








Modelarea piețelor financiare

Seminarul 15: LLM-uri și analiza sentimentului — aplicații



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

Academia de Studii Economice din București
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Academia Română, Institutul de Prognoză Economică
MSCA Digital Finance

Întrebările de azi și structura seminarului

- Întrebările seminarului
  - **poate un model de limbaj să citească sentimentul financiar?**
  - **poate acest sentiment să prognozeze randamente care mai pot fi tranzacționate?**
- Modelele și datele
  - LM: dicționarul Loughran–McDonald, care numără cuvinte financiare fără context
  - FinBERT: BERT (Bidirectional Encoder Representations from Transformers) adaptat textelor financiare; Qwen2.5: un model mare de limbaj, LLM (*Large Language Model*), întrebat zero-shot, fără exemple etichetate
  - FNSPID (Financial News and Stock Price Integration Dataset): titluri legate de prețurile acțiunilor; SPY (*SPDR S&P 500 ETF Trust*): ETF-ul (*Exchange-Traded Fund*, fond tranzacționat la bursă) pe indicele S&P 500 (*Standard & Poor's* 500), reperul randamentelor în exces
- Livrabile
  - răspunsurile la exercițiile A
  - un tabel și un grafic pentru fiecare exercițiu B
  - corecturile din C1
  - o propoziție de interpretare pentru fiecare exercițiu
- Notebook-ul de azi: deschideți notebook-ul seminarului în Colab
  - fiecare cerință indică secțiunea din notebook
- Seminar: prof. Daniel Traian Pele, danpele@ase.ro; drd. Antoaneta Amza, antoaneta.amza@csie.ase.ro

Harta exercițiilor

Exercițiu	Întrebare	Tip	Model
A1	cît micșorează panta o variabilă binară clasificată greșit?	Rezolvat	–
A2	cît de precis și cît de puternic este un test de acuratețe pe perechi?	Rezolvat	–
A3	de la ce cost pe tranzacție nu mai este profitabilă strategia?	Rezolvat	–
B1	este FinBERT mai precis decît dicționarul Loughran–McDonald?	Rezolvat	–
B2	prognozează titlurile randamentele sau le descriu?	Rezolvat	–
B3	își amintește LLM-ul piața?	Rezolvat	–
C1	ce este greșit într-un răspuns AI (*Artificial Intelligence*, inteligență artificială)?	Propus	B1, B2, B3
B4	este un LLM cu 7 miliarde de parametri mai bun decît FinBERT? (temă)	Propus	B1
Pachete de extindere	A4–A8, B5–B11, C2, C3 (opționale, după curs)	ambele	indicat la fiecare enunț

- **[Rezolvat]**: rezolvarea completă în slide-uri și în notebook, un model de urmat; **[Propus]**: îl rezolvați dumneavoastră; rezolvarea se discută la seminar. Rezolvați B4 înainte de a folosi numerele discordante date în A6
- Partea A: derivări pe hîrtie; Partea B: verificare și inferență pe date reale, fiecare cu o întrebare de interpretare; Partea C: întrebări deschise și critica unui răspuns AI (*Artificial Intelligence*, inteligență artificială)

Datele folosite

Date	Conținut	Scoruri salvate	Folosite în
Twitter Financial News (setul de validare, Hugging Face Hub)	2 388 de titluri din 2022: 1 566 neutre, 475 pozitive, 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 de propoziții din știri, 2 264 cu acord deplin	aceleași scoruri	B5
FNSPID Dong, Fan & Peng (2024)	106 820 de titluri despre 16 acțiuni americane, 27.10.2009–18.12.2023	scorurile zilnice LM, FinBERT, Qwen2.5-7B	A3, A8, B2, B9–B11, C1–C3
Acțiunile și SPY	prețurile ajustate zilnice din datele cursului, cîte o serie pe activ (cîmpul folosit: închiderea ajustată, *adjusted close*), 2009–2023	–	B2, B11, C2, C3
S&P 500	închideri lunare, ian. 2000 – aug. 2026	răspunsurile Qwen2.5 despre fiecare lună	B3, C1

- Scorurile sînt salvate împreună cu Quantlet-urile capitolului, `Quantlets/Ch_15`: la seminar nu trebuie rulat niciun model de limbaj
- C3 (pachet de extindere) cere o nouă evaluare a titlurilor FNSPID

Pregătire: rulați și verificați

```
# 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())
```

- Rezultatul așteptat
  - 4 846 și 2 388 de rînduri
  - neutral 1 566, positive 475, negative 347
  - (18 282, 19) și 16 acțiuni
- Fiecare rînd din `tw` conține textul, eticheta umană și probabilitățile fiecărui model (de ex. `finbert_negative`, `qwen7B_positive`)
- Dacă cifrele diferă, opriți-vă și verificați fișierele înainte de orice exercițiu

Notații folosite în toate exercițiile

- Tonul după dicționar τ și listele Loughran–McDonald (LM) Loughran & McDonald (2011)
- Probabilitățile modelului p_k, eticheta $\hat{y}$ și scorul net $s = p_{\mathbf{pos}} - p_{\mathbf{neg}}$ (FinBERT Araci (2019), Qwen2.5 Qwen Team (2024))
- Doi clasificatori A și B pe aceleași texte: numerele discordante $n_{\mathbf{01}}$, $n_{\mathbf{10}}$ și testul McNemar McNemar (1947)
- Ziua d, orizontul h și primul randament tranzacționabil, $d + 2$
- Unități și inferență
  - bp: punct de bază, 0,01%; SE: eroare standard; CI: interval de încredere de 95%
  - t_{NW}: t Newey–West Newey & West (1987); SE grupate Petersen (2009)

Trei moduri de a citi sentimentul

- **Dicționar** (*bag of words*, reprezentarea prin frecvențele cuvintelor) Loughran & McDonald (2011): tonul unui text

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

  - P, N: numărul de cuvinte ale textului găsite într-o listă fixă pozitivă, respectiv negativă
  - Clasa = semnul lui τ ($\tau = 0$: neutru); exemplu $P = 2$, $N = 1$: $\tau = 1/3$, pozitiv
  - Listele LM, construite din rapoartele 10-K: 2 355 de cuvinte negative și doar 354 pozitive; transparent, dar insensibil la context și la negație
- **FinBERT** Araci (2019): BERT, un encoder Transformer (Capitolul 14), antrenat în continuare pe știri financiare
  - Adaptat prin reantrenare pe propoziții etichetate (*fine-tuning*); rezultatul: $p_{\mathbf{pos}}$, $p_{\mathbf{neu}}$, $p_{\mathbf{neg}}$, cu suma 1
  - Propozițiile de adaptare provin din Financial PhraseBank Malo et al. (2014): acuratețea lui acolo este parțial în eșantion (B5)
- **LLM zero-shot** (Qwen2.5 Qwen Team (2024)): un model mare de limbaj întrebat „Is this headline positive, negative or neutral?"
  - Fără antrenare pe aceste etichete (zero-shot); citim logiții z_k ai celor trei cuvinte-răspuns (slide-ul următor)

De la logiți la probabilități

- Logitul z_k: scorul nenormalizat al modelului pentru răspunsul $k \in \{\text{pos}, \text{neu}, \text{neg}\}$; mai mare înseamnă mai probabil
- **Softmax cu temperatura** $T > 0$: transformă logiții în probabilități

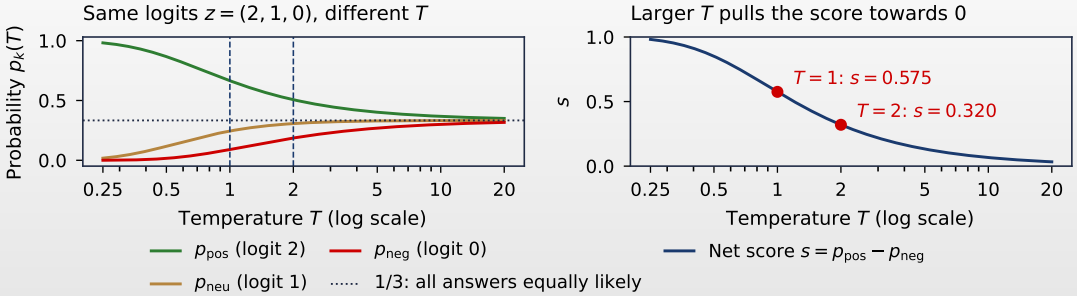
$$p_k(T) = \frac{e^{z_k/T}}{\sum_j e^{z_j/T}}, \qquad \sum_k p_k(T) = 1$$

  - $T = 1$: probabilitățile proprii ale modelului; $T > 1$ le aplatizează spre $1/3$; $T < 1$ le accentuează
  - Contează doar diferențele de logiți: $z_k - z_j = T \ln(p_k/p_j)$, diferența de logiți (A7)
- **Eticheta și scorul net**

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

  - $\arg\max_k$: răspunsul cu probabilitatea cea mai mare; T nu îl schimbă, deoarece împărțirea la $T > 0$ păstrează ordinea logiților
  - $s > 0$: ton pozitiv; s apropiat de 0: neutru sau nedecis; s se schimbă cu T

Efectul temperaturii



- Exemplu: logiții $z = (2, 1, 0)$ pentru (pos, neu, neg)
  - $T = 1$: $p = (0{,}665;\ 0{,}245;\ 0{,}090)$, $s = 0{,}575$
  - $T = 2$: $p = (0{,}506;\ 0{,}307;\ 0{,}186)$, $s = 0{,}320$: aceeași etichetă, scor împins spre 0
- Cînd $T \to \infty$, cele trei probabilități tind la $1/3$ și $s \to 0$; o regulă cu prag pe s depinde deci de T

Evaluarea unui clasificator

- Matricea de confuzie C: C_{jk} = numărul de texte cu eticheta umană j (rîndul, adevărul) și eticheta prezisă k (coloana); n texte, K clase
- **Acuratețea**: proporția textelor de pe diagonală

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

  - Reperul: proporția clasei majoritare (o prezicem mereu), niciodată $1/K$
- **Precizia** și **recall**-ul clasei k

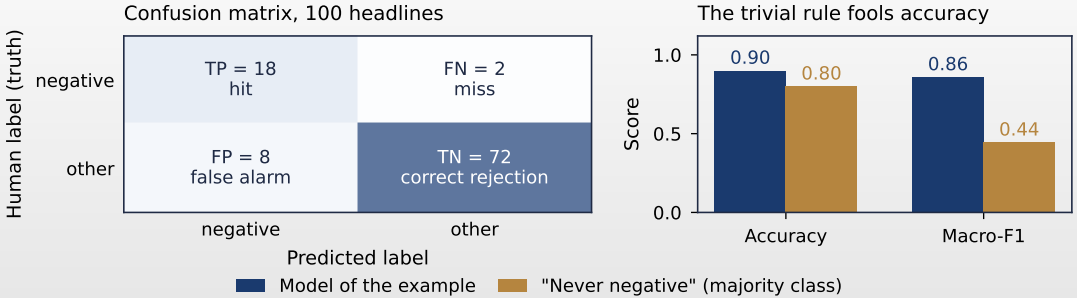
$$\text{prec}_k = \frac{C_{kk}}{\sum_j C_{jk}}, \qquad \text{rec}_k = \frac{C_{kk}}{\sum_j C_{kj}}$$

  - prec: proporția textelor marcate k care sînt cu adevărat k (coloana); rec: proporția textelor k care sînt marcate (rîndul)
- $F_{\mathbf{1}}$ și **macro-F1**

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

  - $F_{\mathbf{1}}$: media armonică, mică dacă oricare dintre ele este mică; macro-F1 dă unei clase rare aceeași pondere ca uneia frecvente

Acuratețea, macro-F1 și clasa majoritară



- Exemplu: 100 de titluri, 20 cu adevărat negative; modelul marchează 18 dintre ele și 8 dintre celelalte 80
  - Acuratețea $(18 + 72)/100 = 0{,}90$, dar „niciodată negativ" are deja 0,80; macro-F1: $(0{,}783 + 0{,}935)/2 = 0{,}86$ față de $(0 + 0{,}889)/2 = 0{,}44$
  - Ratele folosite în A1: $\pi = 0{,}20$ cu adevărat negative, $p = 0{,}26$ marcate, ratări $\alpha_1 = 2/20 = 0{,}10$, alarme false $\alpha_0 = 8/80 = 0{,}10$

Eroarea de măsurare și atenuarea

- Panta estimată prin metoda celor mai mici pătrate, OLS (*Ordinary Least Squares*), pe un singur regresor x: în eșantioane mari, un raport de momente

