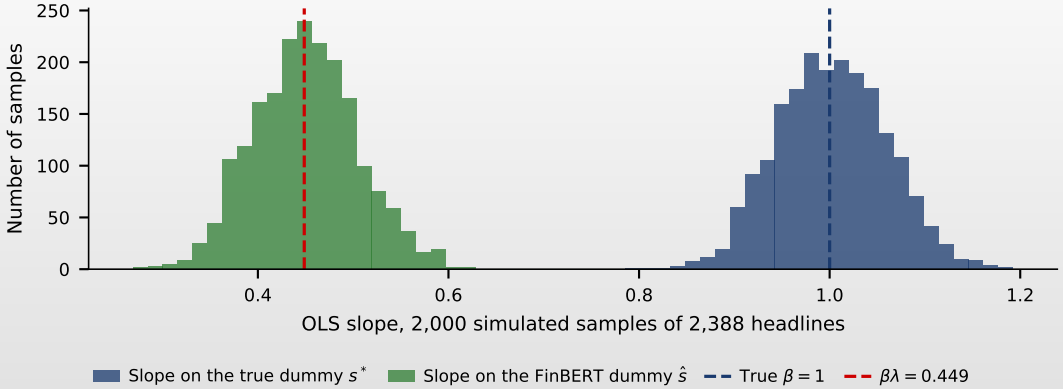
A1 [Solved]: Tasks and Interpretation

- **Tasks**
  (a) Show that plim $\hat{b} = \beta\,\mathrm{Cov}(s^*, \hat{s})/\mathrm{Var}(\hat{s})$.
  (b) Derive $\lambda = \pi(1-\pi)(1-\alpha_0-\alpha_1)/[p(1-p)]$.
  (c) Compute π, p, α_0, α_1 from the matrix, then λ to three decimals.
  (d) Compute $\mathrm{Cov}(u, s^*)$ for $u = \hat{s} - s^*$: is the error classical?
- **Report**: λ, $1/\lambda$ and the sign of $\mathrm{Cov}(u, s^*)$
- **Interpretation**: what share of a true effect survives in a slope on the FinBERT dummy?

A1 [Solved]: Solution (1/3)

- (a) For one regressor plim $\hat{b} = \text{Cov}(r, \hat{s})/\text{Var}(\hat{s})$ and $\text{Cov}(r, \hat{s}) = \beta \,\text{Cov}(s^*, \hat{s}) + \text{Cov}(\varepsilon, \hat{s})$; the last term is 0 by the non-differential assumption
- (b) Both variables are binary: $\text{Cov}(s^*, \hat{s}) = P(s^* = 1, \hat{s} = 1) - \pi p = \pi(1 - \alpha_1) - \pi p$
- Total probability: $p = \pi(1 - \alpha_1) + (1 - \pi)\alpha_0$; substituting, $\text{Cov}(s^*, \hat{s}) = \pi(1 - \pi)(1 - \alpha_0 - \alpha_1)$
- $\hat{s}$ is Bernoulli, so $\text{Var}(\hat{s}) = p(1 - p)$; the ratio of the two is λ

A1 [Solved]: Solution (2/3)



MFM_ch15_sem_partA

A1 [Solved]: Solution (3/3)

- (c) $\pi = 347/2\,388 = 0.1453$, $p = 535/2\,388 = 0.2240$, $\alpha_{\mathbf{1}} = 1 - 264/347 = 0.2392$, $\alpha_{\mathbf{0}} = 271/2\,041 = 0.1328$
- $\lambda = \dfrac{0.1242 \times 0.6280}{0.2240 \times (1 - 0.2240)} = 0.449$, so $\beta = 2.23\,\hat{b}$
- (d) $\mathrm{Cov}(u, s^*) = 0.0780 - 0.1242 = -0.0462 < 0$: the error is negatively correlated with the truth, not classical Aigner (1973)
- Chart: 2 000 simulated samples with these rates and $\beta = 1$; the slope on $\hat{s}$ averages 0.450 (SD 0.054), as λ predicts
- Interpretation: a slope on the FinBERT dummy keeps only about 45% of the true effect

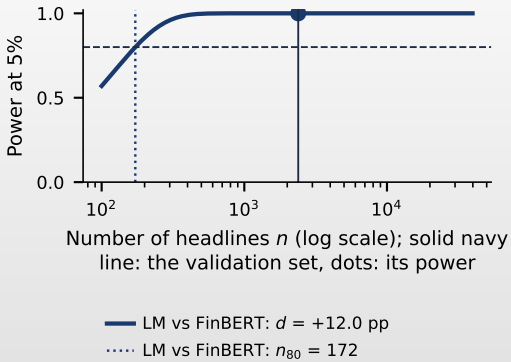
MFM_ch15_sem_partA

A2 [Solved]: Paired Accuracy Difference: Variance, Exact Test, Power — Task

- **Questions**
  - how precise is the accuracy gap between two classifiers scored on the same headlines?
  - how many headlines does it need?
- **Inputs**
  - Twitter validation set, $n = 2\,388$ independent headlines; A = LM dictionary, B = FinBERT
  - Discordant counts: $n_{01} = 539$ (A wrong, B right), $n_{10} = 252$ (A right, B wrong)
  - $D_i = 1$ if A wrong and B right, -1 if A right and B wrong, 0 otherwise; $d = ED_i = \mathrm{acc}_B - \mathrm{acc}_A$, $q = P(D_i \neq 0)$
- **Tasks**
  (a) Show $\hat{d} = (n_{01} - n_{10})/n$ and derive $\mathrm{Var}(\hat{d}) = (q - d^2)/n$.
  (b) Compute $\hat{d}$, SE, the 95% CI and $z = \hat{d}/\mathrm{SE}$.
  (c) Run the exact McNemar test at 5% and state the decision.
  (d) If the true d is $\hat{d}$, compute the power and n_{80}, the n for 80% power, with the formulas on the McNemar Test slide.
- **Report**: $\hat{d}$, SE, CI, z, exact p, the decision, the power and n_{80}
- **Interpretation**: which headlines determine the variance of $\hat{d}$?

Back Exercise list

A2 [Solved]: Solution



- (a) $\hat{d} = \frac{1}{n}\sum_i D_i = (n_{01} - n_{10})/n$; $D_i^2 = 1$ exactly when one is right and the other wrong, so $ED_i^2 = q$, $\mathrm{Var}(D_i) = q - d^2$ and $\mathrm{Var}(\hat{d}) = (q - d^2)/n$
- (b) $\hat{d} = (539 - 252)/2\,388 = 0.1202$, $\hat{q} = 791/2\,388 = 0.3312$, SE = 1.15 pp, CI [9.76; 14.28] pp, $z = 10.43$
- (c) Under H_0 we expect 395.5 on each side; exact $p < 0.001$: reject equal accuracy
- (d) Power $\Phi(+8.47) \approx 100\%$; $n_{80} = 7.84 \times (0.3312 - 0.01444)/0.01444 \approx 172$ headlines
- Interpretation: only the discordant headlines (q) enter the variance; a 12-point gap is detectable with about 170 headlines

