








Modelarea piețelor financiare

Capitolul 15: LLM-uri și analiza sentimentului



Daniel Traian PELE

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

Putem tranzacționa sentimentul citit de modele în știri?

> *"I find that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume."*
>
> — *Tetlock (2007)*

- Capitolul compară dicționare, FinBERT, LLM-uri (*large language models*, modele mari de limbaj) zero-shot și embeddings pe aceleași texte etichetate
  - pe Twitter, cel mai bun este cel mai simplu model supervizat, TF-IDF (frecvențe de cuvinte ponderate) + regresie logistică: acurațețe 83,2%, față de 74,2% pentru Qwen2.5-7B și 72,5% pentru FinBERT
- Pe 16 acțiuni americane mari, cu 106 820 de titluri de știri, scorul explică mișcarea prețului din aceeași zi
  - FinBERT: +39,2 bp pentru o abatere standard a scorului ($t = +17{,}27$); bp: puncte de bază, 1 bp = 0,01%
- O strategie zilnică long-short după semnul scorului nu aduce profit net
  - Qwen2.5-7B: +5,64 bp pe zi înainte de costuri, $p = 0{,}071$ după corecția Holm
  - costul de echilibru, 1,5 bp pe tranzacție, este sub costul presupus de 5 bp
- LLM-urile își amintesc istoria piețelor
  - întrebat doar de lună, Qwen2.5-14B separă lunile de creștere ale S&P 500 (*Standard & Poor's*) de cele de scădere cu AUC 0,69 (aria de sub curba ROC; 0,5 = nicio discriminare)

▸ Graficul central: Randamentele cumulate ale strategiei: niciun profit net

Pot modelele citi sentimentul știrilor și îl putem tranzacționa?

De la text la semnalul de tranzacționare

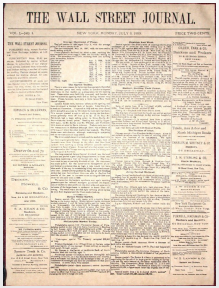
- Partea I: textul ca date, dicționare, evaluarea clasificatorilor, FinBERT, LLM-uri și embeddings
  - LLM-uri: large language models, modele mari de limbaj
- Partea a II-a: deplasarea de anticipare (*look-ahead bias*), semnalul de tranzacționare, scorurile ca estimări
  - Deplasarea de anticipare: folosirea informației din viitor
  - Două studii de caz și bucla de cercetare asistată de AI (artificial intelligence, inteligență artificială)
- Exemplele detaliate și subiectele avansate sînt în Aprofundări, după Concluzii

Rezultatele învățării

- calculați sentimentul unui text cu un dicționar, cu FinBERT și cu un LLM zero-shot
- interpretați un test pe perechi pentru doi clasificatori
- interpretați testele de suprapunere și de memorare dintr-un backtest cu LLM
- interpretați testul unui semnal de sentiment
  - cu aliniere corectă în timp și inferență pentru puține grupuri
  - cu costuri și corecție pentru testarea multiplă
- derivați atenuarea pantei pe un regresor generat
- interpretați intervalele prediction-powered inference
- Aprofundări și Anexa: lectură pentru familiarizare

Piețele reacționează la cuvinte

- Prețurile reacționează la informație, iar cea mai mare parte a ei vine sub formă de text
  - Știri, raportări ale companiilor, conferințe despre rezultate, rapoarte ale analiștilor, comunicate ale băncilor centrale, rețele sociale
- Ziarele financiare sînt la fel de vechi ca piețele organizate
  - The Wall Street Journal: primul număr pe 8 iulie 1889
- Astăzi: milioane de documente pe zi, peste ce poate citi orice analist


THE WALL STREET JOURNAL.

NEW YORK, MONDAY, JULY 8, 1889. PRICE TWO CENTS.

The Wall Street Journal, primul număr, 8 iulie 1889

Foto: Dow Jones & Company (1889); domeniu public; Wikimedia Commons

Textul ca date și definiția sentimentului

- Gentzkow, Kelly & Taddy (2019): un document d este reprezentat numeric prin x_d, folosit într-un model statistic obișnuit
  - x_d: frecvențe, scoruri, vectori
  - x_d devine o mărime de interes (sentiment, temă, risc), folosită apoi pentru a analiza o problemă economică (randamente, volatilitate)
- **Sentimentul unui text**: o veste bună, rea sau neutră pentru activul menționat
  - Trei clase sau scorul net $s = P(\text{pozitiv}) - P(\text{negativ}) \in [-1, 1]$
  - Nu este **sentimentul investitorilor**, starea de spirit măsurată din date de piață (Baker & Wurgler (2006))
- Sinteze pentru finanțe: Kearney & Liu (2014), Loughran & McDonald (2016)
- În acest capitol, scorul este o **estimare** a tonului adevărat, cu eroare de măsurare

Primele dovezi și narațiunile

- Antweiler & Frank (2004): mesajele de pe forumurile despre acțiuni ajută la prognoza volatilității
  - Efectul asupra randamentelor este mic
- Tetlock (2007): pesimismul ridicat dintr-o rubrică a Wall Street Journal anunță presiuni de scădere a prețurilor
  - Presiunile sînt urmate de o revenire
  - Pesimismul neobișnuit de mare sau de mic anunță și volume mari de tranzacționare
- Tetlock, Saar-Tsechansky & Macskassy (2008): ponderea cuvintelor negative din știrile despre o firmă anticipează profituri și randamente scăzute
  - Prețurile încorporează această informație doar treptat
- Shiller (2017): narațiunile populare se răspîndesc ca epidemiile și influențează deciziile economice
  - Ele pot fi măsurate prin text
  - Shiller a primit Premiul Nobel pentru Economie în 2013, împreună cu Fama și Hansen
  - Narațiunile pot îndepărta prețurile de valorile fundamentale (capitolele 2 și 17)



Robert J. Shiller, Stockholm, decembrie 2013

Foto: Bengt Nyman (2013); CC BY 2.0; Wikimedia Commons

Patru moduri de a transforma textul în sentiment

Metoda	Principiul	Etichete necesare	Exemple
Dicționar	numără cuvintele din listă	niciuna	GI, LM, VADER
Model cu fine-tuning	rețea antrenată pe propoziții etichetate	mii	FinBERT
Model mare de limbaj	instrucțiune în limbaj natural (prompt)	niciuna (zero-shot)	Qwen2.5, GPT
Embeddings + clasificator	vectorii textelor într-o regresie	sute pînă la mii	MiniLM

GI: Harvard General Inquirer, o listă de cuvinte de uz general; LM: dicționarul financiar Loughran–McDonald; VADER: Valence Aware Dictionary and sEntiment Reasoner, un dicționar cu reguli pentru rețelele sociale; GPT: Generative Pre-trained Transformer, modelele de limbaj ale OpenAI

- Le comparăm pe toate patru pe aceleași date etichetate, cu intervale de încredere (CI, *confidence intervals*) de 95%
- Apoi testăm dacă vreuna produce cîștiguri pe randamentele zilnice ale acțiunilor
- O a cincea cale, regresiile pe text cu fundamentare statistică, nu este testată aici (▸ Anexa)
  - Autori: Taddy; Ke, Kelly și Xiu

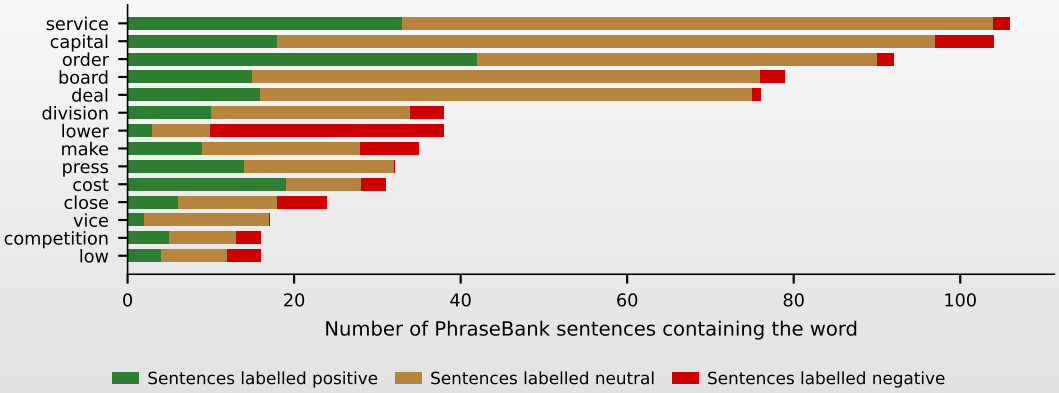
Poate o listă de cuvinte măsura tonul unui text financiar?