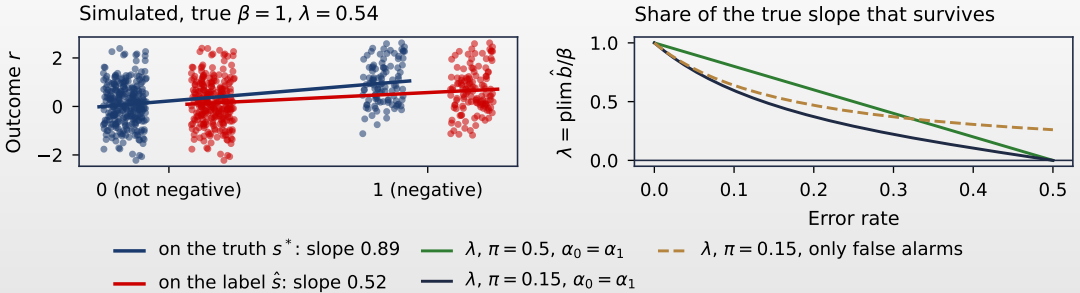
$$\text{plim}_{n\to\infty}\, \hat{b} = \frac{\text{Cov}(r, x)}{\text{Var}(x)}$$

  - plim: limita în probabilitate; r: variabila explicată (un randament)
- Eroare **clasică**: observăm $x = x^* + u$, cu u independent de x^* și de ε

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

  - $\lambda \in (0, 1]$: panta este deplasată spre zero (atenuare); $1/\lambda$ o corectează
- **Variabilă binară (*dummy*) clasificată greșit** Aigner (1973): eticheta $\hat{s} \in \{0, 1\}$ pentru adevăratul $s^* \in \{0, 1\}$
  - $\pi = P(s^* = 1)$; alarme false $\alpha_{\mathbf{0}} = P(\hat{s} = 1 \mid s^* = 0)$; ratări $\alpha_{\mathbf{1}} = P(\hat{s} = 0 \mid s^* = 1)$
  - Probabilitatea totală: $p = P(\hat{s} = 1) = \pi(1 - \alpha_{\mathbf{1}}) + (1 - \pi)\alpha_{\mathbf{0}}$
  - $u = \hat{s} - s^*$ poate fi -1 doar dacă $s^* = 1$: corelată cu s^*, deci nu este clasică; A1 deduce λ
  - Eroare nediferențială: dat fiind s^*, $\hat{s}$ nu aduce informație suplimentară despre r

Atenuarea într-o simulare



- Stînga: 400 de texte simulate, $\pi = 0{,}25$, $\alpha_0 = 0{,}15$, $\alpha_1 = 0{,}25$, $\beta = 1$: panta pe etichetă este aproximativ jumătate din panta pe adevăr
- Dreapta: λ în funcție de rata de eroare; $\lambda = 0$ la 50% erori (aruncarea unei monede); o clasă rară (π mic) pierde mai mult

Doi clasificatori pe aceleași texte

- Date împerecheate: pentru textul i, $D_i = 1$ dacă A greșește și B are dreptate, -1 dacă A are dreptate și B greșește, 0 altfel
  - n_{01}, n_{10}: numărul de texte cu $D_i = 1$ și cu $D_i = -1$ (textele discordante); textele pe care ambii le nimeresc sau ambii le greșesc dau 0
- **Diferența de acuratețe** și varianța ei (deduse în A2)

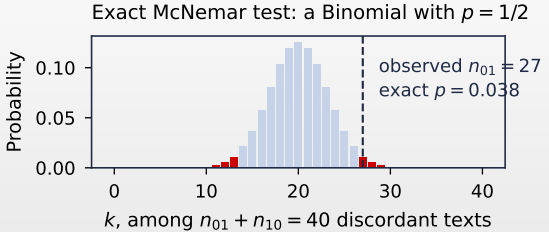
$$\hat{d} = \frac{n_{01} - n_{10}}{n}, \qquad \mathrm{Var}(\hat{d}) = \frac{q - d^2}{n}$$

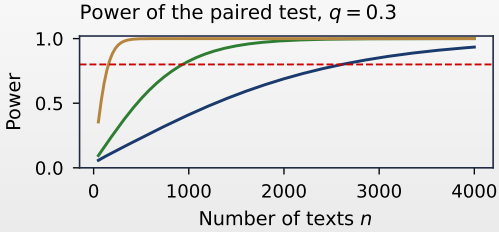
  - $d = \mathrm{acc}_B - \mathrm{acc}_A$; $q = P(D_i \neq 0)$, proporția textelor discordante: doar ele adaugă variantă
- **Testul McNemar exact** McNemar (1947): sub H_0 (acuratețe egală), fiecare text discordant îl favorizează pe B cu probabilitatea $\frac{1}{2}$

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

  - Varianta pentru eșantioane mari: $\chi^2 = (n_{01} - n_{10})^2/m$, comparată cu 3,84, cuantila de 95% a distribuției χ^2_1

Testul McNemar și puterea lui





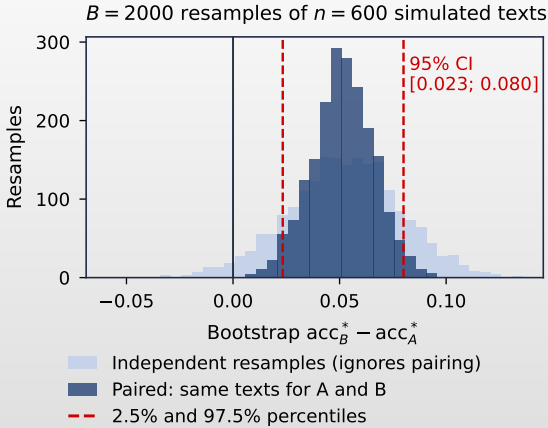
$P(n_{01} = k)$ under H_0: Bin(40, 0.5)
as or more extreme than observed
$d = 0.03$, $n_{80} = 2608$
$d = 0.05$, $n_{80} = 934$
$d = 0.12$, $n_{80} = 156$
80% power

- Puterea testului bilateral de 5% dacă diferența adevărată este d și numărul n care dă puterea 80%:

$$\text{puterea} \approx \Phi\Big(\frac{|d|\sqrt{n}}{\sqrt{q - d^2}} - 1{,}96\Big), \qquad n_{80} = \frac{(1{,}96 + 0{,}84)^2\,(q - d^2)}{d^2}$$

  - Φ: funcția de repartiție Normală standard; 0,84: cuantila ei de 80%; stînga: barele roșii = valoarea p exactă pentru $m = 40$, $n_{01} = 27$
  - Puterea planifică un studiu nou, cu efectul minim util fixat înainte de date; calculată la efectul observat, nu este o dovadă după test

Intervale de încredere bootstrap



- **Bootstrap percentil** pentru o statistică $\hat{\theta}$ (o acuratețe, o diferență)
  - Extragem n texte cu întoarcere, de B ori; recalculăm $\hat{\theta}_b^*$, $b = 1, \ldots, B$
  - CI 95% $= [q_{2{,}5\%};\ q_{97{,}5\%}]$, percentilele valorilor $\hat{\theta}_b^*$
- **Bootstrap împerecheat**: aceleași texte reeșantionate pentru A și B
  - Păstrează dificultatea comună a fiecărui text pentru A și B: un interval mai îngust pentru diferență (graficul)
- **Bootstrap pe blocuri** pentru zile (C2)
  - Reeșantionăm blocuri de zile consecutive, ceea ce păstrează dependența serială

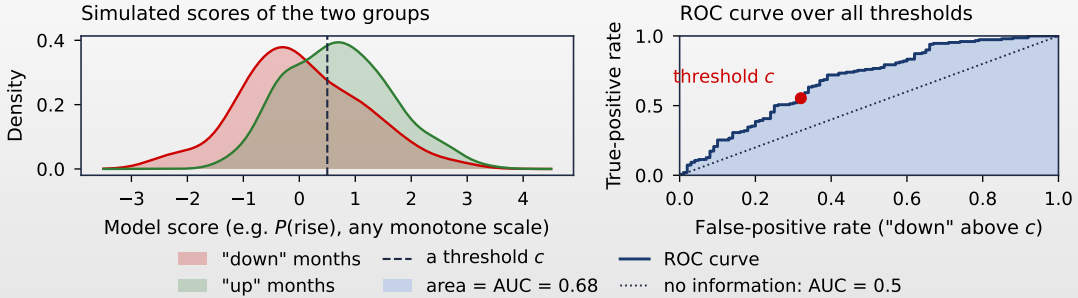
Calitatea ordonării: AUC

- Problema din B3: separăm lunile „în creștere" (n_1) de lunile „în scădere" (n_0) cu un scor S, de ex. $P(\text{creștere})$
- **AUC (Area Under the Curve)**: probabilitatea ca o lună în creștere aleasă la întîmplare să aibă un scor mai mare decît una în scădere

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

  - U: statistica Mann–Whitney, numărul perechilor ordonate corect; $1\{\cdot\}$: 1 dacă este adevărat, altfel 0
  - AUC $= 0{,}5$: nicio informație; 1: ordonare perfectă; contează doar ordinea scorurilor
- **Testul de permutare** pentru AUC $= 0{,}5$
  - Permutăm etichetele creștere/scădere de B ori; $p = (1 + \#\{\text{AUC}_b^* \geq \text{AUC}\})/(B+1)$, #: numărul
- **Două AUC independente** (înainte și după o dată), erorile standard DeLong DeLong et al. (1988)
  - $z = (\text{AUC}_1 - \text{AUC}_2)/\sqrt{\text{SE}_1^2 + \text{SE}_2^2} \sim N(0,1)$ sub H_0; diferența detectabilă cu puterea 80% la 5%: $2{,}80\sqrt{\text{SE}_1^2 + \text{SE}_2^2}$

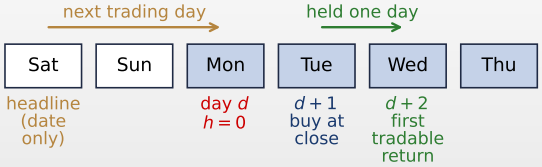
Scoruri, praguri și curba ROC



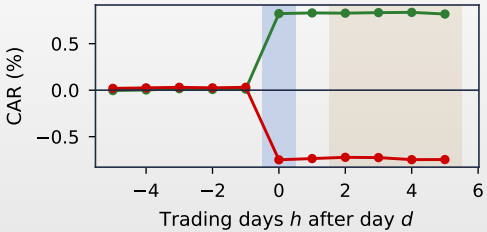
- Stînga: scoruri simulate pentru 150 de luni în creștere și 100 în scădere; un prag c declară „în creștere" orice lună de deasupra lui
- Dreapta: deplasarea lui c trasează curba ROC (Receiver Operating Characteristic), rata adevăraților pozitivi față de rata falșilor pozitivi; aria de sub ea este AUC (aici 0,68)

De la titluri la randamente: calendarul

From a calendar date to the tradable return



Simulated: all news priced on day d



- Ziua d: prima zi de tranzacționare de la data titlului încolo; ora nu se cunoaște, deci $d + 1$ poate conține încă reacția
- O pantă la $h = 0$ măsoară cît de bine *descriu* titlurile ziua; doar $h \geq 2$ este o *prognoză* (B2)
- Studiul de eveniment (B11): randamentul anormal cumulat, CAR (*Cumulative Abnormal Return*), $\mathrm{CAR}[h_1, h_2] = \sum_{h=h_1}^{h_2} \bar{r}^e_{d+h}$, suma randamentelor medii în exces (nu un indice de avere); dreapta: știri încorporate complet în ziua d

Regresia de panel și erorile standard grupate

- Regresia de panel pe zile-acțiune (acțiunea i, ziua d, orizontul h); eroarea $\varepsilon_{i,d}$ poartă indicele zilei știrii d, nu al zilei randamentului $d+h$

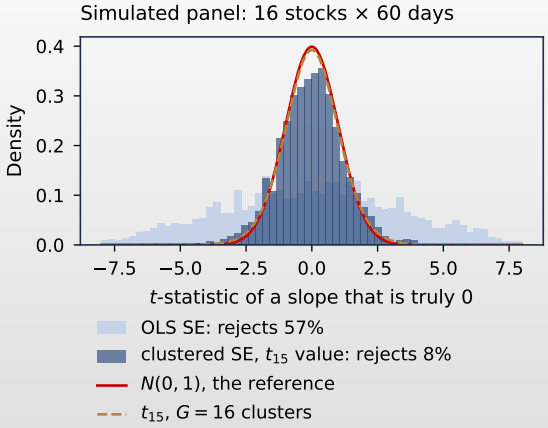
$$r^e_{i,d+h} = a + b\, z_{i,d} + \varepsilon_{i,d}$$