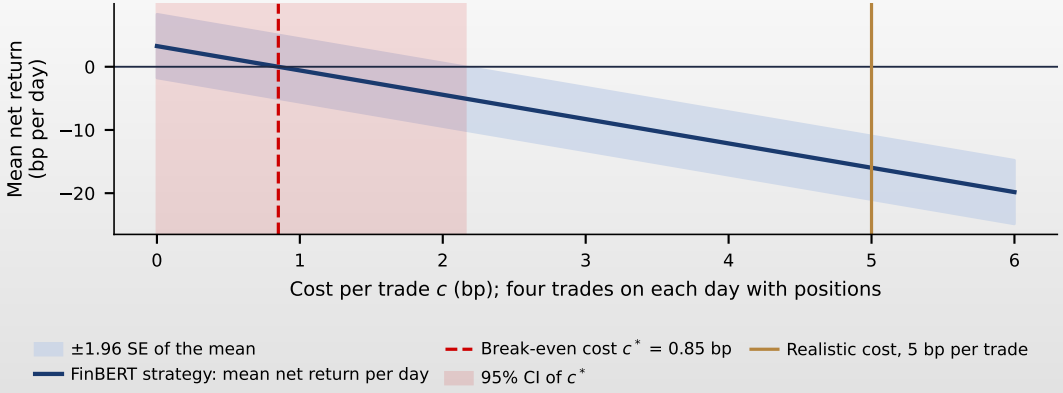
MFM_ch15_sem_partA

A3 [Solved]: Break-Even Cost with a Standard Error — Task

- **Questions**
  - at what cost per trade does the FinBERT strategy stop earning?
  - how precisely do we know that cost?
- **Inputs**
  - Sentiment strategy (the Sentiment Strategy slide): at the close of $d+1$ take $\text{sign}(s_{i,d})$ in every stock with news and a non-zero score on day d, hedged with SPY, closed one day later; four trades on each day with positions (entry and exit, in the stocks and in SPY); 3 568 trading days
  - FinBERT signal: mean gross return over all days $\hat{\mu} = 3.27$ bp, $t_{NW} = 1.27$ ($L = 8$ lags), share of days with positions $a = 0.962$, gross Sharpe ratio $\sqrt{A}\,\hat{\mu}/\sigma = 0.36$, σ the standard deviation of daily returns, $A = 252$ trading days a year
- **Tasks**
  - (a) Holding the gross-return standard deviation and the activity share fixed, write the mean net daily return and the approximate net Sharpe ratio as functions of the cost c per trade.
  - (b) Derive the break-even cost c^* at which the mean net return is zero.
  - (c) Compute $\text{SE}(\hat{\mu}) = \hat{\mu}/t$, then $\text{SE}(c^*)$ by the delta method and the 95% CI of c^*.
  - (d) Recover σ from the gross Sharpe ratio and compute the net Sharpe ratio at $c = 5$ bp.
  - (e) New case: $\hat{\mu} = 2$ bp, $a = 0.9$, $\sigma = 120$ bp, $c = 0.4$ bp; compute c^* and the annualised net Sharpe ratio.
- **Report**: c^*, its SE and CI, the net Sharpe ratio; for (e), c^* and the net Sharpe ratio
- **Interpretation**: is a realistic cost of 5 bp per trade inside the CI of c^*?

Back Exercise list

A3 [Solved]: Solution (1/2)



MFM_ch15_sem_partA

A3 [Solved]: Solution (2/2)

- (a) On a share a of days we pay $4c$: $\mu_{\textbf{net}} = \mu - 4ac$, $SR_{\textbf{net}} = \sqrt{A}\,(\mu - 4ac)/\sigma$, using the gross σ as an approximation
- (b) $c^* = \mu/(4a) = 3.27/3.85 = 0.85$ bp per trade
- (c) $\mathrm{SE}(\hat{\mu}) = 3.27/1.27 = 2.57$ bp; with a fixed by the news calendar, $\mathrm{SE}(c^*) = 2.57/3.85 = 0.67$ bp, CI $[-0.46;\ 2.16]$ bp
- (d) $\sigma = \sqrt{252} \times 3.27/0.362 = 143$ bp, with the unrounded Sharpe ratio 0.362
  - $SR_{\textbf{net}} = \sqrt{252}\,(3.27 - 3.85 \times 5)/143 = -1.77$
- (e) $c^* = 2/(4 \times 0.9) = 0.56$ bp; $SR_{\textbf{net}} = \sqrt{252}\,(2 - 3.6 \times 0.4)/120 = +0.07$: barely profitable
- Interpretation: 5 bp lies far above the CI of c^*; with daily turnover the strategy cannot survive realistic costs

MFM_ch15_sem_partA

B1 [Solved]: Is FinBERT Better than LM? — Task

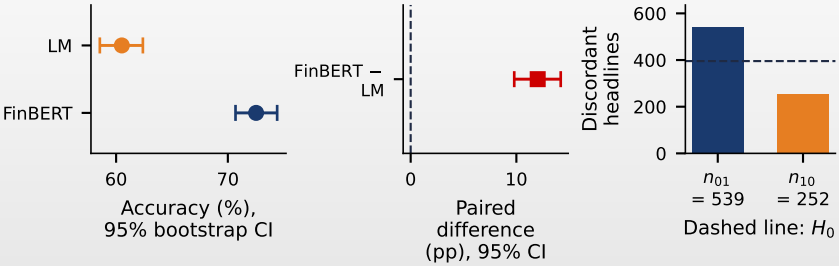
- **Question**: is the fine-tuned FinBERT more accurate than the Loughran–McDonald dictionary on headlines it was not trained on?
- **Data**
  - Twitter validation set of A1: $N = 2\,388$ headlines, human label = truth
  - A = LM dictionary (label = sign of the tone), B = FinBERT (label = largest probability); notebook function `b_pairs`
- **Tasks**
  (a) Compute each accuracy with a percentile bootstrap 95% CI: resample the headlines 2 000 times (seed 0).
  (b) Compute the difference B − A with a paired bootstrap CI: the same resampled headlines for both classifiers.
  (c) Count n_{01}, n_{10}; run the exact McNemar test and $(n_{01} - n_{10})^2/(n_{01} + n_{10})$ against χ^2_1 (3.84).
- **Report**: a table of the two accuracies, their CIs and the difference; n_{01}, n_{10}, both p-values and the decision
- **Interpretation**
  - why is the paired test more powerful than comparing the two intervals?
  - does the gap matter for trading?

B1 [Solved]: Solution (1/2)

	Accuracy (%)	CI 95%
LM	60.5	[58.5; 62.4]
FinBERT	72.5	[70.7; 74.4]
Difference (pp)	+12.0	[+9.8; +14.2]

- Step 1: labels of both classifiers on the 2 388 headlines; step 2: 2 000 bootstrap resamples of the headline indices; step 3: both accuracies and their difference in each resample
- (a)–(b) The two intervals do not overlap and the paired CI of the difference excludes 0; it agrees with the analytic CI of A2

B1 [Solved]: Solution (2/2)



- (c) $n_{01} = 539$, $n_{10} = 252$, 395.5 expected under H_0; exact $p < 0.001$; $\chi^2 = 104.1 > 3.84$: reject equal accuracy
- Pairing removes the variation in headline difficulty: headlines both get right, or both get wrong, carry no information
- Interpretation: FinBERT is clearly better than the dictionary outside its training data; a 12-point gain in labelling says nothing yet about returns (B2)