- **Reprezentarea prin frecvențele cuvintelor (bag of words)**: documentul este redus la frecvențele cuvintelor
  - Ordinea este ignorată
  - Tokeni: cuvinte fără cifre și punctuație, indiferent de majuscule
- Un dicționar listează cuvintele pozitive $\mathcal{P}$ și cuvintele negative $\mathcal{N}$
  - P_d, N_d: numărul de cuvinte pozitive și negative din documentul d
  - **Tonul**: $\tau_d = \dfrac{P_d - N_d}{P_d + N_d} \in [-1, 1]$, iar $\tau_d = 0$ dacă $P_d + N_d = 0$; clasa este semnul lui τ_d (neutră pentru $\tau_d = 0$)
- Avantaje: transparență, viteză, fără date de antrenare
- Limita: ignoră contextul și negația

Dicționare generale, financiare și cu reguli

- **Harvard General Inquirer** (GI), lista Harvard IV-4: un dicționar de psihologie din anii 1960
  - 2 007 cuvinte negative și 1 637 pozitive
- Loughran & McDonald (2011) (LM): construit din rapoartele anuale (formularele 10-K) ale firmelor americane
  - 2 355 de cuvinte negative și doar 354 pozitive
  - Concluzia lor: aproape trei sferturi dintre cuvintele negative din GI nu sînt negative în context financiar
  - În listele folosite aici, 1 552 de cuvinte negative din GI (77%) nu apar printre cuvintele negative din LM
- **VADER** (Hutto & Gilbert (2014)): circa 7 500 de termeni cu scoruri de intensitate, pentru rețelele sociale
  - Reguli pentru negație („not good"), intensificatori („very"), majuscule, semne de exclamare și emoticoane
  - Scorul compus (compound score) ia valori în $[-1, 1]$, cu pragurile uzuale $\pm 0{,}05$
  - Regulile rezolvă o parte din problemele de context, dar vocabularul este general, ca la GI

Cuvinte negative în psihologie, neutre în contabilitate (1/2)



MFM_ch15_dictionaries ▶ Interactiv

Cuvinte negative în psihologie, neutre în contabilitate (2/2)

- Cuvinte negative în GI, dar nu în LM
  - Ordonate după numărul de propoziții din Financial PhraseBank care le conțin
  - Culoarea: eticheta dată propoziției de adnotatorii umani
- „Service", „capital", „board" și „vice" sînt cuvinte uzuale în limbajul de afaceri, nu vești negative
  - Numărul de propoziții: „service" 106, „capital" 104, „board" 79, „vice" 17 (din „vice president")
  - Ponderea propozițiilor care nu sînt etichetate negativ: 98% („service"), 93% („capital"), 96% („board"), 100% („vice")
- „Lower": doar 26% dintre propoziții nu sînt negative
  - Totuși, „costuri mai mici" este o veste bună
  - Semnul depinde de cuvîntul următor, de care nicio listă nu ține cont
- Din cele 964 de propoziții în care singurele cuvinte negative GI lipsesc din lista negativă LM:
  - 89% nu sînt etichetate negativ

MFM_ch15_dictionaries

Pe ce texte judecăm un clasificator? Două corpusuri etichetate

- Financial PhraseBank (Malo et al. (2014)): 4 846 de propoziții din știri
  - Etichetate de un grup de 16 adnotatori cu pregătire în finanțe, cîte 5–8 pentru fiecare propoziție
  - Întrebarea pusă adnotatorilor: este o veste bună, rea sau neutră pentru investitor?
  - Niveluri de acord: cel puțin 50%, 66%, 75% sau 100% dintre adnotatori
  - 2 264 de propoziții cu acord deplin, 629 cu acord sub 66%
  - Etichete: 59% neutre, 28% pozitive, 12% negative
- Twitter Financial News (Hugging Face Hub, licența MIT): titluri din 2022
  - Etichetate bearish, bullish sau neutre
  - 9 543 de titluri de antrenare și 2 388 de validare, folosite pentru evaluare
  - Etichete: 66% neutre, 20% pozitive, 15% negative
  - Texte scurte, zgomotoase, cu simboluri bursiere și linkuri, apropiate de fluxul real de știri
- Două corpusuri, pentru că un rezultat obținut pe unul s-ar putea să nu se confirme pe celălalt

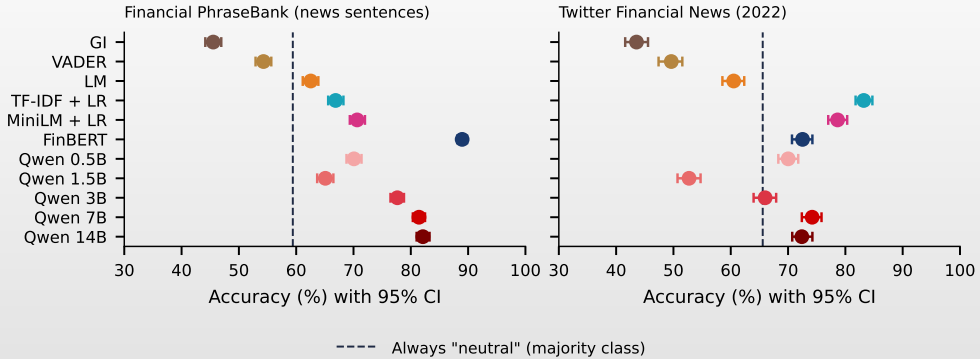
Acurateţe sau F1 macro? Metrici pentru clase dezechilibrate

- Întrebarea: cum comparăm clasificatori cînd o clasă domină?
- Acurateţea: $\text{acc} = \frac{1}{n} \sum_{i=1}^{n} 1\{\hat{y}_i = y_i\}$
- Pentru clasa k:
  - Precizia $P_k = \frac{TP_k}{TP_k + FP_k}$, recall-ul $R_k = \frac{TP_k}{TP_k + FN_k}$, $F1_k = \frac{2P_kR_k}{P_k + R_k}$
  - TP_k: texte din clasa k prezise k; FP_k: texte din alte clase prezise k; FN_k: texte din clasa k prezise altfel
- F1 macro $= \frac{1}{K} \sum_{k=1}^{K} F1_k$, cu $K = 3$ (Capitolul 13)
  - Fiecare clasă contează la fel, oricît de rară
- Reperul „mereu neutru” pe Twitter: acurateţe 66%, dar F1 macro de doar 26,4%
  - $F1_{\textbf{neutru}} = 2 \times 0{,}656 \times 1/(0{,}656 + 1) = 0{,}792$, iar $F1 = 0$ pentru celelalte două clase, deci F1 macro $= 0{,}792/3$

Cînd este un clasificator mai bun? Testul McNemar pe aceleași texte

- Doi clasificatori A și B pe aceleași n texte: acuratețile lor sînt corelate
  - Două CI separate nu răspund la întrebare
- n_{01}: numărul de texte pe care A greșește și B are dreptate
- n_{10}: numărul de texte pe care A are dreptate și B greșește
- Textele pe care ambii clasificatori le clasifică la fel nu aduc informație
  - Diferența de acuratețe este $(n_{01} - n_{10})/n$
- Sub H_0 (acuratețe egală), $n_{01} \sim \text{Binomial}(n_{01} + n_{10}; 0{,}5)$: un test exact
  - Exemplu rezolvat: ▸ Anexa
- Pentru $n_{01} + n_{10}$ mare: $\chi^2 = (|n_{01} - n_{10}| - 1)^2/(n_{01} + n_{10})$ are aproximativ distribuția χ^2_1
  - -1: corecția de continuitate
- Exemplu: A greșește și B are dreptate pe 4 texte, invers pe 1
  - Atunci $p = 2\,P(X \geq 4) = 2 \times 6/32 = 0{,}375$, cu $X \sim \text{Binomial}(5; 0{,}5)$

Acuratețea celor unsprezece clasificatori pe două seturi de date (1/2)



TF-IDF: term frequency–inverse document frequency, frecvențe de cuvinte cu ponderi mai mici pentru cuvintele uzuale; LR: logistic regression, regresie logistică

MFM_ch15_classification ▶ Interactiv

Acuratețea celor unsprezece clasificatori pe două seturi de date (2/2)

- TF-IDF + LR și MiniLM + LR: singurele metode antrenate pe etichete Twitter
- Toate cele trei dicționare rămîn, pe Twitter, sub reperul clasei majoritare
  - Acuratețea pe Twitter: GI 43,6%, VADER 49,6%, LM 60,5%, față de 66% pentru reperul „mereu neutru”
- Acuratețea FinBERT: 89,0% pe PhraseBank și 72,5% pe Twitter
- Acuratețea Qwen2.5-7B zero-shot: 81,4% pe PhraseBank și 74,2% pe Twitter
- Cel mai bun pe Twitter, 83,2%: cel mai simplu model supervizat, TF-IDF + regresie logistică
- Lecția: etichetele din domeniul-țintă contează mai mult decît mărimea modelului
- Tipul de dovadă: comparație între proceduri (arată care metodă funcționează mai bine pe aceleași date, nu de ce)

MFM_ch15_classification

Ajută contextul? De la cuvinte la BERT

- Un dicționar dă cuvîntului „lower" un semn fix
  - Un model de limbaj citește propoziția din jurul lui
  - Fiecare token primește un vector care depinde de toți ceilalți tokeni (atenție, Capitolul 14; Vaswani et al. (2017))
- **BERT** (Bidirectional Encoder Representations from Transformers; Devlin et al. (2019)): un codificator Transformer
  - Pre-antrenat pe cărți și Wikipedia prin predicția cuvintelor mascate aleator
  - Circa 110 milioane de parametri
  - Pentru clasificare se folosește vectorul final $h_{\mathbf{CLS}} \in \mathbb{R}^{\mathbf{768}}$ al tokenului special [CLS (*classification token of BERT*, simbolul de clasificare al modelului BERT)] (de la *classification*, clasificare), pus la începutul fiecărui text
- Fine-tuning: un strat de clasificare peste $h_{\mathbf{CLS}}$
  - Toată rețeaua este antrenată pe propoziții etichetate
  - $p = \text{softmax}(Wh_{\mathbf{CLS}} + c)$, cu $W \in \mathbb{R}^{\mathbf{3}\times\mathbf{768}}$: probabilitățile claselor negativ, neutru și pozitiv
  - Antrenarea minimizează entropia încrucișată $-\frac{1}{n}\sum_i \ln p_{i,y_i}$, cu y_i eticheta umană a propoziției i

FinBERT: BERT adaptat textelor financiare

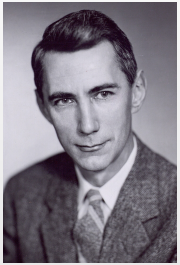
- Araci (2019): FinBERT este BERT pre-antrenat suplimentar pe știri financiare
  - Apoi adaptat prin fine-tuning pentru sentiment
  - Modelul public folosit aici (ProsusAI/finbert) a fost adaptat prin fine-tuning pe Financial PhraseBank
  - Durata: cîteva milisecunde pe propoziție pe un laptop
- Scorul folosit în acest capitol: $s = p_{\text{poz}} - p_{\text{neg}} \in [-1, 1]$, iar eticheta este $\arg\max_k p_k$
- De ce FinBERT ca reper: este public, rapid și larg folosit în literatura de finanțe
- Huang, Wang & Yang (2023): un FinBERT antrenat pe texte financiare depășește dicționarele
  - Texte: rapoarte ale analiștilor, transcrieri ale conferințelor despre rezultate și formulare 10-K
  - Mai ales pe propozițiile pe care acestea le citesc greșit