  - r^e: randamentul simplu în exces față de SPY, din prețurile ajustate; $z_{i,d}$: scorul zilnic standardizat pe panel; b: randamentul la o abatere standard a scorului (SD sau s.d. în grafice)
- Erorile sînt corelate în aceeași zi (știri comune) și pentru aceeași acțiune (în timp): erorile standard OLS sînt prea mici
- **SE grupate** Petersen (2009): însumăm produsele $X_g^\top \hat{u}_g$ în fiecare grup g (o zi sau o acțiune)

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

  - $\hat\beta = (\hat a, \hat b)$; X: cele n rînduri $(1, z_{i,d})$; X_g, $\hat{u}_g$: rîndurile și reziduurile OLS ale grupului g; c: corecția pentru eșantioane mici
  - Puține grupuri ($G = 16$ acțiuni): comparăm $t = \hat b/\mathrm{SE}$ cu t_{G-1} (2,13), nu cu 1,96, sau folosim un wild cluster bootstrap Cameron, Gelbach & Miller (2008)

De ce contează gruparea



- Simulare: 16 acțiuni × 60 de zile, panta adevărată 0; atît scorul, cît și eroarea persistă pentru aceeași acțiune
- SE OLS tratează cele 960 de rînduri ca independente
  - Statistica t este mult prea dispersată: 57% respingeri false la 5%
- SE grupate cu valoarea critică $t_{\mathbf{15}}$
  - Aproximativ 8% respingeri false: aproape de 5%, dar încă peste
  - Cu atît de puține grupuri, un wild cluster bootstrap este mai sigur (B9)
  - Întrebare pentru sală: de ce comparăm t cu $t_{\mathbf{15}}$, nu cu $N(0,1)$?
  - Răspuns: SE grupată se estimează din doar 16 sume pe grupuri, deci are circa $G - 1 = 15$ grade de libertate

O strategie de sentiment, statistica t și costurile

- Strategia (A3): la închiderea din $d+1$, o poziție de semnul lui $s_{i,d}$ în fiecare acțiune cu știri și $s_{i,d} \neq 0$, acoperită cu SPY și închisă după o zi; R_t: randamentul ei în ziua t (0 fără poziții)
- **Statistica t Newey–West** Newey & West (1987): t-ul unei medii cînd zilele sînt dependente serial

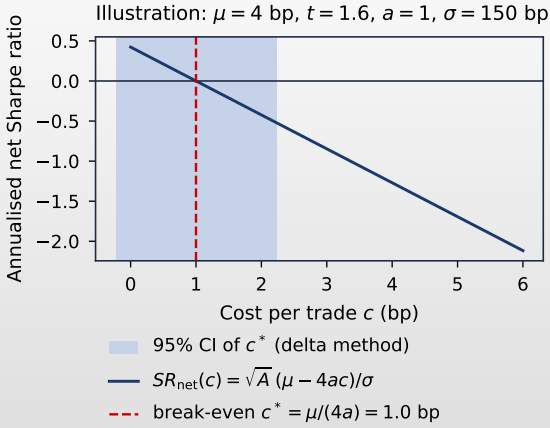
$$t_{NW} = \frac{\hat{\mu}}{\sqrt{\hat{\Omega}/T}}, \qquad \hat{\Omega} = \hat{\gamma}_{\mathbf{0}} + 2\sum_{l=\mathbf{1}}^{L}\Big(1 - \frac{l}{L+1}\Big)\hat{\gamma}_l$$

  - $\hat{\mu} = \bar{R}$; $\hat{\gamma}_l$: autocovarianța de selecție a lui R_t la lag-ul l; L: numărul de lag-uri; T: numărul de zile
- **Costurile**: c pe tranzacție, patru tranzacții în fiecare zi cu poziții, o proporție a de astfel de zile

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

  - $A = 252$ de zile de tranzacționare pe an; σ: abaterea standard a randamentelor zilnice brute; SR: raportul Sharpe; c^*: costul de echilibru
- **Metoda delta**: SE pentru o funcție de o estimație
  - $\mathrm{SE}(g(\hat{\mu})) \approx |g'(\hat{\mu})|\,\mathrm{SE}(\hat{\mu})$; cu $\mathrm{SE}(\hat{\mu}) = \hat{\mu}/t_{NW}$ și a fix: $\mathrm{SE}(c^*) = \mathrm{SE}(\hat{\mu})/(4a)$

Costul de echilibru: o ilustrare

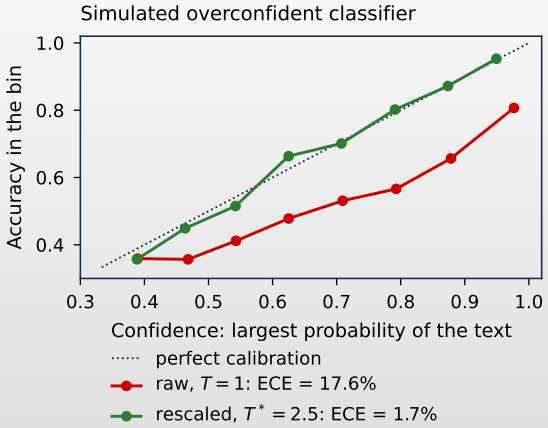


- Date ilustrative, nu cele din A3: $\mu = 4$ bp pe zi, $t_{NW} = 1{,}6$, $a = 1$, $\sigma = 150$ bp
- Raportul Sharpe net scade liniar în c și trece prin 0 la $c^* = 4/4 = 1$ bp
- $\mathrm{SE}(c^*) = (4/1{,}6)/4 = 0{,}63$ bp
  - CI 95% pentru c^*: $1 \pm 1{,}96 \times 0{,}63 = [-0{,}23;\ 2{,}23]$ bp
  - Un cost realist de 5 bp pe tranzacție se află mult la dreapta intervalului

Deplasarea de anticipare și încercările multiple

- **Deplasarea de anticipare** (*look-ahead bias*) Glasserman & Lin (2023): un LLM antrenat pe texte pînă la publicarea lui poate să fi citit ce a urmat după un titlu
  - Un backtest pe știri de dinaintea publicării amestecă prognoza cu memoria
  - Testul de memorie (B3): întrebăm modelul despre luni trecute fără nicio dată; un AUC peste 0,5 înseamnă că își amintește
  - Test fără contaminare: doar texte publicate după publicarea ponderilor (B3, C1)
- **Încercări multiple**: mai multe scoruri, orizonturi sau prompt-uri, apoi este raportat cel mai bun
  - Cu 20 de teste independente la 5%, probabilitatea a cel puțin unei respingeri false este $1 - 0{,}95^{\mathbf{20}} = 0{,}64$
  - Raportăm toate încercările sau corectăm valorile p (B7, B9, B10)

Calibrarea probabilităților



- Calibrarea (A7) Guo et al. (2017): dintre textele cu probabilitatea 0,8, aproximativ 80% ar trebui să fie corecte
- Logaritmul negativ al verosimilității, NLL (*Negative Log-Likelihood*): pierderea folosită pentru alegerea lui T
  - $\mathrm{NLL}(T) = -\sum_i \ln p_{iy_i}(T)$, y_i: eticheta adevărată a textului i
- Scorul Brier: eroarea pătratică medie a probabilităților
  - $\frac{1}{n}\sum_i \sum_k (p_{ik} - 1\{y_i = k\})^2$
- Eroarea medie de calibrare, ECE (*Expected Calibration Error*): $\sum_b \frac{n_b}{n} |\mathrm{acc}_b - \mathrm{conf}_b|$
  - Intervale b ale probabilității maxime; conf_b: media ei, acc_b: acuratețea, n_b: textele din interval
- Graficul: un model supraîncrezător se află sub diagonală; $T^* > 1$ îl aduce pe ea

Holm, wild cluster bootstrap, PPI și κ Fleiss

- **Corecția Holm** pentru M teste: ordonăm $p_{(1)} \le \cdots \le p_{(M)}$
  - $\tilde{p}_{(j)} = \max_{i \le j} \min\{1, (M-i+1)\, p_{(i)}\}$; respingem unde $\tilde{p} < 0{,}05$: cel mult 5% șanse de vreo respingere falsă
- **Wild cluster bootstrap** Cameron, Gelbach & Miller (2008) (B9)
  - $y_i^* = \bar{y} + w_{g(i)}(y_i - \bar{y})$, $w_g = \pm 1$ cu probabilitatea $\frac{1}{2}$ pentru fiecare grup; $p = P^*(|t^*| \ge |t|)$
  - Grupare dublă: $\hat{V} = \hat{V}_{\mathbf{zi}} + \hat{V}_{\mathbf{acțiune}} - \hat{V}_{\mathbf{zi} \cap \mathbf{acțiune}}$
- **Inferența susținută de predicții**, PPI (*Prediction-Powered Inference*) Angelopoulos et al. (2023) (B8): media modelului corectată cu deplasarea lui pe textele etichetate
  - $\hat{\theta} = \frac{1}{N}\sum_{\mathbf{neetich.}} f - \frac{1}{n}\sum_{\mathbf{etich.}} (f - Y)$; f: scorul modelului, Y: eticheta umană
  - N: textele neetichetate, n: cele etichetate, două eșantioane separate
  - Mai multe predicții nu repară un eșantion etichetat nereprezentativ
- κ **al lui Fleiss** Fleiss (1971) (A5): acordul evaluatorilor peste cel întîmplător
  - $\kappa = (\bar{P} - \bar{P}_e)/(1 - \bar{P}_e)$: $\bar{P}$ acordul observat, $\bar{P}_e$ cel întîmplător; 0 = întîmplare, 1 = perfect

A1 [Rezolvat]: atenuarea prin clasificare greșită — model și date

- **Întrebare**: o regresie folosește eticheta FinBERT „negativ" în locul adevărului; cu cît se micșorează panta?
- **Date**
  - Twitter Financial News, setul de validare: $N = 2\,388$ de titluri din 2022, fiecare etichetat negativ, neutru sau pozitiv; eticheta umană este considerată adevărul
  - Modelul $r = a + \beta s^* + \varepsilon$, $s^* = 1$ dacă titlul este cu adevărat negativ; observăm doar $\hat{s} = 1$ cînd FinBERT prezice negativ
  - Eroare nediferențială: dat fiind s^*, $\hat{s}$ nu aduce informație suplimentară despre r, deci $\mathrm{Cov}(\varepsilon, \hat{s}) = 0$
  - π, p, alarmele false α_0 și ratările α_1: ratele definite pe slide-ul Eroarea de măsurare

Matricea de confuzie FinBERT, Twitter

Adevărat \ prezis	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

Înapoi Lista exercițiilor

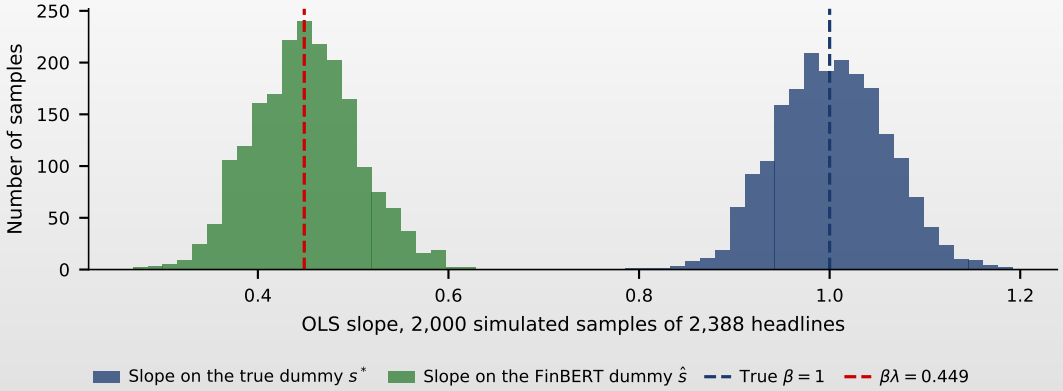
A1 [Rezolvat]: cerințe și interpretare

- **Cerințe**
  (a) Arătați că plim $\hat{b} = \beta\,\mathrm{Cov}(s^*, \hat{s})/\mathrm{Var}(\hat{s})$.
  (b) Deduceți $\lambda = \pi(1-\pi)(1-\alpha_0-\alpha_1)/[p(1-p)]$.
  (c) Calculați π, p, α_0, α_1 din matrice, apoi λ cu trei zecimale.
  (d) Calculați $\mathrm{Cov}(u, s^*)$ pentru $u = \hat{s} - s^*$: este eroarea clasică?
- **Raportați**: λ, $1/\lambda$ și semnul lui $\mathrm{Cov}(u, s^*)$
- **Interpretare**: ce parte din efectul adevărat rămîne într-o pantă pe variabila binară FinBERT?