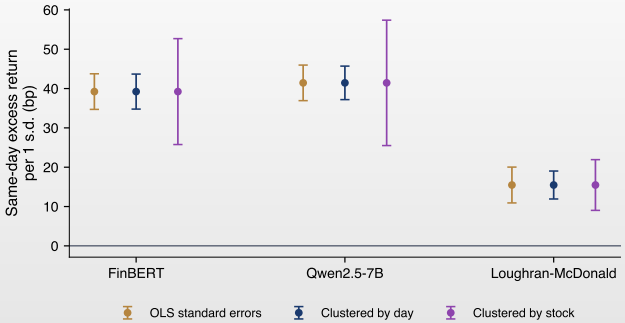
MFM_ch15_sem_partB

B2 [Solved]: Reaction or Prediction? — Task

- **Questions**
  - does the daily headline score explain the return of the news day?
  - does it forecast the first return one can still trade?
  - how wide are the honest intervals?
- **Data**
  - FNSPID panel: 18 282 stock-days with news, 16 US (United States) stocks, 27.10.2009–18.12.2023; $z_{i,d}$: daily net score, standardised over the panel; r^e: excess return over SPY (%), from the adjusted daily closes of the course data
  - Scores: FinBERT, Qwen2.5-7B, LM
- **Tasks**
  (a) Estimate $r^e_{i,d+h} = a + b\, z_{i,d} + \varepsilon$ by OLS for the three scores and $h = 0, 1, 2$; report b in bp per one SD of the score.
  (b) Compute three SEs: OLS, clustered by day ($G = 3\,434$), clustered by stock ($G = 16$), with the small-sample factor c (the Panel Regression slide).
  (c) Compare each t with 1.96, and the stock-clustered t with the $t_{\mathbf{15}}$ critical value 2.13.
- **Report**: a chart of the $h = 0$ slopes with the three intervals, and a table of the regressions
- **Interpretation**
  - why do $h = 0$ and $h = 2$ differ?
  - can a clustered SE with 16 clusters be trusted?

Back Exercise list

B2 [Solved]: Solution (1/2)



- Same-day slopes with the 95% intervals of the three standard errors: the stock-clustered interval is the widest
- Clustering by day is close to OLS: with only 5.3 stocks per day there is little common noise
- Clustering by stock multiplies the FinBERT SE at $h = 0$ by 3.0: the slope differs across stocks and each stock's errors are correlated over time

MFM_ch15_sem_partB

B2 [Solved]: Solution (2/2)

Score	h	b (bp)	SE OLS	SE day	SE stock	t day	t stock
FinBERT	0	+39.2	2.3	2.3	6.9	+17.27	+5.71
FinBERT	1	+3.2	2.2	2.0	2.5	+1.60	+1.25
FinBERT	2	−0.5	2.2	2.2	1.5	−0.25	−0.35
Qwen2.5-7B	0	+41.4	2.3	2.2	8.1	+19.07	+5.10
Qwen2.5-7B	1	+3.7	2.2	2.0	3.0	+1.86	+1.23
Qwen2.5-7B	2	+2.3	2.2	1.9	1.0	+1.19	+2.41
LM	0	+15.5	2.3	1.8	3.3	+8.53	+4.70
LM	1	−0.4	2.2	1.7	1.4	−0.21	−0.26
LM	2	−0.0	2.2	1.7	0.9	−0.01	−0.03

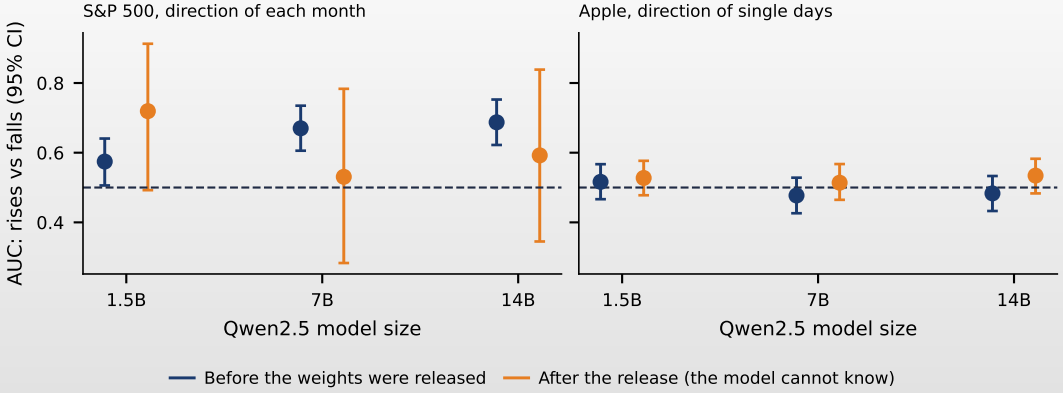
- $h = 0$: large under every SE (t by stock at least 4.7): headlines describe the day's move, often by reporting it ("shares jump")
- $h = 2$: FinBERT -0.5 bp ($t = -0.25$ by day); Qwen $t = +2.41$ by stock, but 16 clusters are few: check with a wild cluster bootstrap Cameron, Gelbach & Miller (2008) and Holm (B9)
- Interpretation: sentiment explains the news day, not the tradable days; with 16 stocks, inference needs few-cluster methods

B3 [Solved]: Does the Model Remember the Market? — Task

- **Question**: asked without any data, can Qwen2.5 tell the months in which the S&P 500 rose from those in which it fell?
- **Data**
  - S&P 500 monthly closes (`GSPC.INDX`), January 2000 to August 2026; a month is "up" if its return is positive
  - Qwen2.5 1.5B, 7B, 14B (weights published 19 September 2024) asked "Did the S&P 500 index rise or fall during ⟨month year⟩?"; score $P(\text{rise})$
  - Before: 296 months to August 2024 (185 up); after: 23 months from October 2024 (13 up); September 2024, the release month, is left out
- **Tasks**
  (a) Compute the AUC $= U/(n_1 n_0)$ (the AUC slide) per model and period, with a 2 000-draw bootstrap CI over months.
  (b) Permutation test for 14B before the release: shuffle the up/down labels 2 000 times.
  (c) For 14B, test $\text{AUC}_{\text{before}} = \text{AUC}_{\text{after}}$ with DeLong variances DeLong et al. (1988) and find the smallest drop detectable with 80% power, with $\text{SE}_0 = \sqrt{(n_1 + n_0 + 1)/(12\, n_1 n_0)}$, the standard error of the post-release AUC under H_0: AUC $= 0.5$ (n_1, n_0: up and down months after the release).
- **Report**: the AUC table with CIs, the permutation p, the DeLong z and p, the detectable drop
- **Interpretation**: what does an AUC above 0.5 mean when the model sees no data?