Distorsiunea evaluării în eșantion

- FinBERT: 89,0% acuratețe pe PhraseBank, pe care a fost antrenat în mare parte
  - Araci (2019): 3 101 din cele 4 845 de propoziții ale versiunii sale (aici 4 846)
- Pe Twitter, FinBERT are 72,5%: o scădere de 16,4 puncte (din valorile nerotunjite)
  - Qwen2.5-7B, fără fine-tuning pe aceste etichete, scade și el cu 7,2 puncte: titlurile Twitter sînt mai greu de clasificat
- Diferența diferențelor: $\Delta = (a_{F,P} - a_{F,T}) - (a_{Q,P} - a_{Q,T}) = 9{,}2$ puncte
  - CI 95% bootstrap [6,7; 11,9]
  - $a_{m,D}$: acuratețea modelului m (F: FinBERT, Q: Qwen) pe datele D (P: PhraseBank, T: Twitter)
  - Ipoteza de identificare, netestabilă: fără suprapunere, ambele modele ar pierde la fel de la PhraseBank la Twitter; CI măsoară doar zgomotul de eșantionare
- Tipul de dovadă: o diferență descriptivă între scăderile de acuratețe
  - Atribuirea lui Δ suprapunerii cere și un comparator necontaminat și nicio altă schimbare diferențială
  - Qwen nu a fost adaptat pe PhraseBank, dar textele pot apărea în datele lui de pretraining
- Regula: evaluați un model de limbaj doar pe texte pe care nu le-a văzut niciodată
  - Ca în testele în afara eșantionului (*out-of-sample*) din capitolele 8 și 13

De la predicția cuvîntului următor la urmarea instrucțiunilor

- Shannon (1951): engleza este foarte predictibilă din literele anterioare
  - Un model mare de limbaj (LLM) estimează $P(w_{t+1} \mid w_1, \ldots, w_t)$ pe tokeni
  - w_t: al t-lea token, un cuvînt sau o parte de cuvînt
  - Brown et al. (2020): la scară mare, sarcinile se rezolvă din cîteva exemple date în prompt (few-shot)
  - Ouyang et al. (2022): ajustarea pe instrucțiuni (*instruction tuning*) și RLHF (*reinforcement learning from human feedback*, învățare prin recompensă din evaluările oamenilor) antrenează modelul să urmeze instrucțiuni
- BloombergGPT (Wu et al. (2023)): 50 de miliarde de parametri, antrenat pe text financiar și general
  - Noțiuni de bază: Quantinar: The Next Word Problem, Quantinar: Natural Language Processing



Claude E. Shannon, fondatorul teoriei informației

Foto: Tekniska museet, Stockholm (fără dată); CC BY 2.0; Wikimedia Commons

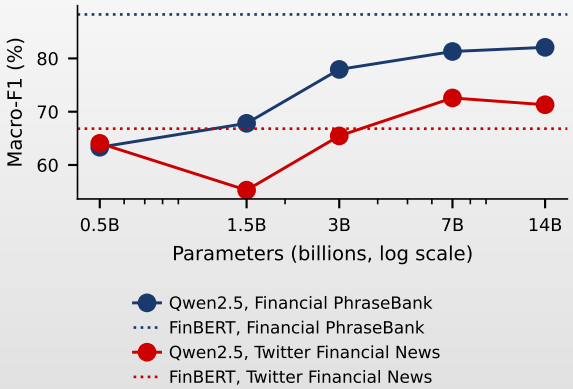
Clasificarea zero-shot cu Qwen2.5: prompt-ul și probabilitățile

- Prompt-ul: o instrucțiune de sistem plus textul de clasificat
  - `You classify the sentiment of financial news for investors. Answer with one word: positive, negative or neutral.`
  - `Headline: <text> Sentiment:`
- Citim logiții z_k (scorurile nenormalizate ale modelului) ai tokenilor „positive", „negative", „neutral" la prima poziție generată
  - $p_k = e^{z_k} / \sum_j e^{z_j}$, normalizat pe cele trei cuvinte-răspuns
  - Eticheta: $\arg\max_k p_k$
  - Scorul: $p_{\mathbf{poz}} - p_{\mathbf{neg}}$, ca la FinBERT

Temperatura și modelele Qwen2.5

- La generare, modelul extrage următorul token din $p_k(T) = e^{z_k/T} / \sum_j e^{z_j/T}$, cu temperatura $T > 0$
  - $T \to 0$: mereu tokenul cel mai probabil (decodare greedy), deci același răspuns la fiecare rulare; Lopez-Lira & Tang (2026) folosesc temperatura 0
  - $T = 1$: distribuția modelului; $T > 1$ aplatizează probabilitățile, deci răspunsurile variază mai mult
- Aici nu se generează text
  - Eticheta $\arg\max_k p_k$ este răspunsul de la $T \to 0$
  - Scorul păstrează probabilitățile de la $T = 1$, fără extragerea etichetei dintr-un răspuns textual
- Qwen Team (2024): modele cu ponderi deschise, de la 0,5 la 72 de miliarde de parametri
  - Licențe: Apache 2.0; Qwen Research License pentru 3B
  - Publicate pe 19 septembrie 2024
- Folosim variantele 0.5B, 1.5B, 3B, 7B și 14B
  - B: miliarde de parametri
  - Timpii de procesare: Anexa
- Ponderi deschise: rezultatele pot fi reproduse
  - Data publicării limitează ce putea ști modelul

Dimensiunea modelului și acuratețea zero-shot

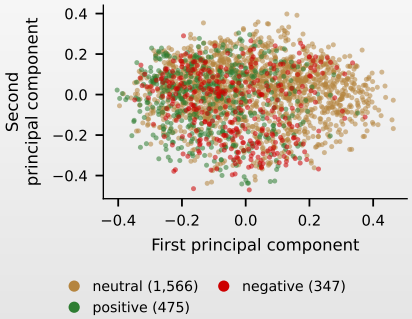


- F1 macro pentru Qwen2.5 zero-shot cu același prompt, iar liniile punctate arată FinBERT
- PhraseBank: F1 macro crește de la 63,3% (0.5B) la 82,1% (14B), tot sub FinBERT (88,3%)
  - Avantajul FinBERT de acolo provine parțial din suprapunerea cu datele de antrenare (Distorsiunea evaluării în eșantion, *in-sample*)
- Pe Twitter, F1 macro nu crește monoton cu mărimea
  - 1.5B este cel mai slab (55,3%)
  - 7B este cel mai bun (72,6%), peste FinBERT (66,8%)
  - Diferența de acuratețe dintre 7B și 1.5B este de +21,5 puncte procentuale (testul McNemar: $p < 0{,}001$)
- Modelele mici sînt foarte sensibile la formulare
  - Modelul 1.5B răspunde rar „neutral” cu acest prompt

MFM_ch15_llm_zero_shot ▶ Interactiv

▸ Aprofundări: prompt, calibrare, exemplu

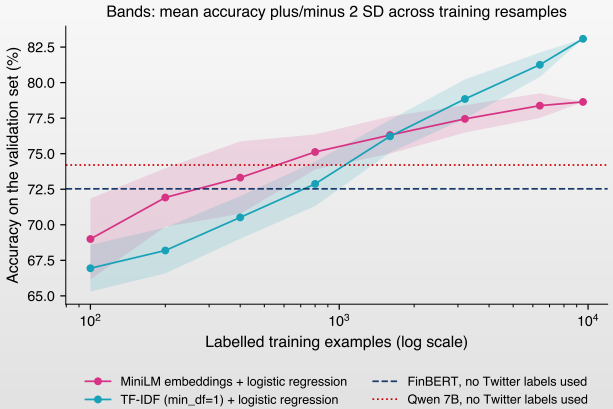
Pot vectorii generici ai textelor măsura sentimentul?



- **Embedding**: un vector $e_d \in \mathbb{R}^k$, apropiat pentru texte asemănătoare
  - word2vec (Mikolov et al. (2013)), Sentence-BERT (Reimers & Gurevych (2019))
  - Aici: MiniLM (Wang et al. (2020)), cu $k = 384$ și vectori de normă 1, deci similaritatea este $\cos(e_a, e_b) = e_a^\top e_b$
  - Embeddings intră într-un clasificator sau într-o regresie ca orice altă variabilă
- Graficul: titlurile din setul de validare Twitter, colorate după etichetă
  - Axele: primele două componente principale ale vectorilor MiniLM
- Primele două componente nu separă sentimentul
  - PCA (principal component analysis, analiza componentelor principale; Capitolul 3) păstrează direcțiile de variață maximă, nu pe cele legate de etichetă
- O regresie logistică pe toate cele 384 de dimensiuni găsește direcția sentimentului
  - 78,6% acurateță pe Twitter

MFM_ch15_embeddings ▶ Interactiv

Numărul necesar de etichete (1/2)



MFM_ch15_embeddings ▶ Interactiv

Numărul necesar de etichete (2/2)

- Regresie logistică antrenată pe submulțimi aleatoare ale setului de antrenare Twitter
  - 10 repetări pentru fiecare mărime a eșantionului, iar banda arată ± 2 abateri standard
- Embeddings + regresie logistică: 69,0% acuratețe cu 100 de etichete
  - Peste FinBERT începînd de la 400 de etichete
  - Vectorii pre-antrenați conțin deja cea mai mare parte a informației necesare
- TF-IDF pornește mai jos (66,9%), dar continuă să crească
  - Circa 83% cu toate etichetele, față de 78,6% pentru embeddings
  - Cu multe etichete, cuvintele exacte („upgrade", „cuts", „misses") conțin semnalul în titluri
- Regulă practică: etichetați cîteva sute de documente din domeniul de interes
  - Înainte de a vă baza pe un model gata antrenat