A1 [Rezolvat]: rezolvare (1/3)

- (a) Pentru un singur regresor plim $\hat{b} = \text{Cov}(r, \hat{s})/\text{Var}(\hat{s})$, iar $\text{Cov}(r, \hat{s}) = \beta \, \text{Cov}(s^*, \hat{s}) + \text{Cov}(\varepsilon, \hat{s})$; ultimul termen este 0 din ipoteza de eroare nediferențială
- (b) Ambele variabile sînt binare: $\text{Cov}(s^*, \hat{s}) = P(s^* = 1, \hat{s} = 1) - \pi p = \pi(1 - \alpha_1) - \pi p$
- Formula probabilității totale: $p = \pi(1 - \alpha_1) + (1 - \pi)\alpha_0$; înlocuind, $\text{Cov}(s^*, \hat{s}) = \pi(1 - \pi)(1 - \alpha_0 - \alpha_1)$
- $\hat{s}$ este Bernoulli, deci $\text{Var}(\hat{s}) = p(1 - p)$; raportul celor două este λ

A1 [Rezolvat]: rezolvare (2/3)



MFM_ch15_sem_partA

A1 [Rezolvat]: rezolvare (3/3)

- (c) $\pi = 347/2\,388 = 0{,}1453$, $p = 535/2\,388 = 0{,}2240$, $\alpha_{\mathbf{1}} = 1 - 264/347 = 0{,}2392$, $\alpha_{\mathbf{0}} = 271/2\,041 = 0{,}1328$
- $\lambda = \dfrac{0{,}1242 \times 0{,}6280}{0{,}2240 \times (1 - 0{,}2240)} = 0{,}449$, deci $\beta = 2{,}23\,\hat{b}$
- (d) $\mathrm{Cov}(u, s^*) = 0{,}0780 - 0{,}1242 = -0{,}0462 < 0$: eroarea este corelată negativ cu adevărul, deci nu este clasică Aigner (1973)
- Graficul: 2 000 de eșantioane simulate cu aceste rate și $\beta = 1$; panta pe $\hat{s}$ are media 0,450 (abaterea standard 0,054), cum prezice λ
- Interpretare: o pantă pe variabila binară FinBERT păstrează doar circa 45% din efectul adevărat

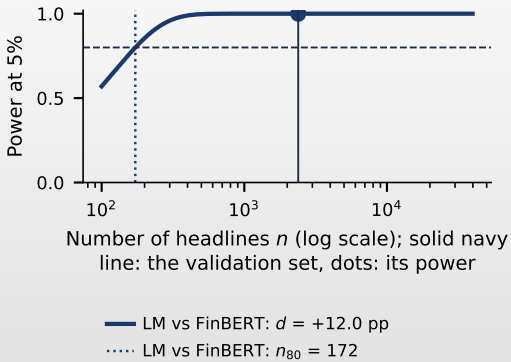
MFM_ch15_sem_partA

A2 [Rezolvat]: diferența de acuratețe pe perechi: variantă, test exact, putere — enunț

- **Întrebări**
  - cît de precisă este diferența de acuratețe dintre doi clasificatori evaluați pe aceleași titluri?
  - de cîte titluri are nevoie?
- **Date de intrare**
  - Setul de validare Twitter, $n = 2\,388$ de titluri independente; A = dicționarul LM, B = FinBERT
  - Numere discordante: $n_{01} = 539$ (A greșește, B are dreptate), $n_{10} = 252$ (A are dreptate, B greșește)
  - $D_i = 1$ dacă A greșește și B are dreptate, -1 dacă A are dreptate și B greșește, 0 altfel; $d = ED_i = \text{acc}_B - \text{acc}_A$, $q = P(D_i \neq 0)$
- **Cerințe**
  - (a) Arătați că $\hat{d} = (n_{01} - n_{10})/n$ și deduceți $\text{Var}(\hat{d}) = (q - d^2)/n$.
  - (b) Calculați $\hat{d}$, SE, CI 95% și $z = \hat{d}/\text{SE}$.
  - (c) Aplicați testul McNemar exact la 5% și formulați decizia.
  - (d) Dacă d adevărat este $\hat{d}$, calculați puterea și n_{80}, valoarea lui n pentru puterea 80%, cu formulele de pe slide-ul Testul McNemar.
- **Raportați**: $\hat{d}$, SE, CI, z, p exact, decizia, puterea și n_{80}
- **Interpretare**: ce titluri determină varianța lui $\hat{d}$?

înapoi Lista exercițiilor

A2 [Rezolvat]: rezolvare



- (a) $\hat{d} = \frac{1}{n}\sum_i D_i = (n_{01} - n_{10})/n$; $D_i^2 = 1$ exact cînd unul are dreptate și celălalt greșește, deci $ED_i^2 = q$, $\mathrm{Var}(D_i) = q - d^2$ și $\mathrm{Var}(\hat{d}) = (q - d^2)/n$
- (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) Sub H_0 ne așteptăm la 395,5 de fiecare parte; p exact $<0{,}001$: respingem acuratețea egală
- (d) Puterea $\Phi(+8{,}47) \approx 100\%$; $n_{80} = 7{,}84 \times (0{,}3312 - 0{,}01444)/0{,}01444 \approx 172$ de titluri
- Interpretare: doar titlurile discordante (q) intră în varianță; o diferență de 12 puncte se detectează cu circa 170 de titluri

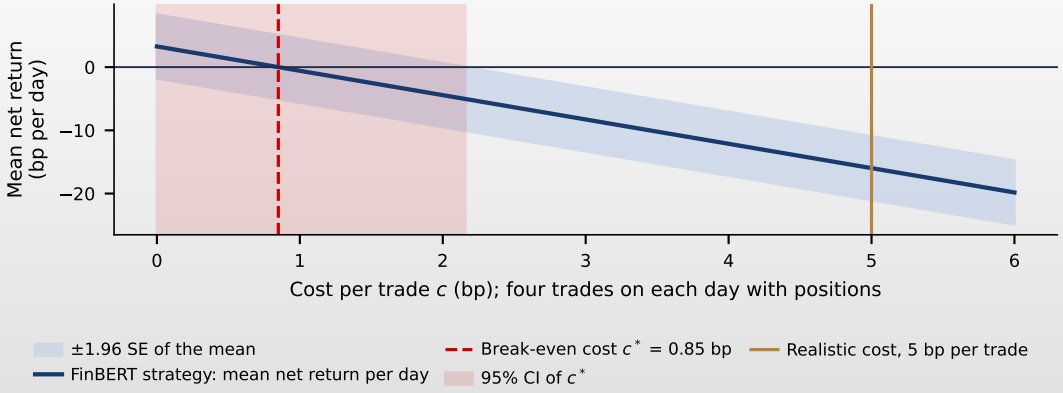
MFM_ch15_sem_partA

A3 [Rezolvat]: costul de echilibru cu eroare standard — enunț

- **Întrebări**
  - de la ce cost pe tranzacție strategia FinBERT nu mai este profitabilă?
  - cît de precis cunoaștem acest cost?
- **Date de intrare**
  - Strategia de sentiment (slide-ul Strategia de sentiment): la închiderea din $d+1$ deschidem o poziție de semn $\text{sign}(s_{i,d})$ în fiecare acțiune cu știri și scor nenul în ziua d, acoperită cu SPY (*hedging*) și închisă după o zi; patru tranzacții în fiecare zi cu poziții (intrarea și ieșirea, pe acțiuni și pe SPY); 3 568 de zile de tranzacționare
  - Semnalul FinBERT: randamentul brut mediu pe toate zilele $\hat{\mu} = 3{,}27$ bp, $t_{NW} = 1{,}27$ ($L = 8$ lag-uri), ponderea zilelor cu poziții $a = 0{,}962$, raportul Sharpe brut $\sqrt{A}\,\hat{\mu}/\sigma = 0{,}36$, cu σ abaterea standard a randamentelor zilnice și $A = 252$ de zile de tranzacționare pe an
- **Cerințe**
  - (a) Păstrînd fixe abaterea standard a randamentului brut și ponderea zilelor active, scrieți randamentul net zilnic mediu și raportul Sharpe net aproximativ ca funcții de costul c pe tranzacție.
  - (b) Deduceți costul de echilibru c^* la care randamentul net mediu este zero.
  - (c) Calculați $\text{SE}(\hat{\mu}) = \hat{\mu}/t$, apoi $\text{SE}(c^*)$ prin metoda delta și CI 95% pentru c^*.
  - (d) Obțineți σ din raportul Sharpe brut și calculați raportul Sharpe net la $c = 5$ bp.
  - (e) Caz nou: $\hat{\mu} = 2$ bp, $a = 0{,}9$, $\sigma = 120$ bp, $c = 0{,}4$ bp; calculați c^* și raportul Sharpe net anualizat.
- **Raportați**: c^*, SE și CI, raportul Sharpe net; pentru (e), c^* și raportul Sharpe net
- **Interpretare**: este un cost realist de 5 bp pe tranzacție în interiorul CI pentru c^*?

◂ Înapoi ▸ Lista exercițiilor

A3 [Rezolvat]: rezolvare (1/2)



MFM_ch15_sem_partA

A3 [Rezolvat]: rezolvare (2/2)

- (a) Într-o fracțiune a din zile plătim $4c$: $\mu_{\text{net}} = \mu - 4ac$, $SR_{\text{net}} = \sqrt{A}\,(\mu - 4ac)/\sigma$, folosind σ brută ca aproximație
- (b) $c^* = \mu/(4a) = 3{,}27/3{,}85 = 0{,}85$ bp pe tranzacție
- (c) $\mathrm{SE}(\hat{\mu}) = 3{,}27/1{,}27 = 2{,}57$ bp; cu a fixat de calendarul știrilor, $\mathrm{SE}(c^*) = 2{,}57/3{,}85 = 0{,}67$ bp, CI $[-0{,}46;\ 2{,}16]$ bp
- (d) $\sigma = \sqrt{252} \times 3{,}27/0{,}362 = 143$ bp, cu raportul Sharpe nerotunjit 0,362
  - $SR_{\text{net}} = \sqrt{252}\,(3{,}27 - 3{,}85 \times 5)/143 = -1{,}77$
- (e) $c^* = 2/(4 \times 0{,}9) = 0{,}56$ bp; $SR_{\text{net}} = \sqrt{252}\,(2 - 3{,}6 \times 0{,}4)/120 = +0{,}07$: abia profitabilă
- Interpretare: 5 bp este mult peste CI pentru c^*; cu pozițiile schimbate zilnic, strategia nu rămîne profitabilă la costuri realiste

MFM_ch15_sem_partA

B1 [Rezolvat]: FinBERT față de LM — enunț

- **Întrebare**: este FinBERT, adaptat prin fine-tuning, mai precis decît dicționarul Loughran–McDonald pe titluri pe care nu a fost antrenat?
- **Date**
  - Setul de validare Twitter din A1: $N = 2\,388$ de titluri, eticheta umană = adevărul
  - A = dicționarul LM (eticheta = semnul tonului), B = FinBERT (eticheta = probabilitatea cea mai mare); funcția `b_pairs` din notebook
- **Cerințe**
  (a) Calculați fiecare acuratețe cu CI bootstrap percentil 95%: reeșantionați titlurile de 2 000 de ori (seed 0).
  (b) Calculați diferența B − A cu CI bootstrap pe perechi: aceleași titluri reeșantionate pentru ambii clasificatori.
  (c) Numărați n_{01}, n_{10}; aplicați testul McNemar exact și $(n_{01} - n_{10})^2/(n_{01} + n_{10})$ față de χ^2_1 (3,84).
- **Raportați**: un tabel cu cele două acurateți, CI și diferența; n_{01}, n_{10}, ambele valori p și decizia
- **Interpretare**
  - de ce este testul pe perechi mai puternic decît compararea celor două intervale?
  - contează diferența pentru tranzacționare?