Back Exercise list

B3 [Solved]: Solution (1/3)



MFM_ch15_lookahead

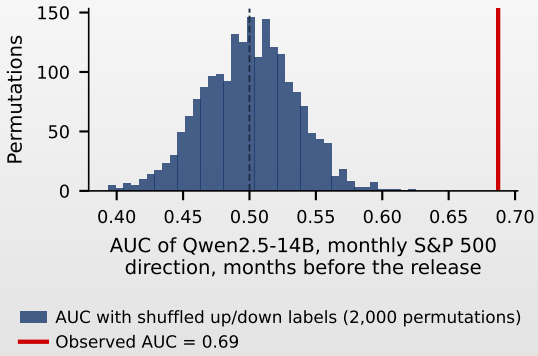
B3 [Solved]: Solution (2/3)

Qwen2.5	AUC before	CI 95%	AUC after	CI 95%
1.5B	0.57	[0.51; 0.64]	0.72	[0.49; 0.91]
7B	0.67	[0.61; 0.73]	0.53	[0.28; 0.78]
14B	0.69	[0.62; 0.75]	0.59	[0.35; 0.84]

- (a) 7B and 14B rank up months above down months far above chance before the release; 1.5B barely; after the release all intervals contain 0.5
- Right panel of the chart: the same question for the direction of 500 Apple days before and 493 after the release
- For days, the AUC lies between 0.48 and 0.53 and every interval contains 0.5: we detect no discrimination above chance in this experiment
- The discrimination of months is compatible with memorisation, but on its own does not separate it from calendar or regime regularities

MFM_ch15_lookahead

B3 [Solved]: Solution (3/3)



- (b) Permutation p for 14B before the release <0.001: no shuffled AUC reaches the observed one (chart)
- (c) DeLong, 14B: SE before 0.032, after 0.126; $z = 0.73$, $p = 0.47$: do not reject equality
  - $\mathrm{SE}_{\mathbf{0}} = \sqrt{24/(12 \times 13 \times 10)} = 0.124$ ($n_{\mathbf{1}} = 13$ up months, $n_{\mathbf{0}} = 10$ down)
  - Detectable drop $2.80\sqrt{\mathrm{SE}^2_{\mathbf{before}} + \mathrm{SE}^2_{\mathbf{0}}} = 0.36$
  - Detecting the fall from 0.69 to 0.5 with 80% power would need about 101 post-release months; we have 23
- Interpretation: without data, the discrimination can only come from what the model learned in training (memory or regularities)
  - A backtest on pre-release news mixes forecasting with remembering Glasserman & Lin (2023)

MFM_ch15_sem_partB

C1: Reproduce the Analysis from Saved Scores

- Use `ch15_twitter_scores.csv` for the labelled comparison and `ch15_news_daily.csv` for daily stock scores and returns, from the chapter Quantlet folder. No new model inference is needed for this exercise.
- In the assistant's code (next slide), `qwen_label` and `qwen_score` stand for those saved scores. Separate three principal errors in the claim from additional inference limitations; count the same underlying error once when it propagates to several lines.
- Day-only clustering is an inference choice to assess using B2, not a guarantee of reliable uncertainty with persistent stock effects. Report that limitation separately from mistakes in labels, timing and claims about information availability.

C1 [Proposed]: Critique an AI Answer — the Answer

- **Student prompt**: "I have FNSPID headlines for 16 US stocks, 2009–2023, and daily excess returns vs SPY. Check that Qwen2.5-7B reads sentiment, then test whether its headline sentiment forecasts returns out of sample."
- **Code of the AI assistant** (`qwen_label`, `qwen_score`: the zero-shot scorer on the Three Ways to Read Sentiment slide)

```
import pandas as pd, statsmodels.formula.api as smf
from sklearn.metrics import accuracy_score
tw['pred'] = [qwen_label(t) for t in tw['text']]   # Twitter validation set
acc = accuracy_score(tw['label'], tw['pred'])
news['day'] = first_trading_day_on_or_after(news['date'])
news['s'] = [qwen_score(t) for t in news['title']]  # P(pos) - P(neg)
d = news.groupby(['ticker', 'day'])['s'].mean().reset_index()
d['z'] = (d['s'] - d['s'].mean()) / d['s'].std()
d = d.merge(rets, on=['ticker', 'day'])     # rets: ticker, day, xret (%)
fit = smf.ols('xret ~ z', data=d).fit(cov_type='cluster',
              cov_kwds={'groups': pd.factorize(d['day'])[0]})
```

Back Exercise list

C1 [Proposed]: Critique an AI Answer — Task

- **Model**: B1, B2 and B3 [Solved]
- **Conclusion of the AI assistant**
  - "Accuracy 74.2%, more than twice the 33% of a random guess among three classes: Qwen2.5-7B reads financial sentiment very well."
  - "Slope +41.4 bp per standard deviation of the score, $t = +19.07$ (clustered by day): the day's headlines forecast the stock's excess return."
  - "Qwen2.5 was never trained on these labels or return series, so the 2009–2023 backtest is fully out of sample."
- **Question**: the code runs and the numbers are right, yet the answer contains **three** errors typical of sentiment backtests; which?
- **Tasks**
  - (a) Find each error: name the line of code or the sentence and explain why it is wrong.
  - (b) Say how you would detect it without knowing the answer: a class count, a timestamp check, a release date, a memory test.
  - (c) Write the corrected code or statement, with the numbers it produces.
  - (d) Log each error in `AI_ERRORS.md` with the fields: what the AI answered, what was wrong, how you found out, the correct version.
- **Report**: the three errors, each with its detection, its correction and its numbers, and the `AI_ERRORS.md` entries

B4 [Proposed]: Is a 7-Billion-Parameter LLM Better than FinBERT? — Task

- **Model**: B1 [Solved]
- **Question**: does a general open LLM, used zero-shot, beat the fine-tuned FinBERT on headlines neither was trained on?
- **Data**
  - Twitter validation set of A1: $N = 2\,388$ headlines, human label = truth
  - A = FinBERT, B = Qwen2.5-7B zero-shot with prompt P1; neither used any Twitter label
- **Tasks**
  (a) Compute each accuracy with a percentile bootstrap 95% CI: resample the headlines 2 000 times (seed 0).
  (b) Compute the difference B − A with a paired bootstrap CI: the same resampled headlines for both classifiers.
  (c) Count n_{01}, n_{10}; run the exact McNemar test and $(n_{01} - n_{10})^2/(n_{01} + n_{10})$ against χ^2_1 (3.84).
- **Report**: the same table, tests and chart as in B1
- **Interpretation**
  - do the two intervals overlap?
  - does that settle the question?
  - what would a much larger, slower model still bring?

Back Exercise list

Team Project Step

- Choose a text source not covered in class (announcements of the BVB (Bursa de Valori București, the Bucharest Stock Exchange), Romanian financial news, central-bank statements, crypto posts)
- Label at least 200 texts with two annotators; report Cohen's kappa (observed agreement corrected for chance agreement). Reserve at least 30% of the texts as a test sample before scoring any model and evaluate only on that held-out part
- Deliver one page: source, label counts, agreement, held-out accuracy and its confidence interval.
- Log AI use in `AI_USE.md` and caught errors in `AI_ERRORS.md`; extension packs are optional; C2 and C3 are project starts
- Assessment criteria
  - documented labels: two annotators and Cohen's kappa
  - accuracy measured only on the held-out sample, with a confidence interval
  - trading results reported after costs
  - one economic interpretation per result

Extension Packs: What Each One Needs

Pack	Exercises	Prerequisites
1. Measurement error	A4, A5	A1
2. Comparing classifiers	A6, B5 [Solved], B6, B7	A2, B1
3. LLM probabilities	A7 [Solved]	the From Logits to Probabilities slide, the Calibration slide
4. Valid inference	B8 [Solved], B9, B10	B2, B3
5. Signals and costs	A8, B11, C2	A3, B2
6. Replication lab	C3	C2, Lopez-Lira & Tang (2026)

- Each exercise has its own notebook section; do them in order within a pack

A4 [Proposed]: Attenuation for the LM Dictionary — Task

- **Model**: A1 [Solved]
- **Question**: how much does the error in the LM dictionary's binary "negative" label shrink the slope? The continuous-score slopes in B2 (15.5 bp for LM, 39.2 bp for FinBERT) answer a different question.
- **Data**
  - Twitter validation set of A1: $N = 2\,388$ headlines, human label = truth
  - Same model and assumption as A1; now $\hat{s} = 1$ when the LM tone of the headline is negative
- **Tasks**
  (a) Compute π, p, α_0 and α_1 from the LM matrix.
  (b) Compute λ and $1/\lambda$ with the formula of A1.
  (c) Compute $\mathrm{Cov}(u, s^*)$: is the LM error classical?
  (d) Compare λ with that of FinBERT (A1) and compute the ratio of the two.
- **Report**: the four rates, λ, $1/\lambda$, the ratio
- **Interpretation**: which classifier attenuates the slope more?