MFM_ch15_embeddings

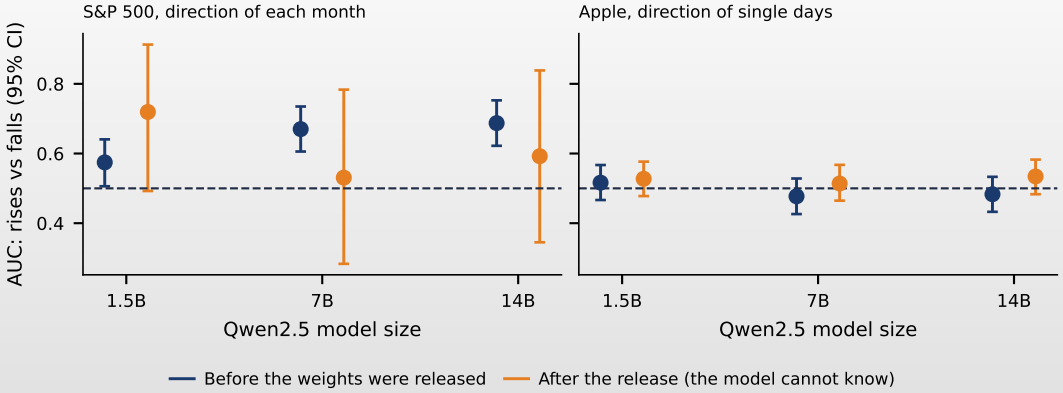
Memorarea datelor de test de către modelele de limbaj

- Un LLM antrenat pe texte de pînă în 2024 poate să fi citit ce s-a întîmplat după un titlu din 2015
  - Un backtest pe știrile din 2015 confundă atunci prognoza cu memorarea
- Glasserman & Lin (2023): două distorsiuni în backtest-urile de sentiment
  - Deplasarea de anticipare: modelul cunoaște randamentele care au urmat știrii
  - Efectul de distragere: cunoștințele generale despre compania numită interferează cu citirea textului
  - În testele lor, distragerea a contat mai mult decît deplasarea de anticipare
- Sarkar & Vafa (2024): teste directe arată deplasarea de anticipare în aplicații ale LLM-urilor
  - La conferințe despre rezultate și la alegeri

Testul memoriei: întrebări despre trecut, fără context

- Întrebarea lunară: „Did the S&P 500 index rise or fall during <lună an>?"
  - Pusă pentru 296 de luni dinainte de publicarea ponderilor și pentru 23 de luni de după
  - Întrebarea zilnică: „Did Apple stock close higher or lower on <dată> than on the previous trading day?", pusă pentru 500 de zile aleatoare din 2010–2023 și pentru 493 de zile de după octombrie 2024
- Scorul: P(creștere), calculat din probabilitățile cuvintelor-răspuns
- Direcția adevărată: din prețurile de închidere lunare ale S&P 500 și din cele zilnice ale Apple
  - AUC (aria de sub curba ROC, *receiver operating characteristic*): probabilitatea ca o perioadă aleatoare de creștere să primească un scor mai mare decît una aleatoare de scădere
  - AUC = 0,5 înseamnă nicio capacitate de discriminare (Capitolul 13)
- După publicare nu mai există nimic de memorat
  - Dacă data nu conține informație predictivă (de exemplu sezonalitate), AUC în populație este 0,5
  - AUC empiric variază în jurul acestei valori

Memoria măsurată: lunile da, zilele nu (1/2)



MFM_ch15_lookahead ▶ Interactiv

Memoria măsurată: lunile da, zilele nu (2/2)

- AUC cu intervale bootstrap de 95%, cu lunile reeșantionate independent
- Lunile dinainte de publicare: AUC 0,67 pentru 7B și 0,69 pentru 14B
  - CI [0,61; 0,73] pentru 7B și [0,62; 0,75] pentru 14B
  - Testul prin permutări pentru 14B: $p < 0{,}001$
- Modelul 1.5B știe mult mai puțin: AUC 0,57
- După publicare (23 de luni) intervalele includ 0,5, dar sînt prea largi pentru o concluzie
  - Ar fi nevoie de circa 101 luni după publicare pentru a detecta o cădere la 0,5
  - Testul formal: ▸ Anexa
- Zile individuale pentru Apple: AUC între 0,48 și 0,52 înainte de publicare, fără discriminare detectabilă peste nivelul aleator
- Deplasarea de anticipare este un pericol pentru evenimentele bine cunoscute
  - Mai puțin pentru zgomotul zilnic
- Tipul de dovadă: o descriere; discriminarea lunilor este compatibilă cu memorarea, dar nu o separă de regularitățile calendaristice sau de regim

MFM_ch15_lookahead

Remedii: modele disponibile la data analizată

- Testați doar după data-limită a datelor de antrenare
  - Lopez-Lira & Tang (2026) evaluează GPT-4 (a patra versiune a modelelor GPT) pe titluri de după data-limită
  - Contaminarea este eliminată, dar fereastra de test începe tîrziu
- Mascați numele companiilor și datele înainte de evaluare (Glasserman & Lin (2023))
  - Apoi comparați pantele cu și fără mascare
  - Este un diagnostic: produsele, evenimentele sau citatele pot identifica în continuare episodul
- Modele cu informație limitată la o dată de referință (Sarkar & Vafa (2024))
  - Antrenate doar pe textul disponibil pînă la data t
  - Drinkall et al. (2024): o serie de LLM-uri cu date de referință succesive, fiecare fără informații despre faptele de după data lui, iar modelele și datele lor de antrenare sînt publice
  - He et al. (2025): ChronoBERT și ChronoGPT folosesc doar textul disponibil la fiecare dată
  - La prognoza randamentelor zilei următoare din știri, rapoartele lor Sharpe sînt apropiate de cele ale unui model mult mai mare: acolo, deplasarea de anticipare are un efect modest
- Raportați versiunea modelului și data publicării ponderilor sale împreună cu fiecare rezultat

Titluri datate, potrivite cu prețurile zilnice

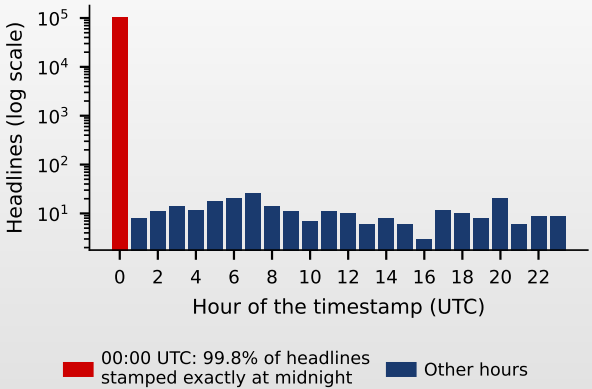
- Întrebarea secțiunii: prognozează tonul titlurilor randamente pe care le mai putem tranzacționa?
- FNSPID (Dong, Fan & Peng (2024)): 15,7 milioane de știri despre firmele din S&P 500, 1999–2023
  - Financial News and Stock Price Integration Dataset
  - Știri de pe patru site-uri de știri
  - Licența Creative Commons CC BY-NC 4.0 (atribuire, utilizare necomercială)
- Aici: 16 acțiuni americane mari cu prețuri zilnice în datele cursului
  - Tehnologie, bănci, GameStop, Strategy
  - 106 820 de titluri distincte, după eliminarea repetițiilor
  - 18 282 de zile-acțiune cu știri (perechi acțiune–zi cu cel puțin un titlu), între 27.10.2009 și 18.12.2023



Bandă de știri pe clădirea Reuters, New York

Foto: bgilliard (2005); CC BY-SA 2.0; Wikimedia Commons

Marcajul de timp al știrilor: data fără ora publicării (1/2)



MFM_ch15_news_data ▶ Interactiv

Marcajul de timp al știrilor: data fără ora publicării (2/2)

- Ora marcajului de timp FNSPID în timpul universal coordonat (UTC), scară logaritmică
- 99,8% dintre titluri au marcajul exact 00:00 UTC: sursa dă data, nu ora publicării
  - Un titlu datat d poate apărea înainte de deschidere, în timpul ședinței sau după închiderea zilei d
- Alinierea folosită aici, fără deplasare de anticipare
  - Ziua d: prima zi de tranzacționare egală cu data titlului sau ulterioară
  - Randamentele din d și $d+1$ pot conține încă reacția
  - Prima prognoză sigură: randamentul din $d+2$, de la închiderea din $d+1$ la închiderea din $d+2$

MFM_ch15_news_data

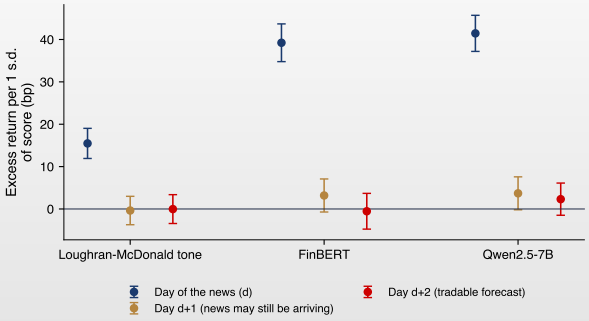
Trei scoruri și randamentul în exces

- Cele trei scoruri:
  - Tonul LM
  - Scorul net FinBERT $P(\text{poz}) - P(\text{neg})$
  - Scorul net Qwen2.5-7B cu prompt-ul P1 (cel de pe slide-ul despre clasificarea zero-shot)
  - LM găsește un cuvînt din listă doar în 27% dintre titluri: restul primesc tonul 0
  - Etichetele FinBERT: 19% negative, 62% neutre, 18% pozitive
  - Corelațiile pe titluri: FinBERT–Qwen 0,62, FinBERT–LM 0,41, Qwen–LM 0,36
- Randamentul în exces: $r^e_{i,t} = r_{i,t} - r_{\mathbf{SPY},t}$
  - SPY: un fond tranzacționat la bursă care urmărește indicele S&P 500
  - Ajustat la piață: $\beta = 1$ pentru fiecare acțiune; alternativa uzuală este $r_{i,t} - \hat{\alpha}_i - \hat{\beta}_i r_{\mathbf{SPY},t}$, cu $\hat{\beta}_i$ estimat înaintea evenimentului