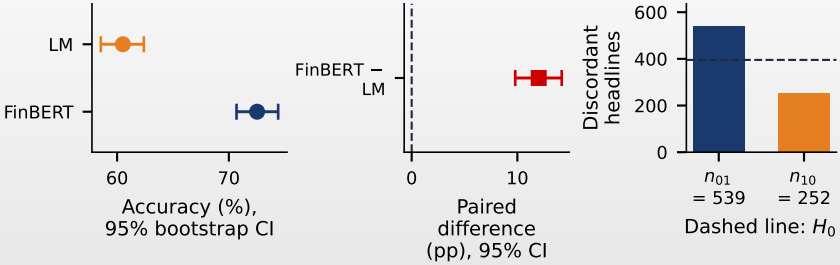
Înapoi Lista exercițiilor

B1 [Rezolvat]: rezolvare (1/2)

	Acurateţe (%)	CI 95%
LM	60,5	[58,5; 62,4]
FinBERT	72,5	[70,7; 74,4]
Diferența (pp)	+12,0	[+9,8; +14,2]

- Pasul 1: etichetele ambilor clasificatori pe cele 2 388 de titluri; pasul 2: 2 000 de reeșantionări bootstrap ale indicilor titlurilor; pasul 3: ambele acurateți și diferența lor în fiecare reeșantionare
- (a)–(b) Cele două intervale nu se suprapun, iar CI pe perechi al diferenței exclude 0; coincide practic cu CI analitic din A2

B1 [Rezolvat]: rezolvare (2/2)



- (c) $n_{01} = 539$, $n_{10} = 252$, 395,5 așteptate sub H_0; p exact $<0{,}001$; $\chi^2 = 104{,}1 > 3{,}84$: respingem acuratețea egală
- Împerecherea elimină variația dificultății titlurilor: titlurile pe care ambii le clasifică la fel nu aduc informație
- Interpretare: FinBERT este clar mai bun decît dicționarul în afara datelor lui de antrenare; un cîștig de 12 puncte în etichetare nu implică încă nimic despre randamente (B2)

MFM_ch15_sem_partB

B2 [Rezolvat]: reacție față de predicție — enunț

- **Întrebări**
  - explică scorul zilnic al titlurilor randamentul din ziua știrii?
  - prognozează el primul randament care mai poate fi tranzacționat?
  - cît de largi sînt intervalele corecte?
- **Date**
  - Panelul FNSPID: 18 282 de zile-acțiune cu știri, 16 acțiuni americane, 27.10.2009–18.12.2023; $z_{i,d}$: scorul net zilnic, standardizat pe panel; r^e: randamentul în exces față de SPY (%), din prețurile ajustate zilnice ale cursului
  - Scoruri: FinBERT, Qwen2.5-7B, LM
- **Cerințe**
  (a) Estimați prin OLS $r^e_{i,d+h} = a + b\, z_{i,d} + \varepsilon$ pentru cele trei scoruri și $h = 0, 1, 2$; raportați b în bp pentru o abatere standard a scorului.
  (b) Calculați trei SE: OLS, grupată pe zile ($G = 3\,434$), grupată pe acțiuni ($G = 16$), cu factorul de eșantion mic c (slide-ul Regresia de panel).
  (c) Comparați fiecare t cu 1,96, iar t grupat pe acțiuni cu valoarea critică $t_{\mathbf{15}}$, 2,13.
- **Raportați**: un grafic al pantelor pentru $h = 0$ cu cele trei intervale și un tabel al regresiilor
- **Interpretare**
  - de ce diferă $h = 0$ și $h = 2$?
  - putem avea încredere într-o SE grupată cu 16 grupuri?

înapoi Lista exercițiilor

B2 [Rezolvat]: rezolvare (1/2)



- Pantele din aceeași zi cu intervalele de 95% ale celor trei erori standard: intervalul grupat pe acțiuni este cel mai larg
- Gruparea pe zile este apropiată de OLS: cu doar 5,3 acțiuni pe zi, zgomotul comun este mic
- Gruparea pe acțiuni înmulțește SE FinBERT la $h = 0$ cu 3,0: panta diferă între acțiuni, iar erorile fiecărei acțiuni sînt corelate în timp

MFM_ch15_sem_partB

B2 [Rezolvat]: rezolvare (2/2)

Scor	h	b (bp)	SE OLS	SE zi	SE acțiune	t zi	t acțiune
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

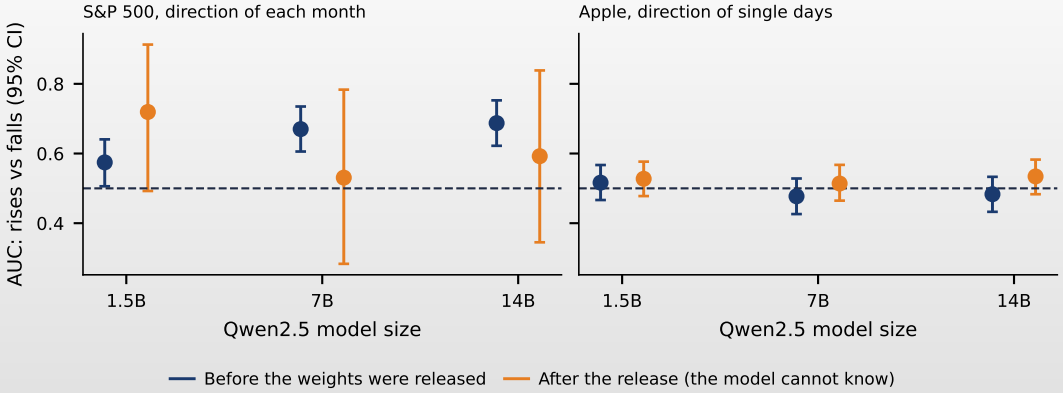
- $h = 0$: mare cu orice SE (t pe acțiuni de cel puțin 4,7): titlurile descriu mișcarea zilei, adesea o relatează explicit („shares jump")
- $h = 2$: FinBERT $-0{,}5$ bp ($t = -0{,}25$ pe zile); Qwen $t = +2{,}41$ pe acțiuni, dar 16 grupuri sînt puține: verificați cu wild cluster bootstrap Cameron, Gelbach & Miller (2008) și Holm (B9)
- Interpretare: sentimentul explică ziua știrii, nu zilele tranzacționabile; cu 16 acțiuni, inferența cere metode pentru puține grupuri

B3 [Rezolvat]: memoria modelului despre piață — enunț

- **Întrebare**: întrebat fără nicio dată, poate Qwen2.5 să distingă lunile în care S&P 500 a crescut de cele în care a scăzut?
- **Date**
  - Închideri lunare S&P 500 (`GSPC.INDX`), ianuarie 2000 – august 2026; o lună este „în creștere" dacă randamentul ei este pozitiv
  - Qwen2.5 1.5B, 7B, 14B (ponderi publicate pe 19 septembrie 2024) întrebate „Did the S&P 500 index rise or fall during ⟨lună an⟩?"; scorul $P(\text{rise})$
  - Înainte: 296 de luni pînă în august 2024 (185 în creștere); după: 23 de luni din octombrie 2024 (13 în creștere); septembrie 2024, luna publicării, este exclusă
- **Cerințe**
  (a) Calculați AUC $= U/(n_1 n_0)$ (slide-ul AUC) pe model și perioadă, cu CI bootstrap din 2 000 de extrageri pe luni.
  (b) Test prin permutare pentru 14B înainte de publicare: amestecați etichetele creștere/scădere de 2 000 de ori.
  (c) Pentru 14B, testați $\text{AUC}_{\textbf{înainte}} = \text{AUC}_{\textbf{după}}$ cu variațele DeLong DeLong et al. (1988) și găsiți cea mai mică scădere detectabilă cu puterea 80%, cu $\text{SE}_0 = \sqrt{(n_1 + n_0 + 1)/(12\, n_1 n_0)}$, eroarea standard a AUC după publicare sub H_0: AUC $= 0{,}5$ (n_1, n_0: lunile în creștere și în scădere după publicare).
- **Raportați**: tabelul AUC cu CI, p prin permutare, z și p DeLong, scăderea detectabilă
- **Interpretare**: ce înseamnă un AUC peste 0,5 cînd modelul nu vede date?

înapoi Lista exercițiilor

B3 [Rezolvat]: rezolvare (1/3)



MFM_ch15_lookahead

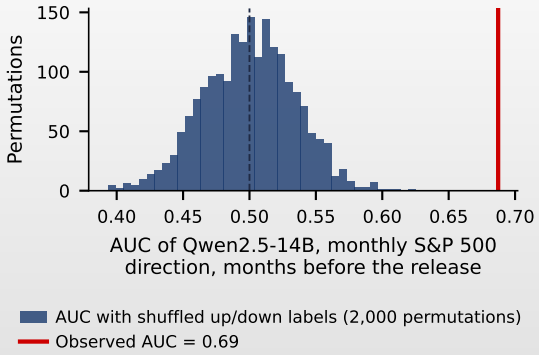
B3 [Rezolvat]: rezolvare (2/3)

Qwen2.5	AUC înainte	CI 95%	AUC după	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 și 14B ordonează lunile de creștere peste cele de scădere mult peste nivelul întîmplător înainte de publicare; 1.5B doar marginal; după publicare toate intervalele conțin 0,5
- Panoul din dreapta al graficului: aceeași întrebare pentru direcția a 500 de zile Apple înainte și 493 după publicare
- Pentru zile, AUC este între 0,48 și 0,53 și toate intervalele conțin 0,5: nu detectăm discriminare peste nivelul aleator în acest experiment
- Discriminarea lunilor este compatibilă cu memorarea, dar singură nu o separă de regularitățile calendaristice sau de regim

MFM_ch15_lookahead

B3 [Rezolvat]: rezolvare (3/3)



- (b) p prin permutare pentru 14B înainte de publicare $<0{,}001$: niciun AUC amestecat nu atinge valoarea observată (graficul)
- (c) DeLong, 14B: SE înainte 0,032, după 0,126; $z = 0{,}73$, $p = 0{,}47$: nu respingem egalitatea
  - $\mathrm{SE}_{\mathbf{0}} = \sqrt{24/(12 \times 13 \times 10)} = 0{,}124$ ($n_{\mathbf{1}} = 13$ luni în creștere, $n_{\mathbf{0}} = 10$ în scădere)
  - Scăderea detectabilă $2{,}80\sqrt{\mathrm{SE}^2_{\text{înainte}} + \mathrm{SE}^2_{\mathbf{0}}} = 0{,}36$
  - Detectarea revenirii de la 0,69 la 0,5 cu puterea 80% ar cere circa 101 luni după publicare; avem 23
- Interpretare: fără date, discriminarea poate proveni doar din ce a învățat modelul la antrenare (memorie sau regularități)
  - Un backtest pe știri anterioare publicării confundă prognoza cu memorarea Glasserman & Lin (2023)

MFM_ch15_sem_partB

C1: reproducerea analizei din scorurile salvate

- Folosiți `ch15_twitter_scores.csv` pentru comparația cu etichete și `ch15_news_daily.csv` pentru scoruri și randamente zilnice pe acțiune, din folderul Quantlet al capitolului. Exercițiul nu cere o nouă rulare a modelului.
- În codul asistentului (slide-ul următor), `qwen_label` și `qwen_score` țin locul scorurilor salvate. Separați trei erori principale ale afirmației de limitele suplimentare ale inferenței; numărați o singură dată aceeași eroare care se propagă pe mai multe rînduri.
- Gruparea numai pe zile este o alegere de inferență de evaluat cu B2, nu o garanție a incertitudinii corecte cînd efectele pe acțiune persistă. Raportați limita separat de greșelile privind etichetele, calendarul și disponibilitatea informației.