LM confusion matrix, Twitter

True \ predicted	neg	neu	pos	Total
neg	172	166	9	347
neu	247	1 192	127	1 566
pos	45	349	81	475
Total	464	1 707	217	2 388

A5 [Proposed]: Fleiss' κ and One Dawid–Skene Step — Task

- **Model**: the κ formula on the Holm, PPI and κ slide
- **Questions**
  - how much do raters agree beyond chance?
  - what is the most likely true label of a disputed sentence?
- **Inputs**
  - Illustrative votes (made up for the exercise): 6 sentences, $m = 5$ raters each
  - $P_i = \sum_j n_{ij}(n_{ij} - 1)/[m(m-1)]$, $\bar{P} = \frac{1}{N}\sum_i P_i$, p_j = share of votes for j, $\bar{P}_e = \sum_j p_j^2$; κ Fleiss (1971)
- **Tasks**
  - (a) Compute every P_i and $\bar{P}$.
  - (b) Compute p_j, $\bar{P}_e$ and κ.
  - (c) E-step of Dawid & Skene (1979): each rater gives the true class with probability 0.8 and each other class with 0.1, uniform prior; compute $P(\text{true} = j \mid \text{votes}) \propto \prod_l \pi_{jl}^{n_{il}}$ for sentences 4 and 5.
  - (d) Compare the posteriors with the vote shares and describe the M-step.
- **Report**: κ and the two posteriors
- **Interpretation**: how would you treat the disputed sentences in a test set?

Sentence	neg	neu	pos
1	5	0	0
2	0	5	0
3	0	1	4
4	1	3	1
5	2	3	0
6	0	2	3

Back Exercise list

A6 [Proposed]: Is "No Difference" a Power Problem? — Task

- **Model**: A2 [Solved]
- **Question**: FinBERT and Qwen2.5-7B look tied: is that evidence of equal accuracy, or of a weak test?
- **Inputs**
  - Twitter validation set, $n = 2\,388$; A = FinBERT, B = Qwen2.5-7B (prompt P1); $n_{01} = 395$, $n_{10} = 355$
- **Tasks**
  (a) Compute $\hat{d}$, $\hat{q}$, SE, the 95% CI and z with the formulas of A2.
  (b) Run the exact McNemar test at 5%.
  (c) Compute the power at $n = 2\,388$ if the true d is $\hat{d}$, and n_{80}.
  (d) Use the CI to state which differences the data rule out.
- **Report**: the numbers of (a)–(d)
- **Interpretation**: may you conclude that the two models are equally accurate?

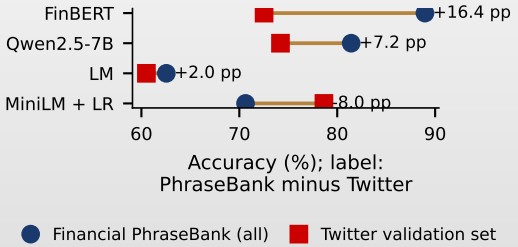
B5 [Solved]: Training-Data Leakage — Task

- **Question**: how much of FinBERT's accuracy on Financial PhraseBank comes from having been trained on it?
- **Data**
  - Financial PhraseBank Malo et al. (2014): 4 846 sentences (2 264 with full agreement); FinBERT was fine-tuned on a split of them (3 101 of 4 845 in Araci (2019); the release used here has 4 846)
  - Twitter validation set, 2 388 headlines unseen by FinBERT; classifiers: FinBERT, Qwen2.5-7B, LM, MiniLM embeddings Wang et al. (2020) + logistic regression, trained on the Twitter training set
- **Tasks**
  (a) Compute accuracy (2 000-draw bootstrap CI) and macro-F1 on PhraseBank with full agreement, all of PhraseBank and Twitter.
  (b) Compute the drop from all of PhraseBank to Twitter for each classifier.
  (c) Compute the difference in differences FinBERT minus Qwen2.5-7B with a bootstrap CI (each data set resampled independently).
- **Report**: a table of accuracies and drops, the difference in differences with its CI
- **Interpretation**: how do you separate "harder data" from "memory of the training set"?

Back Exercise list

B5 [Solved]: Solution

Accuracy (%)	PhraseBank, 100% agr.	PhraseBank, all	Twitter	Drop (pp)	Macro-F1, Twitter
FinBERT	97.2	89.0	72.5	16.4	66.8
Qwen2.5-7B	93.6	81.4	74.2	7.2	72.6
LM	65.0	62.5	60.5	2.0	46.2
MiniLM + log. reg.	81.4	70.6	78.6	−8.0	69.9



- (c) $16.4 - 7.2 = 9.2$ pp from unrounded accuracies, bootstrap CI [6.7; 11.9]: excludes 0
- The result: a descriptive difference between the accuracy drops
  - Qwen2.5-7B was not fine-tuned on PhraseBank, but the texts may be in its pretraining data
  - Attributing the 9.2 points to overlap needs an uncontaminated comparator and no other differential change (distribution, difficulty, class mix)
- Interpretation: MiniLM moves the other way because it is trained on Twitter: every model is best at home

MFM_ch15_sem_partB

B6 [Proposed]: How Many Labels Do You Need? — Task

- **Model**: B5 [Solved]
- **Question**: how many labelled headlines does a simple supervised model need to beat FinBERT?
- **Data**
  - Twitter Financial News: training set of 9 543 headlines, validation set of 2 388; benchmark FinBERT, 72.5%, no Twitter label used
  - Logistic regression ($C = 4$) on 384-dimensional MiniLM embeddings; logistic regression ($C = 8$) on TF-IDF (Term Frequency–Inverse Document Frequency): words and word pairs weighted by frequency and rarity
  - Implementation: one `default_rng(15)`, simple subsets without stratification, in increasing n; MiniLM `all-MiniLM-L6-v2` kept fixed; `TfidfVectorizer(ngram_range=(1,2), sublinear_tf=True)`; `max_iter=3000`
- **Tasks**
  (a) For $n \in \{100, 200, 400, \ldots, 6\,400, 9\,543\}$ draw 10 random training subsets (the full set once), fit both models and average the validation accuracy.
  (b) Find the smallest n at which each model beats FinBERT on average.
  (c) Report the SD of the embedding accuracy over the 10 repetitions at $n = 100$.
- **Report**: the learning curve and its table
- **Interpretation**: why do embeddings start higher but grow more slowly?