Specificația regresiei și erorile standard grupate pe zile

- Regresie pe toate zilele-acțiune cu știri: $r^{e}_{i,d+h} = a + b\, z_{i,d} + \varepsilon_{i,d+h}$, $h = 0, 1, 2$
  - z: scorul zilnic standardizat pe panel (medie 0, abatere standard 1), deci b este randamentul în exces (%) asociat unei abateri standard a scorului
  - $h = 0$ măsoară reacția, $h = 2$ măsoară predictibilitatea
- Erorile acțiunilor din aceeași zi sînt corelate (șocuri comune ale pieței)
  - Erorile standard (SE) obișnuite din OLS (*ordinary least squares*, metoda celor mai mici pătrate) le ignoră
  - Grupăm pe zile (Petersen (2009))
  - $\widehat{V} = (X^\top X)^{-1} \left(\sum_g X_g^\top \hat{\varepsilon}_g \hat{\varepsilon}_g^\top X_g \right) (X^\top X)^{-1}$
  - g indexează zilele de tranzacționare, X conține regresorii (constanta și z), iar X_g, $\hat{\varepsilon}_g$ sînt rîndurile și reziduurile OLS din ziua g

Reacție față de predicție: titlurile descriu prețurile



- Panta b în puncte de bază pentru o abatere standard a scorului (s.d. în grafic)
  - bp: 1 bp = 0,01%
  - CI 95% grupat pe zile
- Ziua d: FinBERT +39,2 bp ($t = +17{,}27$), Qwen +41,4 bp ($t = +19{,}07$), LM +15,5 bp ($t = +8{,}53$)
  - Multe titluri relatează chiar mișcarea prețului („shares jump", „stock falls"): cauzalitatea merge în ambele sensuri
- Ziua $d + 1$: FinBERT +3,2 bp ($t = +1{,}60$), Qwen +3,7 bp ($t = +1{,}86$)
  - O parte din acest efect provine din știri publicate după închiderea zilei d, nu dintr-o prognoză
- Ziua $d + 2$: FinBERT −0,5 bp ($t = -0{,}25$), Qwen +2,3 bp ($t = +1{,}19$), LM −0,0 bp ($t = -0{,}01$)

MFM_ch15_news_signal ▶ Interactiv

Inferență cu 16 acțiuni: grupare pe acțiuni și grupare dublă

- Întrebarea: rezistă pantele dacă erorile sînt corelate și în timp, pentru aceeași acțiune?
- Efectele de firmă persistă, deci erorile aceleiași acțiuni sînt corelate între zile: grupăm și pe acțiuni
- Gruparea dublă (Cameron, Gelbach & Miller (2011)): $\widehat{V}_{\textbf{dublu}} = \widehat{V}_{\textbf{zi}} + \widehat{V}_{\textbf{acțiune}} - \widehat{V}_{\textbf{zi}\cap\textbf{acțiune}}$
  - Fiecare termen este formula sandwich de pe slide-ul anterior, cu grupurile date de zi, de acțiune, respectiv de perechea acțiune–zi
  - O pereche acțiune–zi are aici o singură observație, deci ultimul termen este estimatorul White, robust doar la heteroscedasticitate
  - Scăderea lui evită numărarea de două ori a variației fiecărei observații
- Driscoll & Kraay (1998) țin cont, în plus, de corelația dintre zile diferite

Puține grupuri: wild cluster bootstrap

- Cu doar $G = 16$ acțiuni, SE grupată pe acțiuni este subestimată, iar testul t respinge prea des (Cameron, Gelbach & Miller (2008))
- Wild cluster bootstrap (reeșantionare cu semne aleatoare pe grupuri) cu ipoteza nulă impusă
  - Estimăm modelul restricționat ($b = 0$), cu constanta $\tilde{a}$ și reziduurile $\tilde{\varepsilon}_{i,t}$
  - Pentru $j = 1, \ldots, B$, $B = 9\,999$: ponderi Rademacher $w_g = \pm 1$ cu probabilitatea $\frac{1}{2}$, o extragere pentru fiecare acțiune g
  - $r^{e*}_{i,t} = \tilde{a} + w_{g(i)}\,\tilde{\varepsilon}_{i,t}$; reestimăm regresia și calculăm t^*_j grupat pe acțiuni
  - $p = \frac{1}{B}\sum_j 1\{|t^*_j| \geq |t|\}$
- Același semn pentru toate reziduurile unei acțiuni păstrează corelația din interiorul grupului
- Bootstrap-ul presupune acțiuni independente între ele, deci citiți-l împreună cu t cu grupare dublă

Nouă regresii: corecția Holm

- Nouă regresii: 3 scoruri $\times$ 3 orizonturi ($h = 0, 1, 2$)
- Dacă niciun scor nu ar avea efect, nouă teste la 5% ar da totuși respingeri false
  - În medie, $9 \times 0{,}05 = 0{,}45$ respingeri false
- Corecția Holm (Capitolul 2) menține FWER la cel mult 5%
  - FWER (family-wise error rate): probabilitatea vreunei respingeri false
  - Ordonăm valorile p: $p_{(\mathbf{1})} \leq \cdots \leq p_{(m)}$, cu $m = 9$
  - $\tilde{p}_{(k)} = \max_{j \leq k} \min\{1, (m - j + 1)\, p_{(j)}\}$; respingem $H_{(k)}$ cînd $\tilde{p}_{(k)} \leq 0{,}05$
  - Mai puternică decît Bonferroni ($m\, p_{(k)}$ pentru toate) și validă oricare ar fi dependența dintre teste

Inferență cu 16 acțiuni și nouă regresii: rezultatele

Scor, orizont	b (bp)	t pe acțiuni	t grupare dublă	p wild bootstrap	p Holm
FinBERT, d	+39,2	+5,71	+5,67	<0,001	<0,001
Qwen, d	+41,4	+5,10	+5,07	<0,001	<0,001
FinBERT, $d+2$	−0,5	−0,35	−0,30	0,739	1,00
Qwen, $d+2$	+2,3	+2,41	+2,01	0,061	0,37
LM, $d+2$	−0,0	−0,03	−0,02	0,979	1,00

- Qwen în $d+2$: $t = +2{,}41$, cu grupare pe acțiuni, pare semnificativ
  - $p = 0{,}029$ cu distribuția t cu 15 grade de libertate
  - Dar nu și după wild bootstrap și corecția Holm
  - Corecția Holm pas cu pas: ordonăm cele 9 valori p bootstrap; Qwen în $d+2$ este a 4-a, după cele trei valori din ziua d (toate sub 0,001), deci p Holm $= (9-4+1) \times 0{,}061 \approx 0{,}37$
  - Rîndurile care nu apar în tabel: LM în ziua d, p <0,001; în ziua $d+1$, $p = 0{,}272$ (FinBERT), 0,317 (Qwen), 0,802 (LM)
- Tipul de dovadă: asociere (variabilele se mișcă împreună; nu dovedește o relație cauzală)
  - Treimea cu scorurile cele mai pozitive are deja un randament în exces cumulat de +0,63% în cele 5 zile dinainte de știre (Aprofundări)

O strategie zilnică long-short de sentiment

- Întrebarea: aduce semnul scorului un profit pe care îl putem încasa efectiv?
- La închiderea din $d+1$: pentru fiecare acțiune cu știri și scor nenul în ziua d
  - Poziție long dacă scorul este pozitiv și short dacă este negativ
  - Fiecare poziție este acoperită (*hedging*) cu SPY; acoperirea unu-la-unu elimină mișcarea pieței doar pentru $\beta = 1$
  - Pozițiile au ponderi egale și se închid după o zi, la închiderea din $d+2$
  - Scorul LM este zero în 48% dintre perechile acțiune–zi: acestea nu intră în portofoliu
- Randamentul brut în $d+2$: $G_{d+2} = \frac{1}{N_d} \sum_{i \in A_d} \text{sign}(s_{i,d}) \, (r_{i,d+2} - r_{\text{SPY},d+2})$
  - $A_d = \{i : |s_{i,d}| > \delta\}$, cu pragul $\delta = 0$: acțiunile cu știri și scor nenul în ziua d
  - $N_d = |A_d|$: pozițiile deschise la închiderea din $d+1$; dacă A_d este vidă, $G_{d+2} = 0$ și nu se plătesc costuri

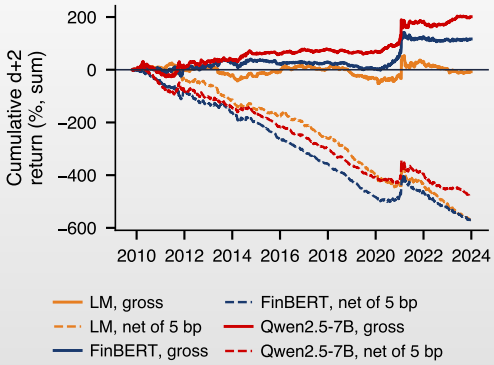
Costurile de tranzacționare: dus-întors și costul de echilibru

- $c = 5$ bp pe tranzacție într-un singur sens, pe unitatea de valoare nominală tranzacționată
  - Nivel presupus, nu estimat, plauzibil pentru acțiuni americane mari și lichide
- O tranzacție dus-întors (intrare și ieșire) costă $2c$
- Într-o zi cu poziții: $4c = 20$ bp
  - Un dus-întors pentru portofoliul de acțiuni și unul pentru acoperirea SPY
  - Fiecare cu valoare nominală 1
- Randamentul net: $G^{\text{net}}_{d+2} = G_{d+2} - 4c\,1\{N_d > 0\}$
- Costul de echilibru c^*: valoarea lui c la care randamentul net mediu este 0
  - Adică $c^* = \bar{G}_{\text{activ}}/4$
  - $\bar{G}_{\text{activ}}$: randamentul brut mediu în zilele cu poziții