C1 [Propus]: critica unui răspuns AI — răspunsul

- **Cererea studentului**: „Am titlurile FNSPID pentru 16 acțiuni americane, 2009–2023, și randamentele zilnice în exces față de SPY. Verifică dacă Qwen2.5-7B citește sentimentul, apoi testează dacă sentimentul titlurilor prognozează randamentele în afara eșantionului.”
- **Codul asistentului AI** (`qwen_label`, `qwen_score`: evaluatorul zero-shot de pe slide-ul Trei moduri de a citi sentimentul)

```
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]})
```

◂ Înapoi ▸ Lista exercițiilor

C1 [Propus]: critica unui răspuns AI — enunț

- **Model**: B1, B2 și B3 [Rezolvat]
- **Concluzia asistentului AI**
  - „Acuratețe 74,2%, de peste două ori cei 33% ai unei ghiciri la întîmplare între trei clase: Qwen2.5-7B citește foarte bine sentimentul financiar."
  - „Panta $+41{,}4$ bp pentru o abatere standard a scorului, $t = +19{,}07$ (grupat pe zile): titlurile zilei prognozează randamentul în exces al acțiunii."
  - „Qwen2.5 nu a fost antrenat pe aceste etichete sau serii de randamente, deci backtest-ul 2009–2023 este complet în afara eșantionului."
- **Întrebare**: codul rulează și cifrele sînt corecte, dar răspunsul conține **trei** erori tipice pentru backtesting-ul sentimentului; care sînt?
- **Cerințe**
  - (a) Identificați fiecare eroare: numiți rîndul de cod sau propoziția și explicați de ce este greșită.
  - (b) Precizați cum ați detecta-o fără să cunoașteți răspunsul: numărul de observații pe clase, o verificare a orelor, o dată de publicare, un test de memorie.
  - (c) Scrieți codul sau afirmația corectată, cu cifrele pe care le produce.
  - (d) Notați fiecare eroare în `AI_ERRORS.md`, cu cîmpurile: ce a răspuns AI-ul, ce era greșit, cum ați identificat eroarea, varianta corectă.
- **Raportați**: cele trei erori, cu detectarea, corectura și cifrele fiecăreia, și intrările din `AI_ERRORS.md`

B4 [Propus]: un LLM cu 7 miliarde de parametri față de FinBERT — enunț

- **Model**: B1 [Rezolvat]
- **Întrebare**: depășește un LLM deschis general, folosit zero-shot, modelul FinBERT adaptat prin fine-tuning, pe titluri pe care niciunul nu a fost antrenat?
- **Date**
  - Setul de validare Twitter din A1: $N = 2\,388$ de titluri, eticheta umană = adevărul
  - A = FinBERT, B = Qwen2.5-7B zero-shot cu promptul P1; niciunul nu a folosit etichete Twitter
- **Cerințe**
  (a) Calculați fiecare acuratețe cu CI bootstrap percentil 95%: reeșantionați titlurile de 2 000 de ori (seed 0).
  (b) Calculați diferența B − A cu CI bootstrap pe perechi: aceleași titluri reeșantionate pentru ambii clasificatori.
  (c) Numărați n_{01}, n_{10}; aplicați testul McNemar exact și $(n_{01} - n_{10})^2/(n_{01} + n_{10})$ față de χ^2_1 (3,84).
- **Raportați**: același tabel, aceleași teste și același grafic ca în B1
- **Interpretare**
  - se suprapun cele două intervale?
  - este întrebarea astfel tranșată?
  - ce ar aduce totuși un model mult mai mare și mai lent?

◂ Înapoi ▸ Lista exercițiilor

Pasul pentru proiectul de echipă

- Alegeți o sursă de texte nediscutată la seminar (anunțuri ale BVB (Bursa de Valori București), știri financiare românești, comunicate ale băncilor centrale, postări despre criptoactive)
- Etichetați cel puțin 200 de texte cu doi adnotatori; raportați kappa Cohen (acordul observat corectat pentru acordul întîmplător). Rezervați cel puțin 30% dintre texte ca eșantion de test înainte de a evalua modelele și raportați acuratețea numai pe acel eșantion
- Predați o pagină: sursa, numărul etichetelor, acordul, acuratețea pe eșantionul rezervat și intervalul ei de încredere.
- Consemnați folosirea AI în `AI_USE.md` și erorile depistate în `AI_ERRORS.md`; pachetele de extindere sînt opționale; C2 și C3 sînt puncte de plecare pentru proiect
- Criterii de evaluare
  - etichete documentate: doi adnotatori și kappa Cohen
  - acuratețea măsurată numai pe eșantionul rezervat, cu interval de încredere
  - rezultatele de tranzacționare raportate după costuri
  - o interpretare economică pentru fiecare rezultat

Pachetele de extindere: condițiile fiecăruia

Pachet	Exerciții	Cunoștințe necesare
1. Eroarea de măsurare	A4, A5	A1
2. Compararea clasificatorilor	A6, B5 [Rezolvat], B6, B7	A2, B1
3. Probabilitățile LLM	A7 [Rezolvat]	slide-ul De la logiți la probabilități, slide-ul Calibrarea
4. Inferență validă	B8 [Rezolvat], B9, B10	B2, B3
5. Semnale și costuri	A8, B11, C2	A3, B2
6. Laborator de replicare	C3	C2, Lopez-Lira & Tang (2026)

- Fiecare exercițiu are secțiunea lui în notebook; rezolvați-le în ordine în cadrul unui pachet

A4 [Propus]: atenuarea pentru dicționarul LM — enunț

- **Model**: A1 [Rezolvat]
- **Întrebare**: cît micșorează panta erorile etichetei binare „negativ" a dicționarului LM? Pantele pe scoruri continue din B2 (15,5 bp pentru LM, 39,2 bp pentru FinBERT) răspund la altă întrebare.
- **Date**
  - Setul de validare Twitter din A1: $N = 2\,388$ de titluri, eticheta umană = adevărul
  - Același model și aceeași ipoteză ca în A1; acum $\hat{s} = 1$ cînd tonul LM al titlului este negativ
- **Cerințe**
  (a) Calculați π, p, α_0 și α_1 din matricea LM.
  (b) Calculați λ și $1/\lambda$ cu formula din A1.
  (c) Calculați $\mathrm{Cov}(u, s^*)$: este eroarea LM clasică?
  (d) Comparați λ cu cel al FinBERT (A1) și calculați raportul celor doi factori.
- **Raportați**: cele patru rate, λ, $1/\lambda$, raportul
- **Interpretare**: care clasificator atenuează mai mult panta?

Matricea de confuzie LM, Twitter

Adevărat \ prezis	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

înapoi Lista exercițiilor

A5 [Propus]: κ al lui Fleiss și un pas Dawid–Skene — enunț

- **Model**: formula lui κ de pe slide-ul Holm, PPI și κ
- **Întrebări**
  - cît de mult sînt de acord evaluatorii peste nivelul întîmplător?
  - care este eticheta adevărată cea mai probabilă a unei propoziții disputate?
- **Date de intrare**
  - Voturi ilustrative (alese pentru exercițiu): 6 propoziții, cîte $m = 5$ evaluatori
  - $P_i = \sum_j n_{ij}(n_{ij}-1)/[m(m-1)]$, $\bar{P} = \frac{1}{N}\sum_i P_i$, $p_j =$ ponderea voturilor pentru j, $\bar{P}_e = \sum_j p_j^2$; κ Fleiss (1971)
- **Cerințe**
  - (a) Calculați fiecare P_i și $\bar{P}$.
  - (b) Calculați p_j, $\bar{P}_e$ și κ.
  - (c) Pasul E din Dawid & Skene (1979): fiecare evaluator dă clasa adevărată cu probabilitatea 0,8 și fiecare altă clasă cu 0,1, distribuție a priori uniformă; calculați $P(\text{adevărat} = j \mid \text{voturi}) \propto \prod_l \pi_{jl}^{n_{il}}$ pentru propozițiile 4 și 5.
  - (d) Comparați probabilitățile a posteriori cu ponderile voturilor și descrieți pasul M.
- **Raportați**: κ și cele două distribuții a posteriori
- **Interpretare**: cum ați trata propozițiile disputate într-un set de test?

Propoziția	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

◂ Înapoi ▸ Lista exercițiilor

A6 [Propus]: „nicio diferență” și puterea testului — enunț

- **Model**: A2 [Rezolvat]
- **Întrebare**: FinBERT și Qwen2.5-7B par la egalitate: este asta o dovadă de acuratețe egală sau a unui test slab?
- **Date de intrare**
  - Setul de validare Twitter, $n = 2\,388$; A = FinBERT, B = Qwen2.5-7B (promptul P1); $n_{01} = 395$, $n_{10} = 355$
- **Cerințe**
  (a) Calculați $\hat{d}$, $\hat{q}$, SE, CI 95% și z cu formulele din A2.
  (b) Aplicați testul McNemar exact la 5%.
  (c) Calculați puterea la $n = 2\,388$ dacă d adevărat este $\hat{d}$ și n_{80}.
  (d) Folosiți CI pentru a preciza ce diferențe sînt excluse de date.
- **Raportați**: cifrele din (a)–(d)
- **Interpretare**: puteți concluziona că cele două modele sînt la fel de precise?

◂ Înapoi ▸ Lista exercițiilor

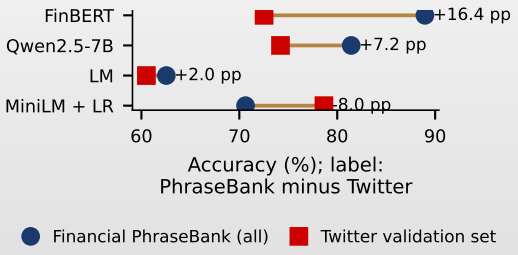
B5 [Rezolvat]: suprapunerea cu datele de antrenare — enunț

- **Întrebare**: cît din acuratețea FinBERT pe Financial PhraseBank provine din faptul că a fost antrenat pe acest set?
- **Date**
  - Financial PhraseBank Malo et al. (2014): 4 846 de propoziții (2 264 cu acord deplin); FinBERT a fost adaptat prin fine-tuning pe o parte a lor (3 101 din 4 845 în Araci (2019); versiunea folosită aici are 4 846)
  - Setul de validare Twitter, 2 388 de titluri nevăzute de FinBERT; clasificatori: FinBERT, Qwen2.5-7B, LM, vectorii MiniLM Wang et al. (2020) + regresie logistică, antrenată pe setul de antrenare Twitter
- **Cerințe**
  (a) Calculați acuratețea (CI bootstrap din 2 000 de extrageri) și macro-F1 pe PhraseBank cu acord deplin, pe tot PhraseBank și pe Twitter.
  (b) Calculați scăderea de la tot PhraseBank la Twitter pentru fiecare clasificator.
  (c) Calculați diferența diferențelor FinBERT minus Qwen2.5-7B cu CI bootstrap (fiecare set reeșantionat independent).
- **Raportați**: un tabel cu acurateți și scăderi, diferența diferențelor cu CI
- **Interpretare**: cum separați „date mai dificile" de „memoria setului de antrenare"?

◂ Înapoi ▸ Lista exercițiilor

B5 [Rezolvat]: rezolvare

Acuratețe (%)	PhraseBank, acord 100%	PhraseBank, tot	Twitter	Scădere (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 + reg. log.	81,4	70,6	78,6	−8,0	69,9



- (c) $16{,}4 - 7{,}2 = 9{,}2$ pp din acuratețile nerotunjite, CI bootstrap [6,7; 11,9]: exclude 0
- Rezultatul: o diferență descriptivă între scăderile de acuratețe
  - Qwen2.5-7B nu a fost adaptat prin fine-tuning pe PhraseBank, dar textele pot apărea în datele lui de preantrenare
  - Atribuirea celor 9,2 puncte suprapunerii cere un comparator necontaminat și nicio altă schimbare diferențială (distribuție, dificultate, compoziția claselor)
- Interpretare: MiniLM variază în sens invers pentru că este antrenat pe Twitter: fiecare model este cel mai bun pe datele pe care a fost antrenat

MFM_ch15_sem_partB

B6 [Propus]: numărul necesar de etichete — enunț

- **Model**: B5 [Rezolvat]
- **Întrebare**: de cîte titluri etichetate are nevoie un model supervizat simplu pentru a-l depăși pe FinBERT?
- **Date**
  - Twitter Financial News: set de antrenare cu 9 543 de titluri, set de validare cu 2 388; reperul FinBERT, 72,5%, fără etichete Twitter
  - Regresie logistică ($C = 4$) pe vectorii MiniLM cu 384 de dimensiuni; regresie logistică ($C = 8$) pe TF-IDF (Term Frequency–Inverse Document Frequency): cuvinte și perechi ponderate prin frecvență și raritate
  - Implementare: un singur `default_rng(15)`, subseturi simple fără stratificare, în ordinea crescătoare a lui n; MiniLM `all-MiniLM-L6-v2` fix; `TfidfVectorizer(ngram_range=(1,2), sublinear_tf=True)`; `max_iter=3000`
- **Cerințe**
  (a) Pentru $n \in \{100, 200, 400, \ldots, 6\,400, 9\,543\}$ extrageți 10 subseturi aleatoare de antrenare (setul complet o singură dată), estimați ambele modele și mediați acuratețea pe validare.
  (b) Găsiți cel mai mic n la care fiecare model îl depășește în medie pe FinBERT.
  (c) Raportați abaterea standard a acurateței vectorilor pe cele 10 repetiții la $n = 100$.
- **Raportați**: curba de învățare și tabelul ei
- **Interpretare**: de ce vectorii pornesc mai sus, dar cresc mai încet?