Back Exercise list

B7 [Proposed]: Prompt Sensitivity — Task

- **Model**: B1 [Solved]
- **Question**: does the answer of the same LLM change when the same question is worded differently?
- **Data**
  - Twitter validation set of A1: $N = 2\,388$ headlines, human label = truth
  - Qwen2.5-7B and 14B with three prompts: P1 "classify the sentiment ... for investors", P2 "good, bad or neutral for the stock price", P3 "what is the sentiment of this text"
- **Tasks**
  (a) Compute accuracy with a 2 000-draw bootstrap CI and the share of "neutral" answers for each prompt and model.
  (b) Run the exact McNemar test of P1 against P3 for each model.
  (c) Compute the share of headlines with the same label under all three prompts.
- **Report**: a table and a chart of the three prompts
- **Interpretation**: if you picked the best prompt on this set, what would you be doing?

Back Exercise list

A7 [Solved]: From Label Probabilities to a Score, and the Optimal Temperature — Task

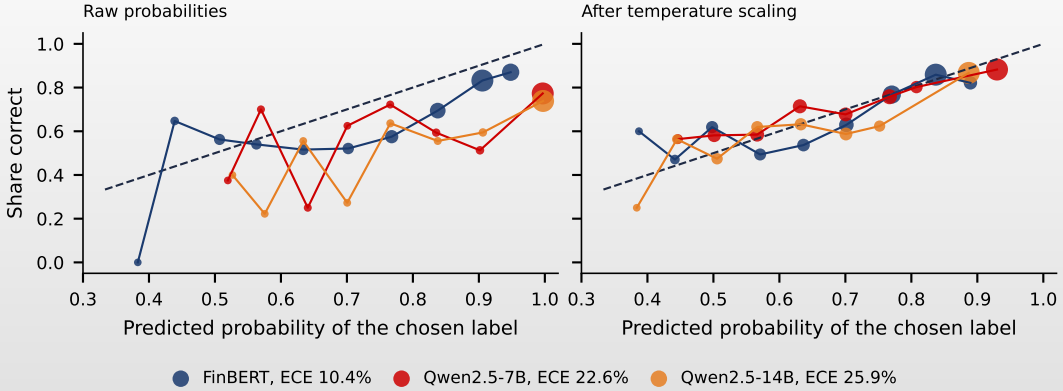
- **Questions**
  - how does the temperature T change an LLM's probabilities, its score and its calibration?
  - which T fits the data?
- **Inputs**
  - One Twitter headline labelled **neutral**; Qwen2.5-7B (P1) gives $p_{\mathbf{neg}} = 0.608$, $p_{\mathbf{neu}} = 0.392$, $p_{\mathbf{pos}} = 0.000$ (rounded to 0 in the file, replaced numerically by $1.0 \times 10^{-\mathbf{12}}$)
  - A random half of the validation set ($n = 1\,194$, seed 15): logits $z_{ik} = \ln p_{ik}$, rescaled $p_{ik}(T)$ as on the From Logits to Probabilities slide, loss $\mathrm{NLL}(T)$ Guo et al. (2017) (the Calibration slide)
- **Tasks**
  (a) Give the label, the net score s and the logit gap $\ln(p_{\mathbf{neu}}/p_{\mathbf{neg}})$ of the headline, then its probabilities and s at $T = 2$.
  (b) Explain why labels and accuracy do not depend on T, while the Brier score does.
  (c) Derive $\mathrm{NLL}'(T)$, evaluate the first-order condition at $T = 1$ and solve for T^* (Brent's method on $[0.5, 50]$).
  (d) On the other half, compute ECE and Brier before and after scaling.
- **Report**: the label, s, $s(T)$, T^*, ECE and Brier before and after
- **Interpretation**
  - is the LLM over- or under-confident?
  - what happens to a rule with a threshold on s?

Back Exercise list

A7 [Solved]: Solution (1/3)

- (a) Largest probability p_{neg}: label negative, wrong; $s = 0.000 - 0.608 = -0.608$; logit gap $\ln(0.392/0.608) = -0.439$ (-0.437 from the unrounded probabilities)
- At $T = 2$, $p_k(T) \propto \sqrt{p_k}$: $\sqrt{p} = (0.780,\ 0.626,\ 1.0 \times 10^{-6})$, sum 1.406, $p(T) = (0.554,\ 0.446,\ 0.000)$, $s(T) = -0.554$: same label, score pulled towards 0
- (b) $z \mapsto z/T$ is increasing for $T > 0$: the order of the logits, hence arg max and accuracy, do not change; the probabilities, and with them Brier and NLL, do

A7 [Solved]: Solution (2/3)



MFM_ch15_valid_inference

A7 [Solved]: Solution (3/3)

- (c) $\ln p_{iy}(T) = z_{iy}/T - \ln \sum_k e^{z_{ik}/T}$ gives $\mathrm{NLL}'(T) = \frac{1}{T^2} \sum_i \left(z_{iy_i} - \sum_k p_{ik}(T) z_{ik}\right)$
- At $T = 1$ the average bracket is $-2.47 < 0$, so NLL falls as T grows; the root is $T^* = 8.43$
- The file's probabilities are rounded to 6 decimals (2 539 zeros in 1 750 headlines): with the floor 10^{-6} instead of 10^{-12}, $T^* = 5.58$; the overconfidence remains
- (d) ECE falls from 22.6% to 4.3%, Brier from 0.468 to 0.366 (chart: raw and rescaled reliability)
- Interpretation: raw LLM probabilities are heavily overconfident; a threshold on s selects different headlines after rescaling

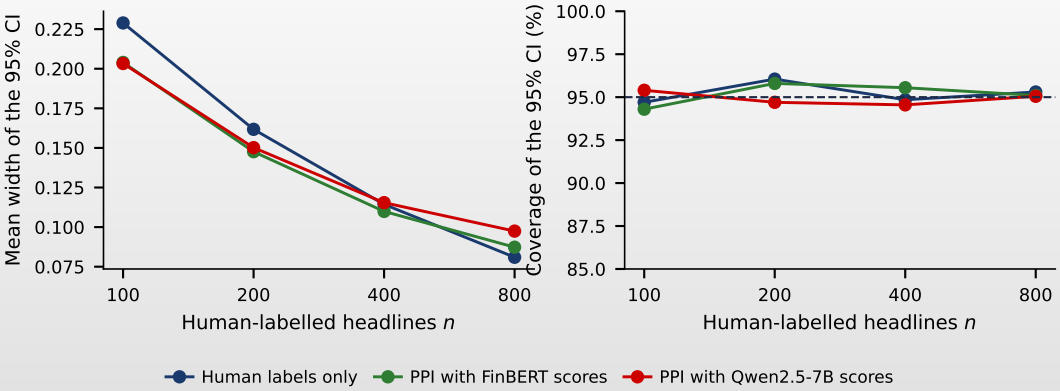
MFM_ch15_valid_inference

B8 [Solved]: Prediction-Powered Inference with a Few Human Labels — Task

- **Question**: can a few hundred human labels plus a biased model give a valid and narrower interval for the mean tone of the headlines?
- **Data**
  - Target θ = mean human net tone $(-1, 0, 1)$ of the 2 388 Twitter validation headlines, which act as the population; f = net score of FinBERT or of Qwen2.5-7B, fixed
  - Design: n labelled headlines (with Y) and $N = 2\,388 - n$ unlabelled ones (only f), drawn independently with replacement
- **Tasks**
  (a) For one draw with $n = 200$, compute the labels-only estimate $\bar{Y}_n \pm 1.96\, s_Y/\sqrt{n}$ ($\bar{Y}_n$, s_Y: mean and standard deviation of the n labels).
  (b) Compute the PPI estimate Angelopoulos et al. (2023): $\hat{\theta}$ (the Holm, PPI and κ slide), $\text{SE} = \sqrt{\hat{V}(f)/N + \hat{V}(f - Y)/n}$, for both models.
  (c) For $n = 100, 200, 400, 800$, repeat 2 000 times and record the coverage of θ and the mean CI width.
- **Report**: the one-draw estimates and a coverage and width chart
- **Interpretation**: why is the plain model mean $\frac{1}{N}\sum f$ not an estimate of θ?