Citirea rezultatelor strategiei: raportul Sharpe și randamentele însumate

- Inferența pentru randamentul zilnic mediu
  - t Newey–West (Capitolul 2)
  - CI bootstrap pe blocuri
- Raportul Sharpe $SR = \sqrt{252}\,\bar{G}/\hat{\sigma}$ (anualizat)
  - $\bar{G}$ și $\hat{\sigma}$: media și abaterea standard a randamentului zilnic G_t
  - Pe toate cele 3 568 de zile de tranzacționare, zilele fără poziții intrînd cu 0
- După costuri, G_t se înlocuiește cu randamentul net
  - Exemplu (Qwen2.5-7B): media +5,64 bp și $SR = +0{,}63$ implică $\hat{\sigma} = \sqrt{252}\times$ media$/SR \approx 143$ bp pe zi
  - După costuri: poziții în 95% dintre zile, deci media scade cu $20\times$ 0,95 bp, la $-13{,}36$ bp pe zi
- Graficul următor însumează randamentele zilnice simple în exces
  - O măsură aditivă a profitului, nu averea compusă

Randamentele cumulate ale strategiei: niciun profit net



- Suma randamentelor zilnice în exces (%), înainte de costuri și cu 5 bp pe tranzacție
- FinBERT: +3,27 bp pe zi ($t = +1{,}27$), nu diferă de zero
  - CI 95% [−1,51; +8,93], iar LM obține −0,21 bp pe zi ($t = -0{,}07$)
- Qwen2.5-7B: +5,64 bp ($t = +2{,}26$, $p = 0{,}024$)
  - CI 95% [+0,95; +10,61], dar este cel mai bun dintre cele trei scoruri testate: după corecția Holm pentru cele trei, $p = 0{,}071$, nesemnificativ
  - Costul de echilibru, +1,5 bp pe tranzacție, este mult sub 5 bp: raportul Sharpe este +0,63 înainte de costuri și −1,48 după costuri
- Tipul de dovadă: asociere
  - Doar 2020–2021 are cîștiguri mari, nesemnificative la 5% (Aprofundări)

MFM_ch15_news_signal ▶ Interactiv

Strategia: ziua netranzacționabilă și memorarea

- Ziua netranzacționabilă $d+1$: FinBERT $-0{,}22$ bp ($t=-0{,}10$), Qwen $+4{,}06$ bp ($t=+1{,}36$)
  - Niciun cîștig semnificativ nici aici: aproape toată reacția are loc în ziua d
- O parte din avantajul Qwen poate proveni din memorare: ponderile Qwen2.5 au fost publicate pe 19 septembrie 2024, după titlurile din 2010–2023
  - Un titlu care numește episodul poate declanșa amintirea, spre deosebire de testul Apple, care dădea doar data

MFM_ch15_news_signal

Sentimentul ca regresor generat: eroarea clasică

- Întrebarea secțiunii: ce se întîmplă cu panta și cu intervalul?
  - Cînd regresorul este ieșirea unui model, nu tonul adevărat
- Modelul: $r = a + \beta s^* + \varepsilon$, cu tonul adevărat s^*
  - Observăm ieșirea unui model $\hat{s} = s^* + u$
  - Eroare clasică (u independentă de s^* și ε): plim $\hat{b} = \beta\lambda$, cu $\lambda = \text{Var}(s^*)/[\text{Var}(s^*) + \text{Var}(u)] < 1$
  - λ: fiabilitatea scorului, partea din varianța lui care este semnal
  - Testul pentru $\beta = 0$ rămîne valid, dar estimările sînt atenuate spre 0

Scor standardizat: corecția prin corelația cu tonul adevărat

- Scor standardizat z, eroare nediferențială ($E[r \mid s^*, \hat{s}] = E[r \mid s^*]$): $\mathrm{Cov}(r, z) = \beta \, \mathrm{Cov}(s^*, z)$
  - Cum $\mathrm{Var}(z) = 1$, panta este $b_z = \mathrm{Cov}(r, z) = \beta \, \sigma_{s^*} \, \rho(s^*, \hat{s})$
  - Panta pe z = (efectul unei abateri standard a lui s^*) $\times$ $\rho(s^*, \hat{s})$: pentru corecție, împărțim panta la corelația dintre tonul adevărat și scor
  - O etichetă binară nu are eroare clasică: clasificarea greșită reduce panta unei variabile binare cu alt factor (Aigner (1973); ▸ Anexa și Seminarul 15, A1)
- Econometria variabilelor generate de AI
  - Battaglia et al. (2024): corecția distorsiunii, estimare comună
  - Ludwig, Mullainathan & Rambachan (2024): utilizări ale LLM-urilor pentru predicție față de estimare

Atenuarea pantelor estimate

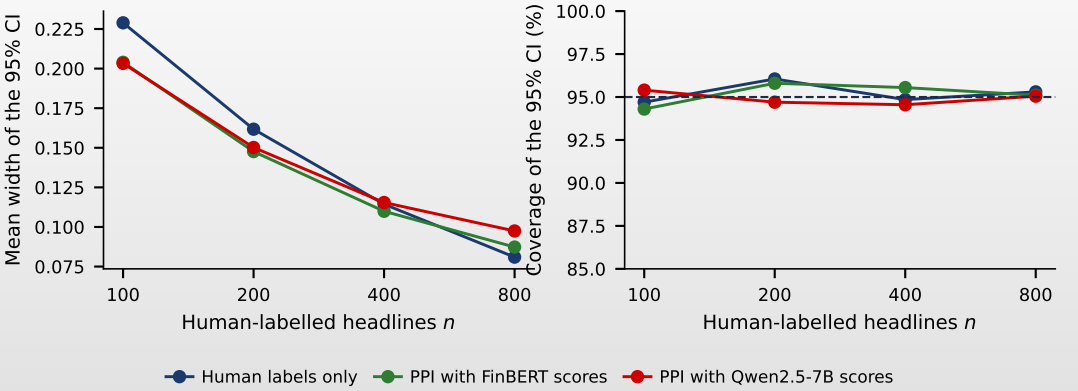
Scor	$\rho(s^*, \hat{s})$, Twitter	b în ziua d (bp)	b/ρ (bp)	ℓ_j, încărcare (un factor)
LM	0,29	15,5	54	0,53
FinBERT	0,58	39,2	67	0,85
Qwen2.5-7B	0,71	41,4	58	0,78

- ρ: corelația scorului net cu eticheta umană $(-1, 0, 1)$, titlu cu titlu
- b/ρ: efectul în aceeași zi al unei abateri standard a tonului adevărat
  - După corecție cele trei metode sînt mult mai apropiate: panta mică a LM este în mare parte eroare de măsurare, dacă ipotezele din Aprofundări sînt îndeplinite
- ℓ_j: încărcarea (*loading*) scorului j pe factorul comun f al celor trei scoruri zilnice
  - $\ell_j = \text{corr}(z_j, f)$
  - Modelul factorial și variabilele instrumentale: Anexa
  - Efectul unei abateri standard a lui f, b/ℓ_j: 29 bp (LM), 46 bp (FinBERT), 53 bp (Qwen)

Inferență validă cu etichete de model: prediction-powered inference

- Datele: n texte etichetate de oameni $(X_j,\ Y_j)$ și N texte neetichetate $\tilde{X}_i$
  - Eșantioane aleatoare independente din aceeași populație
- Ținta: $\theta = E[Y]$, media etichetei umane în populație
  - $f(\cdot)$ este scorul unui model fix, antrenat pe alte date
- PPI (prediction-powered inference, Angelopoulos et al. (2023)): $\hat{\theta}_{\mathbf{PPI}} = \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{j=1}^{n}\left(f(X_j) - Y_j\right)$
  - Primul termen: media predicțiilor pe cele N texte neetichetate
  - Al doilea termen: distorsiunea medie $f - Y$, estimată pe textele etichetate și scăzută din predicție
- $\widehat{\mathrm{Var}}(\hat{\theta}_{\mathbf{PPI}}) = \widehat{\mathrm{Var}}(f)/N + \widehat{\mathrm{Var}}(f - Y)/n$
  - Un CI valid asimptotic sub această schemă de eșantionare, oricare ar fi calitatea lui f

PPI pe titlurile Twitter: corecția distorsiunii este cîștigul principal (1/3)



MFM_ch15_valid_inference ▶ Interactiv

PPI pe titlurile Twitter: corecția distorsiunii este cîștigul principal (2/3)

- Ținta: tonul net uman mediu $(-1, 0, 1)$
- Cele 2 388 de titluri de validare au rolul populației
- 2 000 de replicări cu n titluri etichetate și 2 388 $-n$ neetichetate
  - Extrase independent, cu întoarcere
- Tonul mediu adevărat: $\theta = 0{,}054$
- Mediile simple ale scorurilor modelelor sînt distorsionate
  - FinBERT $-0{,}014$ (semn greșit), Qwen2.5-7B $+0{,}101$
  - Creșterea eșantionului nu elimină această eroare
- PPI păstrează acoperirea nominală: 94,3% la $n = 100$

MFM_ch15_valid_inference

PPI pe titlurile Twitter: corecția distorsiunii este cîștigul principal (3/3)

- Lățimea medie a intervalului la $n = 100$: 0,229 doar cu etichete, 0,204 cu PPI (FinBERT)
- Reducerea variației, $1 - \text{Var}(f - Y)/\text{Var}(Y)$, este modestă
  - 24% pentru FinBERT și 27% pentru Qwen, pentru că modelele sînt zgomotoase
- La $n = 800$ PPI este mai larg (0,087 față de 0,081)
  - Cu doar $2\,388 - 800$ de texte neetichetate, $\text{Var}(f)/N$ nu mai este neglijabil
- Pentru regresiile randamentelor, același eșantion etichetat permite corectarea ambelor condiții de moment OLS (Egami et al. (2023))
- Tipul de dovadă: comparație între proceduri (PPI, doar etichete umane, media scorurilor)

MFM_ch15_valid_inference

Citește un LLM știrile la fel ca piața?