Înapoi Lista exercițiilor

B7 [Propus]: sensibilitatea la prompt — enunț

- **Model**: B1 [Rezolvat]
- **Întrebare**: se schimbă răspunsul aceluiași LLM cînd aceeași întrebare este formulată altfel?
- **Date**
  - Setul de validare Twitter din A1: $N = 2\,388$ de titluri, eticheta umană = adevărul
  - Qwen2.5-7B și 14B cu trei prompturi: P1 „classify the sentiment ... for investors", P2 „good, bad or neutral for the stock price", P3 „what is the sentiment of this text"
- **Cerințe**
  (a) Calculați acuratețea cu CI bootstrap din 2 000 de extrageri și ponderea răspunsurilor „neutral" pentru fiecare prompt și model.
  (b) Aplicați testul McNemar exact pentru P1 față de P3, pentru fiecare model.
  (c) Calculați ponderea titlurilor cu aceeași etichetă sub toate cele trei prompturi.
- **Raportați**: un tabel și un grafic pentru cele trei prompturi
- **Interpretare**: dacă ați alege cel mai bun prompt pe acest set, ce ați face de fapt?

Înapoi Lista exercițiilor

A7 [Rezolvat]: de la probabilitățile etichetelor la un scor și temperatura optimă — enunț

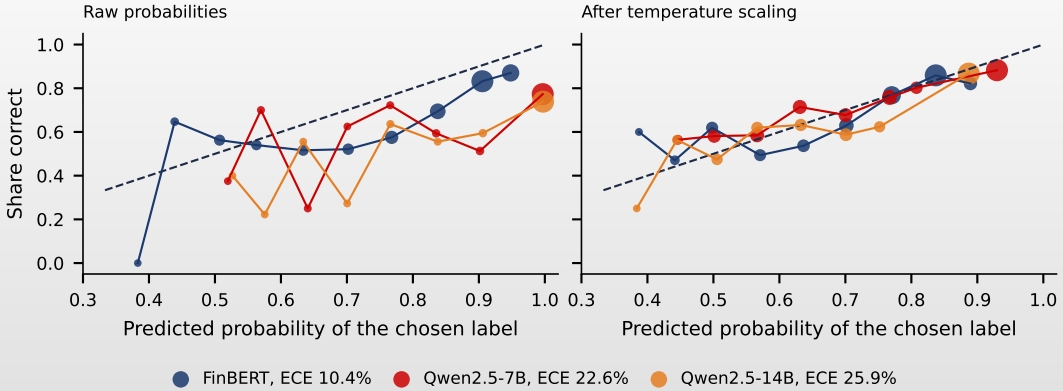
- **Întrebări**
  - cum schimbă temperatura T probabilitățile unui LLM, scorul și calibrarea lui?
  - ce T este optim pentru date?
- **Date de intrare**
  - Un titlu Twitter etichetat **neutru**; Qwen2.5-7B (P1) dă $p_{\mathbf{neg}} = 0{,}608$, $p_{\mathbf{neu}} = 0{,}392$, $p_{\mathbf{pos}} = 0{,}000$ (rotunjit la 0 în fișier, înlocuit numeric cu $1{,}0 \times 10^{-\mathbf{12}}$)
  - O jumătate aleatoare a setului de validare ($n = 1\,194$, seed 15): logiții $z_{ik} = \ln p_{ik}$, rescalați $p_{ik}(T)$ ca pe slide-ul De la logiți la probabilități, pierderea NLL(T) Guo et al. (2017) (slide-ul Calibrarea)
- **Cerințe**
  (a) Dați eticheta, scorul net s și diferența de logit $\ln(p_{\mathbf{neu}}/p_{\mathbf{neg}})$ ale titlului, apoi probabilitățile și s la $T = 2$.
  (b) Explicați de ce etichetele și acuratețea nu depind de T, dar scorul Brier da.
  (c) Deduceți NLL$'(T)$, evaluați condiția de ordinul întîi la $T = 1$ și rezolvați pentru T^* (metoda Brent pe $[0{,}5; 50]$).
  (d) Pe cealaltă jumătate, calculați ECE și Brier înainte și după scalare.
- **Raportați**: eticheta, s, $s(T)$, T^*, ECE și Brier înainte și după
- **Interpretare**
  - este LLM-ul prea încrezător sau prea puțin încrezător?
  - ce se întîmplă cu o regulă cu prag pe s?

◂ Înapoi ▸ Lista exercițiilor

A7 [Rezolvat]: rezolvare (1/3)

- (a) Cea mai mare probabilitate p_{neg}: eticheta negativă, greșită; $s = 0{,}000 - 0{,}608 = -0{,}608$; diferența de logit $\ln(0{,}392/0{,}608) = -0{,}439$ ($-0{,}437$ din probabilitățile nerotunjite)
- La $T = 2$, $p_k(T) \propto \sqrt{p_k}$: $\sqrt{p} = (0{,}780,\ 0{,}626,\ 1{,}0 \times 10^{-6})$, suma 1,406, $p(T) = (0{,}554,\ 0{,}446,\ 0{,}000)$, $s(T) = -0{,}554$: aceeași etichetă, scorul se apropie de 0
- (b) $z \mapsto z/T$ este crescătoare pentru $T > 0$: ordinea logiților, deci argmax și acuratețea, nu se schimbă; probabilitățile, deci Brier și NLL, se schimbă

A7 [Rezolvat]: rezolvare (2/3)



MFM_ch15_valid_inference

A7 [Rezolvat]: rezolvare (3/3)

- (c) $\ln p_{iy}(T) = z_{iy}/T - \ln \sum_k e^{z_{ik}/T}$ dă $\mathrm{NLL}'(T) = \frac{1}{T^2} \sum_i \left(z_{iy_i} - \sum_k p_{ik}(T) z_{ik} \right)$
- La $T = 1$ media parantezei este $-2{,}47 < 0$, deci NLL scade cînd T crește; rădăcina este $T^* = 8{,}43$
- Probabilitățile din fișier sînt rotunjite la 6 zecimale (2 539 de zerouri, în 1 750 de titluri): cu pragul 10^{-6} în locul lui 10^{-12}, $T^* = 5{,}58$; supraîncrederea rămîne
- (d) ECE scade de la 22,6% la 4,3%, Brier de la 0,468 la 0,366 (graficul: fiabilitatea brută și rescalată)
- Interpretare: probabilitățile brute ale LLM sînt mult prea încrezătoare; un prag pe s selectează alte titluri după rescalare

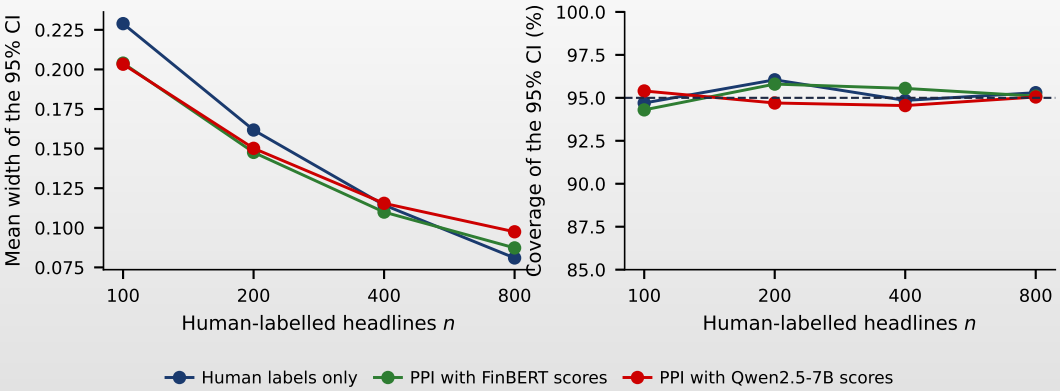
MFM_ch15_valid_inference

B8 [Rezolvat]: inferența susținută de predicții cu puține etichete umane — enunț

- **Întrebare**: pot cîteva sute de etichete umane plus un model distorsionat să dea un interval valid și mai îngust pentru tonul mediu al titlurilor?
- **Date**
  - Ținta θ = tonul net uman mediu $(-1, 0, 1)$ al celor 2 388 de titluri Twitter de validare, care au rolul populației; f = scorul net FinBERT sau Qwen2.5-7B, fix
  - Schema: n titluri etichetate (cu Y) și $N = 2\,388 - n$ neetichetate (doar cu f), extrase independent, cu întoarcere
- **Cerințe**
  (a) Pentru o extragere cu $n = 200$, calculați estimarea doar din etichete $\bar{Y}_n \pm 1{,}96\, s_Y/\sqrt{n}$ ($\bar{Y}_n$, s_Y: media și abaterea standard a celor n etichete).
  (b) Calculați estimatorul PPI Angelopoulos et al. (2023): $\hat{\theta}$ (slide-ul Holm, PPI și κ), SE $= \sqrt{\hat{V}(f)/N + \hat{V}(f-Y)/n}$, pentru ambele modele.
  (c) Pentru $n = 100, 200, 400, 800$, repetați de 2 000 de ori și înregistrați acoperirea lui θ și lățimea medie a CI.
- **Raportați**: estimările dintr-o extragere și un grafic cu acoperirea și lățimea
- **Interpretare**: de ce media simplă a modelului $\frac{1}{N}\sum f$ nu estimează θ?

înapoi Lista exercițiilor

B8 [Rezolvat]: rezolvare (1/2)



MFM_ch15_valid_inference

B8 [Rezolvat]: rezolvare (2/2)

- În simulare $\theta = 0{,}054$; mediile simple ale modelelor FinBERT $-0{,}014$ (semn greșit), Qwen $+0{,}101$: distorsiunea nu scade cînd N crește
- (a) Doar etichete: $0{,}125 \pm 1{,}96 \times 0{,}040$; (b) PPI-FinBERT: media lui $f - Y$ pe etichetate $-0{,}068$, $\hat{\theta} = 0{,}065 \pm 1{,}96 \times 0{,}037$; PPI-Qwen: media lui $f - Y$ $+0{,}036$, $\hat{\theta} = 0{,}075 \pm 1{,}96 \times 0{,}039$
- (c) $n = 100$: acoperirea 94,7%, 94,3%, 95,4%, lățimea 0,229, 0,204, 0,203 (etichete, PPI-FinBERT, PPI-Qwen)
- Reducerea variației asimptotice ($N \to \infty$), $1 - \mathrm{Var}(f - Y)/\mathrm{Var}(Y)$: 24% (FinBERT) și 27% (Qwen)
- La $n = 800$ PPI este mai larg: 0,087 (FinBERT) și 0,097 (Qwen), față de 0,081 doar din etichete
- Interpretare: media modelului converge la o valoare greșită; eșantionul etichetat o corectează, iar PPI este mai îngust doar cînd $\mathrm{Var}(f - Y) < \mathrm{Var}(Y)$ și multe texte sînt neetichetate

MFM_ch15_valid_inference

B9 [Propus]: wild cluster bootstrap cu 16 acțiuni — enunț

- **Model**: B2 [Rezolvat]
- **Întrebare**: cu doar 16 grupuri de acțiuni, care dintre cele nouă pante din B2 rămîn semnificative?
- **Date**
  - Panelul FNSPID din B2: 18 282 de zile-acțiune, 16 acțiuni americane; scorul $z_{i,d}$ și randamentul în exces r^e ca în B2
  - Cele nouă regresii din B2 (trei scoruri $\times$ $h = 0, 1, 2$), SE grupate pe acțiuni ($G = 16$)
- **Cerințe**
  (a) Calculați valoarea p a fiecărei pante cu valori critice $t_{\mathbf{15}}$.
  (b) Aplicați wild cluster bootstrap cu ipoteza nulă impusă Cameron, Gelbach & Miller (2008) (slide-ul Holm, PPI și κ), ponderi Rademacher $w_g = \pm 1$, $g(i)$: acțiunea observației i, $\bar{y}$: valoarea ajustată sub H_0 (doar termenul liber), 9 999 de extrageri.
  (c) Calculați t cu grupare pe două dimensiuni: fiecare componentă cu propriul $G/(G-1)$, fără factorul $(n-1)/(n-2)$ din B2; componenta intersecției, un rînd pe zi-acțiune, este covariança robustă la heteroscedasticitate.
  (d) Ajustați cele nouă valori p bootstrap cu Holm la 5%.
- **Raportați**: un tabel cu t, cele trei valori p și t pe două dimensiuni
- **Interpretare**: rămîne semnificativă panta Qwen la $h = 2$?