Back Exercise list

B8 [Solved]: Solution (1/2)



MFM_ch15_valid_inference

B8 [Solved]: Solution (2/2)

- True $\theta = 0.054$; plain model means FinBERT -0.014 (wrong sign), Qwen $+0.101$: the bias does not shrink as N grows
- (a) Labels only: $0.125 \pm 1.96 \times 0.040$; (b) PPI-FinBERT: mean of $f - Y$ on the labelled texts -0.068, $\hat{\theta} = 0.065 \pm 1.96 \times 0.037$; PPI-Qwen: mean of $f - Y$ $+0.036$, $\hat{\theta} = 0.075 \pm 1.96 \times 0.039$
- (c) $n = 100$: coverage 94.7%, 94.3%, 95.4%, width 0.229, 0.204, 0.203 (labels, PPI-FinBERT, PPI-Qwen)
- Reduction of the asymptotic variance ($N \to \infty$), $1 - \mathrm{Var}(f - Y)/\mathrm{Var}(Y)$: 24% (FinBERT) and 27% (Qwen)
- At $n = 800$ PPI is wider: 0.087 (FinBERT) and 0.097 (Qwen), against 0.081 from labels only
- Interpretation: the model mean converges to the wrong number; the labelled sample corrects it, and PPI is narrower only when $\mathrm{Var}(f - Y) < \mathrm{Var}(Y)$ and many texts are unlabelled

MFM_ch15_valid_inference

B9 [Proposed]: Wild Cluster Bootstrap with 16 Stocks — Task

- **Model**: B2 [Solved]
- **Question**: with only 16 stock clusters, which of the nine slopes of B2 remain significant?
- **Data**
  - FNSPID panel of B2: 18 282 stock-days, 16 US stocks; score $z_{i,d}$ and excess return r^e as in B2
  - The nine regressions of B2 (three scores $\times$ $h = 0, 1, 2$), SE clustered by stock ($G = 16$)
- **Tasks**
  (a) Compute the p-value of each slope with $t_{\mathbf{15}}$ critical values.
  (b) Run the wild cluster bootstrap with the null imposed Cameron, Gelbach & Miller (2008) (the Holm, PPI and κ slide), Rademacher $w_g = \pm 1$, $g(i)$: the stock of observation i, $\bar{y}$: the fit under H_0 (intercept only), 9 999 draws.
  (c) Compute t with two-way clustering: each component with its own $G/(G-1)$, without the $(n-1)/(n-2)$ factor of B2; the intersection component, one row per stock-day, is the heteroskedasticity-robust covariance.
  (d) Adjust the nine bootstrap p-values with Holm at 5%.
- **Report**: a table with t, the three p-values and the two-way t
- **Interpretation**: does Qwen at $h = 2$ survive?

Back Exercise list

B10 [Proposed]: Specification Curve for the Sentiment Strategy — Task

- **Model**: the Bootstrap Confidence Intervals slide
- **Question**: over all reasonable variants of the strategy, is there more evidence of predictability than chance would give?
- **Data**
  - The sentiment strategy of A3 (the Sentiment Strategy slide): position $\text{sign}(s_{i,d})$ from the close of $d+1$, hedged with SPY, held one day; four trades on each day with positions; 3 568 trading days
  - Grid: three scores $\times$ thresholds $|s| > b$, $b \in \{0, 0.25, 0.5\}$, $\times$ return day $d+2, \ldots, d+5$: 36 strategies
  - Gross returns, before costs, and thresholds on the raw scores; for day $d+k$, the position opens at the close of $d+k-1$ and is held one day
- **Tasks**
  (a) Compute the mean daily return and t_{NW} of each strategy and plot the sorted t-statistics Simonsohn, Simmons & Nelson (2020).
  (b) Simulate the joint null: flip the signs of the positions in blocks of 10 consecutive trading days (zero-return days included) (dependent wild bootstrap Shao (2010)), the same flips for all 36, 2 000 times.
  (c) Compute the joint p-values of the number of $t > 1.96$ and of the median t.
- **Report**: the sorted curve and the two joint p-values
- **Interpretation**: what would a "best specification" report hide?

Back Exercise list

A8 [Proposed]: The Qwen Strategy after Costs — Task

- **Model**: A3 [Solved]
- **Question**: the Qwen2.5-7B strategy has a statistically positive gross mean: is it tradable?
- **Inputs**
  - The sentiment strategy of A3 (the Sentiment Strategy slide): position $\text{sign}(s_{i,d})$ from the close of $d+1$, hedged with SPY, held one day; four trades on each day with positions; 3 568 trading days
  - Qwen2.5-7B signal: $\hat{\mu} = 5.64$ bp, $t_{NW} = 2.26$ ($L = 8$), $a = 0.950$, gross Sharpe ratio 0.63
- **Tasks**
  (a) Compute the break-even cost c^*.
  (b) Compute $\text{SE}(c^*)$ by the delta method and the 95% CI.
  (c) Recover σ and approximate the net Sharpe ratio at 5 bp, holding the gross volatility fixed as in A3.
- **Report**: c^*, SE, CI, net Sharpe ratio
- **Interpretation**: does a significant gross mean imply a profit after costs?

Back Exercise list

B11 [Proposed]: Event Study with the Loughran–McDonald Tone — Task

- **Model**: B2 [Solved]
- **Questions**
  - around negative and positive news days, when does the price move?
  - does the dictionary separate the two groups as well as FinBERT?
- **Data**
  - FNSPID panel of B2: 18 282 stock-days, 16 US stocks; score $z_{i,d}$ and excess return r^e as in B2
  - Keep repeated news days: windows for the same stock may overlap. Average first by event date and use the stated Newey–West correction on that date series.
  - The daily LM tone is exactly 0 on 48% of stock-days: groups tone < 0 (5 196 stock-days) and tone > 0 (4 288); FinBERT: bottom and top terciles
- **Tasks**
  (a) For each date, average the excess returns of days $-5, \ldots, +5$ over that date's stock-days in each group.
  (b) Compute the mean over dates of the CAR for $[-5, -1]$, day 0, day 1 and $[+2, +5]$, with t_{NW}.
  (c) Plot the mean cumulative paths of the four groups.
- **Report**: the table of the four windows and the chart
- **Interpretation**: is the post-event return positive in both groups? Distinguish survivorship bias (the 16 stocks are today's large firms) from the reversal mechanism in Tetlock (2007).