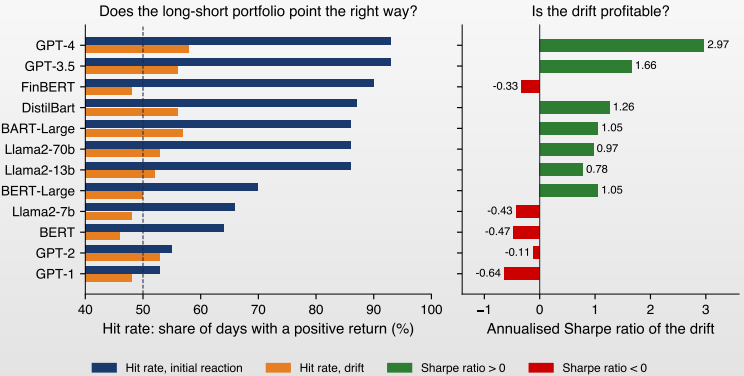
- Lopez-Lira & Tang (2026) (*Journal of Financial Economics*), întrebarea studiului:
  - Poate un LLM de uz general, fără o antrenare financiară explicită, să determine dacă un titlu este o veste bună sau rea pentru acțiune?
- Prognozează răspunsul lui randamentele ulterioare?
- Reacția inițială: mișcarea prețului imediat după știre
  - Nu este tranzacționabilă, dar arată dacă LLM-ul înțelege știrea
- Drift-ul: randamentul de după reacție, în sensul știrii
  - Este tranzacționabil și măsoară reacția insuficientă a pieței
- Față de testul nostru pe FNSPID: aceeași evaluare zero-shot a titlurilor
  - Dar pe știri de după data-limită de antrenare și cu ora exactă a publicării
- Relevanța: LLM-ul devine un instrument de măsură pentru viteza cu care prețurile absorb știrile
  - Modelul teoretic al autorilor: doar LLM-urile care depășesc un anumit prag de acuratețe prognozează randamentele, iar utilizarea lor pe scară largă ar trebui să reducă predictibilitatea

GPT-4 ghicește reacția și câștigă din drift

Model	Rata de succes, reacție	Rata de succes, drift	Drift, % pe zi	*SR*
GPT-4	0,93	0,58	0,34	2,97
GPT-3.5	0,93	0,56	0,29	1,66
FinBERT	0,90	0,48	−0,07	−0,33
GPT-1	0,53	0,48	−0,07	−0,64

- Portofolii long-short pentru știrile dintre ședințe
  - GPT-4 are o rată de succes de 93% pentru reacția inițială, la fel ca GPT-3.5 (versiunea 3,5 a modelelor GPT)
  - GPT-4 are cel mai mare randament ulterior
  - Drift-ul componentei short este 0,26% pe zi (*SR* 2,01), iar al componentei long 0,08% pe zi (*SR* 0,78)
  - Piața reacționează insuficient mai ales la veștile rele
- Regresia randamentului pe scorul GPT-4, știrile dintre ședințe: $\hat{\gamma} = 1{,}326$ ($t = 26{,}52$) pentru reacție
  - 0,161 ($t = 6{,}59$) pentru drift
- Dacă se adaugă scorul de sentiment al furnizorului de știri:
  - Coeficientul GPT-4 rămîne 0,158 ($t = 5{,}27$)
  - Cel al furnizorului este 0,010 ($t = 0{,}15$)
  - Acțiunile mici: coeficientul interacțiunii cu variabila „firmă mică” este 0,404 ($t = 4{,}55$): drift-ul se concentrează acolo unde arbitrajul este costisitor
- Tipul de dovadă: asociere
  - Raportul Sharpe GPT-4 scade de la 6,54 (trimestrul IV 2021) la 1,22 (ianuarie–mai 2024): Aprofundări

Modelele: reacția inițială și randamentul ulterior (1/2)



BART: Bidirectional and Auto-Regressive Transformer, un model de limbaj codificator–decodor; DistilBart: un BART comprimat (distilat); Llama2: LLM-uri cu ponderi deschise publicate de Meta

MFM_ch15_llt_case ▶ Interactiv

Modelele: reacția inițială și randamentul ulterior (2/2)

- Portofolii long-short pentru știrile dintre ședințe, douăsprezece modele (sursa: Lopez-Lira & Tang (2026))
  - Aceeași construcție, octombrie 2021 – mai 2024
- Pragul: GPT-1 (*Generative Pre-trained Transformer 1*, Transformer generativ pre-antrenat, versiunea 1), GPT-2 (primele versiuni GPT), BERT și Llama2-7b ghicesc direcția reacției doar în 53–66% din zile
  - Au rapoarte Sharpe ale drift-ului negative (între $-0{,}64$ și $-0{,}11$)
- Șapte modele ajung la 86–93%
  - Doar GPT-4 (2,97) și GPT-3.5 (1,66) au un raport Sharpe al drift-ului mare; la toate celelalte modele el este cel mult 1,26
- FinBERT ghicește reacția, dar are drift negativ
  - Indică o direcție doar pentru 25% dintre titlurile dintre ședințe
- Concluzia: informația deja reflectată în prețul de deschidere nu aduce cîștiguri ulterioare
- Numărul mediu de acțiuni din componenta short
  - 2 pentru BERT-Large (Sharpe 1,05), 6 pentru GPT-3.5, 20 pentru GPT-4
- Un raport Sharpe obținut din cîteva poziții short este fragil

MFM_ch15_llt_case

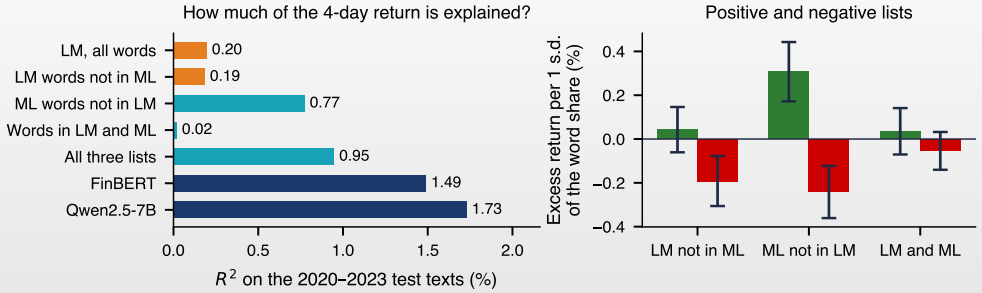
Replicare în acest curs, limite și o extindere deschisă

- Strategia Qwen2.5-7B din acest capitol, pe 16 acțiuni FNSPID
  - +5,64 bp pe zi, față de 34 bp pentru GPT-4
  - Replicarea (Seminarul 15, C3): prompt-ul publicat, temperatura 0, scorurile $\{-1, 0, 1\}$, regula celor două firme (o componentă doar cu cel puțin două firme) și aceeași regresie, pe titlurile FNSPID și prețurile de închidere ajustate din datele cursului
  - Limitele datelor noastre: date fără oră (primul randament sigur este cel din $d + 2$), doar 16 acțiuni mari, titluri mai vechi decît modelul
- Limite enunțate de autori: la 20 bp pe tranzacție dus-întors strategia pierde bani
  - Deci se potrivește celor cu costuri mici, de exemplu unui formator de piață (*market maker*)
  - Costul nostru de 5 bp într-un singur sens înseamnă 10 bp dus-întors
  - Prompt-urile și embeddings nu au fost optimizate, deci comparațiile sînt limite inferioare
- Extindere deschisă, pe titluri publicate după lansarea unui model cu ponderi deschise
  - se mai păstrează drift-ul?
  - continuă să scadă raportul Sharpe, așa cum prezice modelul teoretic al autorilor?

Al doilea studiu de caz: culoarea cuvintelor financiare

- Garcia, Hu & Rohrer (2023) (*Journal of Financial Economics*), întrebarea studiului:
  - Poate o listă de cuvinte aleasă de reacția prețurilor să concureze cu listele alese de cercetători, de exemplu Loughran–McDonald (LM)?
- Corpusul lucrării: transcrierile a 85 530 de conferințe de prezentare a rezultatelor trimestriale (*earnings calls*), 3 229 de firme, 13.10.2005–07.10.2020
- Eticheta fiecărui text: randamentul acțiunii minus randamentul pieței, de la închiderea din ziua $t-1$ la închiderea din ziua $t+2$
- Rezultatul principal al lucrării: liste învățate pe apelurile dinainte de 2016, evaluate pe 2016–2020, cu variabilele de control ale lucrării
  - R^2 ajustat: 2,1% cu cuvintele LM care nu sînt în listele ML (*Machine Learning*, învățare automată), 4,6% cu cuvintele ML care nu sînt în LM
  - Doar cu variabilele de control: 1,7%
  - Listele învățate pe tot eșantionul lucrării au în comun cu LM doar 18 cuvinte pozitive și 30 negative
- Relevanța: lista rămîne transparentă, dar eticheta cuvintelor vine din piață, nu din judecata cercetătorului
  - Este reperul simplu cu care trebuie comparate FinBERT și LLM-urile pe aceleași texte

Listele învățate din randamente, pe titlurile FNSPID



ML (machine learning): listele învățate din randamente; listele sînt construite pe 2009–2019, R^2 și pantele sînt estimate pe 2020–2023