◂ Înapoi ▸ Lista exercițiilor

B10 [Propus]: curba specificațiilor pentru strategia de sentiment — enunț

- **Model**: slide-ul Intervale de încredere bootstrap
- **Întrebare**: pe toate variantele rezonabile ale strategiei, există mai multe dovezi de predictibilitate decît ar produce întîmplarea?
- **Date**
  - Strategia de sentiment din A3 (slide-ul Strategia de sentiment): poziția $\text{sign}(s_{i,d})$ de la închiderea din $d+1$, acoperită cu SPY, ținută o zi; patru tranzacții în fiecare zi cu poziții; 3 568 de zile de tranzacționare
  - Grila: trei scoruri $\times$ praguri $|s| > b$, $b \in \{0; 0{,}25; 0{,}5\}$, $\times$ ziua randamentului $d+2, \ldots, d+5$: 36 de strategii
  - Randamente brute, înainte de costuri, și praguri pe scorurile brute; pentru ziua $d+k$, poziția se deschide la închiderea din $d+k-1$ și se ține o zi
- **Cerințe**
  (a) Calculați randamentul zilnic mediu și t_{NW} pentru fiecare strategie și reprezentați statisticile t sortate Simonsohn, Simmons & Nelson (2020).
  (b) Simulați ipoteza nulă comună: schimbați semnul pozițiilor în blocuri de 10 zile de tranzacționare consecutive din calendarul complet de la prima știre, inclusiv zilele cu randament zero (dependent wild bootstrap Shao (2010)), aceleași schimbări pentru toate 36, de 2 000 de ori.
  (c) Calculați valorile p comune pentru numărul de $t > 1{,}96$ și pentru t median.
- **Raportați**: curba sortată și cele două valori p comune
- **Interpretare**: ce ar ascunde raportarea „celei mai bune specificații”?

înapoi Lista exercițiilor

A8 [Propus]: strategia Qwen după costuri — enunț

- **Model**: A3 [Rezolvat]
- **Întrebare**: strategia Qwen2.5-7B are un randament brut mediu pozitiv statistic: este ea tranzacționabilă?
- **Date de intrare**
  - Strategia de sentiment din A3 (slide-ul Strategia de sentiment): poziția $\text{sign}(s_{i,d})$ de la închiderea din $d+1$, acoperită cu SPY, ținută o zi; patru tranzacții în fiecare zi cu poziții; 3 568 de zile de tranzacționare
  - Semnalul Qwen2.5-7B: $\hat{\mu} = 5{,}64$ bp, $t_{NW} = 2{,}26$ ($L = 8$), $a = 0{,}950$, raportul Sharpe brut 0,63
- **Cerințe**
  (a) Calculați costul de echilibru c^*.
  (b) Calculați SE(c^*) prin metoda delta și CI 95%.
  (c) Obțineți σ și aproximați raportul Sharpe net la 5 bp, cu volatilitatea brută fixă, ca în A3.
- **Raportați**: c^*, SE, CI, raportul Sharpe net
- **Interpretare**: un randament brut semnificativ implică profit după costuri?

◂ Înapoi ▸ Lista exercițiilor

B11 [Propus]: studiu de eveniment cu tonul Loughran–McDonald — enunț

- **Model**: B2 [Rezolvat]
- **Întrebări**
  - în jurul zilelor cu știri negative și pozitive, cînd se mișcă prețul?
  - separă dicționarul cele două grupuri la fel de bine ca FinBERT?
- **Date**
  - Panelul FNSPID din B2: 18 282 de zile-acțiune, 16 acțiuni americane; scorul $z_{i,d}$ și randamentul în exces r^e ca în B2
  - Păstrați zilele repetate de știri: ferestrele aceleiași acțiuni se pot suprapune. Mediați mai întîi pe data evenimentului și aplicați corecția Newey–West indicată pe seria acestor date.
  - Tonul zilnic LM este exact 0 în 48% dintre zilele-acțiune: grupurile ton < 0 (5 196 de zile-acțiune) și ton > 0 (4 288); FinBERT: tercila inferioară și cea superioară
- **Cerințe**
  (a) Pentru fiecare dată, mediați randamentele în exces din zilele $-5, \ldots, +5$ pe zilele-acțiune ale acelei date, în fiecare grup.
  (b) Calculați media pe date a CAR pentru $[-5, -1]$, ziua 0, ziua 1 și $[+2, +5]$, cu t_{NW}.
  (c) Reprezentați traiectoriile cumulate medii ale celor patru grupuri.
- **Raportați**: tabelul celor patru ferestre și graficul
- **Interpretare**: este randamentul de după eveniment pozitiv în ambele grupuri? Distingeți deplasarea de supraviețuire (*survivorship bias*: cele 16 acțiuni sînt firme mari de azi) de mecanismul de revenire din Tetlock (2007).

◂ Înapoi ▸ Lista exercițiilor

C2 [Propus]: profitabilitatea sentimentului titlurilor după costuri — enunț

- **Model**: A3, B2 [Rezolvat]
- **Întrebare**: este profitabilă după costuri realiste o strategie zilnică bazată pe sentimentul titlurilor FNSPID?
- **Date**
  - Strategia de sentiment din A3 (slide-ul Strategia de sentiment): poziția $\text{sign}(s_{i,d})$ de la închiderea din $d+1$, acoperită cu SPY, ținută o zi; patru tranzacții în fiecare zi cu poziții; 3 568 de zile de tranzacționare
  - Poziții ponderate egal; zilele fără știri: randament 0; 5 bp pe tranzacție; scorurile FinBERT, Qwen2.5-7B, LM din tabelele salvate
- **Cerințe**
  (a) Pentru FinBERT, calculați randamentul brut zilnic mediu, t_{NW} și un CI bootstrap pe blocuri suprapuse de 95% (blocuri de 10 zile, 2 000 de extrageri).
  (b) Calculați raportul Sharpe înainte și după costuri, costul de echilibru și raportul Sharpe la 0, 1, 2, 5, 10 bp.
  (c) Repetați pentru Qwen2.5-7B și LM și pentru știrile puternice $|s| > 0{,}5$ (exploratoriu); Holm pe cele trei strategii nefiltrate sau pe toate șase dacă alegeți dintre ele.
  (d) Pas suplimentar: corectați panta din aceeași zi cu un eșantion aleator de zile-acțiune etichetat de oameni (împărțiți la $\rho(s^*, \hat{s})$ sau corectați momentele OLS ca în B8); 200 de zile-acțiune extrase cu seed 15 înainte de a vedea randamentele, etichetate independent de doi evaluatori.
- **Raportați**: o notă de două pagini cu un tabel de rezultate și grafice și un folder Quantlet
- **Interpretare**: discutați calendarul, deplasarea de anticipare (B3) și de supraviețuire, *crowding* Lopez-Lira & Tang (2026)

◂ Înapoi ▸ Lista exercițiilor

C3 [Propus]: replicarea Lopez-Lira și Tang pe FNSPID — enunț

- **Model**: C2, B2, B3, B9
- **Întrebare**: aplicat cu un model deschis pe titlurile cursului, găsește protocolul publicat pentru GPT-4 (Generative Pre-trained Transformer 4) Lopez-Lira & Tang (2026) aceeași reacție și același drift?
- **Date de intrare**
  - Lucrarea: GPT-4 răspunde `YES`, `NO` sau `UNKNOWN` despre fiecare titlu; $s = 1, -1, 0$ determină un portofoliu long-short zilnic; drift-ul pentru știrile dintre ședințe 0,34% pe zi, raport Sharpe 2,97 (tabelul 1)
  - Date: titlurile FNSPID ale celor 16 acțiuni, prețurile ajustate din datele cursului; primul randament sigur $d + 2$
- **Cerințe**
  (a) Evaluați fiecare titlu, extrageți doar prima linie, transformați `YES`/`UNKNOWN`/`NO` în $1/0/-1$; raportați ponderile și răspunsurile nevalide.
  (b) Mediați pe zi-acțiune și formați portofoliul zilnic cu convențiile de pe slide-ul următor, cu randamentele din ziua $d + 2$.
  (c) Raportați media zilnică (bp), t_{NW}, rata de succes, raportul Sharpe, ponderile zilelor cu o componentă și fără poziție.
  (d) Estimați ec. (1) a lucrării, $r_{i,d+h} = a_i + \delta_d + b\, s_k + \varepsilon_k$ pe titlurile k, cu efecte fixe de acțiune a_i și de dată δ_d, $h = 0, 2$, SE grupate pe date și acțiuni, p wild cluster bootstrap (B9).
  (e) Recalculați raportul Sharpe după 5, 10, 20 bp pe dus-întors (*round trip*: intrare și ieșire; figura 4) și comparați cu C2 pe aceleași zile-acțiune.
- **Raportați**: cifrele obținute lîngă tabelele 1 și 2 din lucrare, randamentele cumulate brute și nete, un folder Quantlet
- **Interpretare**
  - ce diferențe față de lucrare pot fi separate cu datele cursului?
  - ce indică B3 despre distorsiune?

◂ Înapoi ▸ Lista exercițiilor

C3 [Propus]: convențiile pentru portofoliu, calendar și costuri

- Fixate înaintea oricărei rulări
  - Modelul: Qwen2.5-7B-Instruct Qwen Team (2024), temperatura 0, prompt-ul din secțiunea 5 a lucrării, nemodificat
  - Ponderile (tabelul 1): componenta long = randamentul mediu, cu ponderi egale, al acțiunilor cu $s > 0$, componenta short = la fel pentru $s < 0$, fiecare doar cu cel puțin două acțiuni; long-short = long minus short (expunere brută 2); dacă se califică o singură componentă, se deține doar aceasta (nota 20)
  - Randamentele: randamentele brute ale acțiunilor, ca în tabelul 1; varianta acoperită cu SPY (ca în C2) este o adaptare a cursului, raportată ca analiză de sensibilitate: cu ambele componente acoperirea se anulează, deci contează doar în zilele cu o singură componentă
  - Calendarul: toate zilele de tranzacționare, ca în C2, cu randament 0 fără poziție; media, t_{NW} și raportul Sharpe pe toate zilele; rata de succes pe zilele cu poziții
  - Costurile (figura 4): c bp pe dus-întors pentru fiecare unitate de valoare nominală: $2c$ pe zi cu două componente, c cu una, plus, în varianta acoperită, c pe valoarea nominală netă SPY (1 în zilele cu o singură componentă)
  - Regresiile (d): randamente brute în % pe zi, ca în tabelul 2; efectele fixe de dată absorb randamentul pieței; wild cluster bootstrap păstrează ambele efecte fixe în modelul nul

Acronime folosite în acest material

- **AI**: Artificial Intelligence — *inteligență artificială*
- **AUC**: Area Under the (ROC) Curve — *aria de sub curba ROC*
- **BERT**: Bidirectional Encoder Representations from Transformers — *reprezentări bidirecționale produse de codificatorul unui Transformer*
- **BVB**: Bursa de Valori București
- **CAR**: Cumulative Abnormal Return — *randamentul anormal cumulat*
- **CI**: Confidence Interval — *interval de încredere*
- **ECE**: Expected Calibration Error — *eroarea medie de calibrare*
- **FNSPID**: Financial News and Stock Price Integration Dataset — *setul de date care integrează știri financiare și prețuri de acțiuni*
- **GPT-4**: Generative Pre-trained Transformer 4 — *Transformer generativ pre-antrenat, versiunea 4*
- **LLM**: Large Language Model — *model mare de limbaj*
- **LM**: Loughran-McDonald (financial sentiment dictionary) — *dicționarul financiar de sentiment Loughran-McDonald*
- **LR**: Logistic Regression — *regresie logistică*
- **NLL**: Negative Log-Likelihood — *logaritmul negativ al verosimilității*
- **OLS**: Ordinary Least Squares — *metoda celor mai mici pătrate*
- **PPI**: Prediction-Powered Inference — *inferență susținută de predicții*
- **ROC**: Receiver Operating Characteristic (curve) — *curba caracteristică de operare a receptorului*
- **SD**: Standard Deviation — *abatere standard*
- **SE**: Standard Error — *eroare standard*
- **TF-IDF**: Term Frequency - Inverse Document Frequency — *frecvența termenului înmulțită cu inversul frecvenței în documente*
- **UTC**: Coordinated Universal Time (Temps Universel Coordonné) — *timpul universal coordonat*

Bibliografie (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.

Bibliografie (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.

Bibliografie (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.