Back Exercise list

C2 [Proposed]: Does Headline Sentiment Pay after Costs? — Task

- **Model**: A3, B2 [Solved]
- **Question**: does a daily strategy on FNSPID headline sentiment earn money after realistic costs?
- **Data**
  - The sentiment strategy of A3 (the Sentiment Strategy slide): position $\text{sign}(s_{i,d})$ from the close of $d+1$, hedged with SPY, held one day; four trades on each day with positions; 3 568 trading days
  - Positions equally weighted; days without news: return 0; 5 bp per trade; scores FinBERT, Qwen2.5-7B, LM from the saved tables
- **Tasks**
  (a) For FinBERT, compute the mean daily gross return, t_{NW} and a moving-block bootstrap 95% CI (blocks of 10 days, 2 000 draws).
  (b) Compute the Sharpe ratio before and after costs, the break-even cost and the Sharpe ratio at 0, 1, 2, 5, 10 bp.
  (c) Repeat for Qwen2.5-7B and LM, and for strong news $|s| > 0.5$ (exploratory); Holm over the three unfiltered strategies, or over all six if you select among them.
  (d) Further step: correct the same-day slope with a random human-labelled sample of stock-days (divide by $\rho(s^*, \hat{s})$, or rectify the OLS moments as in B8); 200 stock-days drawn with seed 15 before looking at returns, labelled independently by two annotators.
- **Report**: a two-page note with a results table and charts, and a Quantlet folder
- **Interpretation**: discuss timing, look-ahead (B3), survivorship of the 16 stocks and crowding Lopez-Lira & Tang (2026)

Back Exercise list

C3 [Proposed]: Replicating Lopez-Lira and Tang on FNSPID — Task

- **Model**: C2, B2, B3, B9
- **Question**: does the published GPT-4 (Generative Pre-trained Transformer 4) design Lopez-Lira & Tang (2026), run with an open model on the course headlines, find the same reaction and the same drift?
- **Inputs**
  - Paper: GPT-4 answers `YES`, `NO` or `UNKNOWN` about each headline; $s = 1, -1, 0$ drives a daily long-short portfolio; overnight drift 0.34% per day, Sharpe ratio 2.97 (Table 1)
  - Data: FNSPID headlines of the 16 stocks, adjusted daily closes of the course data; first safe return $d + 2$
- **Tasks**
  (a) Score every headline, parse the first line only, map `YES`/`UNKNOWN`/`NO` to $1/0/-1$; report the shares and invalid answers.
  (b) Average by stock-day and form the daily portfolio with the conventions of the next slide, earning the returns of day $d+2$.
  (c) Report the mean daily return in bp, t_{NW}, the hit rate, the Sharpe ratio, and the shares of days with one leg and with no position.
  (d) Estimate the paper's Eq. (1), $r_{i,d+h} = a_i + \delta_d + b\, s_k + \varepsilon_k$ over headlines k, with stock and date fixed effects a_i and δ_d, for $h = 0, 2$, SE clustered by date and stock and the wild cluster bootstrap p (B9).
  (e) Recompute the Sharpe ratio net of 5, 10, 20 bp per round trip (Figure 4) and compare with C2 on the same stock-days.
- **Report**: your numbers next to Tables 1 and 2 of the paper, cumulative returns before and after costs, a Quantlet folder
- **Interpretation**
  - which differences from the paper can your data separate?
  - what does B3 say about the bias?

Back Exercise list

C3 [Proposed]: Portfolio, Calendar and Cost Conventions

- Fixed before any run
  - Model: Qwen2.5-7B-Instruct Qwen Team (2024), temperature 0, the prompt of Section 5 of the paper word for word
  - Weights (Table 1): long leg = equal-weighted mean return of the stocks with $s > 0$, short leg = same for $s < 0$, a leg only with at least two stocks; long-short = long minus short (gross exposure 2); if only one leg qualifies, hold it alone (footnote 20)
  - Returns: raw stock returns, as in Table 1; the SPY-hedged version (as in C2) is a course adaptation, reported as a sensitivity: with both legs the hedge cancels, so it matters only on one-leg days
  - Calendar: all trading days, as in C2, with return 0 without a position; mean, t_{NW} and Sharpe ratio over all days; hit rate over days with a position
  - Costs (Figure 4): c bp per round trip on each unit of notional: $2c$ per day with two legs, c with one, plus, in the hedged version, c on the net SPY notional (1 on one-leg days)
  - Regressions (d): raw returns in % per day, as in Table 2; the date fixed effects absorb the market return; the wild cluster bootstrap keeps both fixed effects in the null model

Acronyms used in this material

- **AI**: Artificial Intelligence
- **AUC**: Area Under the (ROC) Curve
- **BERT**: Bidirectional Encoder Representations from Transformers
- **BVB**: Bursa de Valori București — *the Bucharest Stock Exchange*
- **CAR**: Cumulative Abnormal Return
- **CI**: Confidence Interval
- **ECE**: Expected Calibration Error
- **FNSPID**: Financial News and Stock Price Integration Dataset
- **GPT-4**: Generative Pre-trained Transformer 4
- **LLM**: Large Language Model
- **LM**: Loughran-McDonald (financial sentiment dictionary)
- **LR**: Logistic Regression
- **NLL**: Negative Log-Likelihood
- **OLS**: Ordinary Least Squares
- **PPI**: Prediction-Powered Inference
- **ROC**: Receiver Operating Characteristic (curve)
- **SD**: Standard Deviation
- **SE**: Standard Error
- **TF-IDF**: Term Frequency - Inverse Document Frequency
- **US**: United States
- **UTC**: Coordinated Universal Time (Temps Universel Coordonné)

References (1/3)

- Aigner, D.J. (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1(1), 49–59.
- Angelopoulos, A.N., Bates, S., Fannjiang, C., Jordan, M.I., Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671), 669–674.
- Araci, D. (2019). FinBERT: financial sentiment analysis with pre-trained language models. arXiv:1908.10063.
- Cameron, A.C., Gelbach, J.B., Miller, D.L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.
- Dawid, A.P., Skene, A.M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1), 20–28.
- DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 837–845.
- Dong, Z., Fan, X., Peng, Z. (2024). FNSPID: a comprehensive financial news dataset in time series. *Proceedings of the 30th ACM SIGKDD Conference*, 4918–4927.
- Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.

References (2/3)

- Franke, J., Härdle, W.K., Hafner, C.M. (2019). Statistics of Financial Markets: An Introduction, 5th ed. Springer.
- Glasserman, P., Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. arXiv:2309.17322.
- Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q. (2017). On calibration of modern neural networks. *ICML*; arXiv:1706.04599.
- Lopez-Lira, A., Tang, Y. (2026). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Journal of Financial Economics*, 184, 104335.
- Loughran, T., McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.
- Malo, P., Sinha, A., Korhonen, P., Wallenius, J., Takala, P. (2014). Good debt or bad debt: detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782–796.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
- Newey, W.K., West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.

References (3/3)

- Petersen, M.A. (2009). Estimating standard errors in finance panel data sets: comparing approaches. *Review of Financial Studies*, 22(1), 435–480.
- Qwen Team (2024). Qwen2.5 technical report. arXiv:2412.15115.
- Sarkar, S.K., Vafa, K. (2024). Lookahead bias in pretrained language models. SSRN Working Paper 4754678.
- Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489), 218–235.
- Simonsohn, U., Simmons, J.P., Nelson, L.D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214.
- Tetlock, P.C. (2007). Giving content to investor sentiment: the role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168.
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M. (2020). MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained Transformers. *NeurIPS*; arXiv:2002.10957.