MFM_ch15_colour_words ▶ Interactiv

Metoda: robust MNIR și ce schimbă replicarea din curs

- MNIR (multinomial inverse regression, Taddy (2013), Anexă): numărul de apariții ale cuvîntului j în textul i este Poisson, cu media $m_i e^{\alpha_j + \varphi_j v_i}$
  - m_i: numărul de cuvinte al textului; v_i: randamentul standardizat
- Pasul robust: MNIR pe 500 de subeșantioane de cîte 5 000 de texte, extrase fără întoarcere
  - D_j^+: procentul subeșantioanelor cu $\varphi_j > 0$ minus procentul celor cu $\varphi_j < 0$
  - Lista pozitivă: $D_j^+ \geq 80\%$; lista negativă: $D_j^+ \leq -80\%$
- Scorul unui text: numărul de cuvinte din listă împărțit la numărul de cuvinte, scalat la abaterea standard 1
- Replicarea din curs, pe datele capitolului:
  - Textul: toate titlurile FNSPID ale unei acțiuni într-o zi, 18 282 de perechi acțiune–zi
  - Învățare: oct. 2009 – dec. 2019, fără textele a căror fereastră de randament ajunge în 2020 (10 286 de texte); testare: 2020–2023 (7 980 de texte)
  - Vocabularul (15 376 de cuvinte) și pragurile de winsorizare: doar din textele de învățare
  - Semnul lui φ_j: aproximarea prin testul scorului (*score test*) a selecției AIC (criteriul informațional Akaike, $z^2 > 2$), în locul penalizării gamma-lasso
  - Fără variabilele de control ale lucrării; numele celor 16 firme sînt eliminate în locul filtrului de nume proprii

Rezultatul: cuvintele alese de piață explică mai mult decît LM

- R^2 pe textele de test, 2020–2023, cu pantele estimate pe aceeași perioadă: 0,20% cu toate cuvintele LM, 0,77% cu cele 94 de cuvinte ML care nu sînt în LM
  - Diferența: +0,58 puncte procentuale, CI 95% [0,14; 1,02] (bootstrap pe blocuri de 10 zile consecutive, 1 999 de extrageri, cu listele învățate fixate)
  - Pentru pragurile lui D_j^+ între 50% și 95% (în loc de 80%), R^2 al listelor ML este între 0,44% și 1,02%
- O abatere standard în plus a ponderii cuvintelor ML (t cu erori standard Driscoll–Kraay, HAC pe zile)
  - Pozitive: +0,31% ($t = 4{,}45$); negative: −0,24% ($t = -3{,}98$)
- Coeficientul cuvintelor LM pozitive nu diferă semnificativ de zero: +0,04% ($t = 0{,}81$)
  - Cuvintele LM negative: −0,19% ($t = -3{,}29$), semnificative și alături de listele ML: −0,17% ($t = -2{,}90$)
- FinBERT (1,49%) și Qwen2.5-7B (1,73%) explică mai mult decît orice listă
  - Qwen față de listele ML: +0,96 puncte procentuale, CI 95% [0,20; 1,73]
- Tipul de dovadă: comparație între proceduri, pe o asociere din aceeași fereastră de 4 zile, nu o prognoză
  - Multe cuvinte ML descriu chiar mișcarea prețului („jumped”, „slumped”, „gainers”, „losers”): titlul poate urma randamentul
- Limitele: doar 1 cuvînt pozitiv și 5 negative comune cu LM (positive; claims, downgrades, weak, worse, worst)
  - Qwen poate ști deja ce a urmat știrilor din 2020–2023 (Deplasarea de anticipare)

O întrebare deschisă: putem avea încredere în ipotezele generate de LLM-uri?

- Ludwig & Mullainathan (2024): generarea ipotezelor ca procedură statistică
  - Un model antrenat pe date descoperă un tipar, oamenii îl numesc, date noi îl testează
- Lopez-Lira, Tang & Zhu (2025): înainte de data-limită a cunoștințelor, un LLM își poate aminti rezultatele
  - Deci o ipoteză propusă de el despre acea perioadă nu este independentă de datele care o vor testa
- Întrebarea deschisă: cînd un asistent propune semnalul, prompt-ul, pragul și perioada de deținere
  - care este familia de teste?
  - ce inferență rămîne validă?
- Agenții care rulează singuri experimentele (Lu et al. (2024)) fac familia mai mare și mai puțin vizibilă

Familia de teste: FWER și FDR

- Formal: $H_{\mathbf{0},j} : E[r_t^{(j)}] = 0$ pentru strategiile candidate $j = 1, \ldots, M$
  - Cu controlul FWER (Holm) sau al FDR (Benjamini & Hochberg (1995))
  - FDR (false discovery rate): media proporției respingerilor false printre toate respingerile
  - BH (procedura Benjamini–Hochberg) la nivelul q: respingem ipotezele cu $p_{(\mathbf{1})}, \ldots, p_{(k^*)}$, unde $k^* = \max\{k : p_{(k)} \leq kq/M\}$
  - M numără fiecare variantă propusă de asistent, inclusiv pe cele abandonate
  - Dacă ipotezele au fost construite pe rezultatele de test, p-value-urile cer date de test neatinse sau inferență selectivă (secvențială)

Bucla descoperirii și verificările cercetătorului

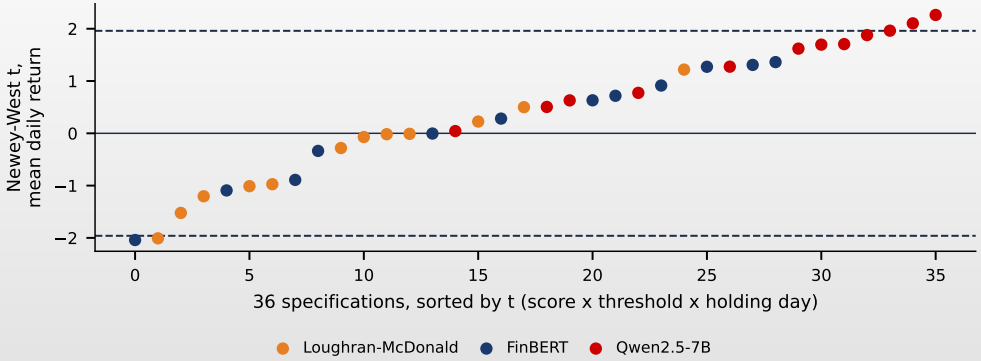
- Literatura, ipotezele, datele și codul, robustețea fixată dinainte, critica, raportul
  - Un asistent (Claude, ChatGPT, Gemini, GitHub Copilot) redactează fiecare pas, cercetătorul îl verifică
  - Prompt-uri reutilizabile: ▸ Anexa
- **Memorarea**: modelul știe ce a urmat după titlu
  - Teste: știri de după publicarea ponderilor, text anonimizat ca diagnostic (Glasserman & Lin (2023)) și testul de memorie din acest capitol
- **Drumurile ramificate** (*forking paths*): prompt, prag, orizont ajustate pe perioada de test
  - Test: o grilă preînregistrată și o rată de eroare declarată: Holm (FWER) sau BH (FDR, sub independență sau dependență pozitivă)
  - $|t| > 3$ (Harvey, Liu & Zhu (2016)) este o regulă empirică, nu o garanție pentru rata de eroare
- **HARKing**: un mecanism plauzibil scris după rezultat, cu referințe care pot să nu existe
  - HARKing: hypothesising after the results are known, formularea ipotezei după ce rezultatul este cunoscut
  - Test: formulați mecanismul și ce l-ar infirma înaintea testului și verificați fiecare DOI (digital object identifier, identificatorul digital al lucrării)
- **Alinierea în timp și agenții care rulează cod**
  - Marcaje de timp doar la nivel de dată și modificări neanunțate ale datelor, ferestrelor sau seed-urilor
  - Teste: primul randament tranzacționabil este cel din $d + 2$, date doar pentru citire, apeluri de instrumente jurnalizate și verificarea modificărilor (diff) fiecărui script înainte de a accepta un rezultat

Mini-studiu de caz: rafinările propuse de un asistent

- O propunere tipică a unui asistent:
  - „semnalul Qwen funcționează; rafinează-l cu un prag mai strict sau cu o zi de intrare mai tîrzie"
- Testăm toată grila pe scorurile salvate
- Grila: 3 scoruri $\times$ 3 praguri $\times$ 4 zile de randament $=$ 36 de strategii
  - Scorurile: LM, FinBERT, Qwen
  - Pragurile: $|s_{i,d}| > \delta$, $\delta \in \{0; 0{,}25; 0{,}5\}$, pe scorul brut
  - Zilele: $d+2-d+5$, fiecare o poziție de o zi deschisă la închiderea precedentă
  - Cazul de bază: 5,64 bp pe zi, $t = 2{,}26$
- $|t| > 1{,}96$ pentru 5 din 36 de strategii (2 cu semnul greșit)
  - Față de 1,8 așteptate sub ipoteza nulă globală la 5%
- Numărul de strategii care rămîn semnificative:
  - 0 cu Holm (FWER 5%), 0 cu BH (FDR 10%), 0 cu regula $|t| > 3$
- Pragul Bonferroni este $|t| > 3{,}20$
- Interpretare: rafinările nu adaugă dovezi
  - Cea mai bună variantă este cea de la care am pornit
  - Aleasă ea însăși dintre trei scoruri, pe o perioadă pe care modelul și-o poate aminti

MFM_ch15_ai_discovery

Curba specificațiilor față de zgomot (1/2)



MFM_ch15_valid_inference ▶ Interactiv

Curba specificațiilor față de zgomot (2/2)

- Toate cele 36 de strategii din mini-studiu, ordonate după t (Simonsohn, Simmons & Nelson (2020)), cu linii punctate la $\pm 1{,}96$
- Ipoteza nulă comună: nicio specificație nu prognozează randamentele
  - Schimbăm aleator semnul pozițiilor în blocuri de 10 zile de tranzacționare consecutive, inclusiv zilele cu randament zero, de 2 000 de ori, cu aceleași schimbări pentru toate cele 36 de specificații (dependent wild bootstrap, Shao (2010))
  - Valid dacă, sub ipoteza nulă, randamentele cu semn sînt simetrice în jurul lui 0, iar blocurile sînt aproximativ independente
- Observat: 3 specificații cu $t > 1{,}96$, iar t median este 0,50
  - Sub ipoteza nulă: probabilitatea a 3 sau mai multe specificații cu $t > 1{,}96$ este 0,113, iar a unei mediane a lui t cel puțin la fel de mari, 0,207
- Interpretare: curba este compatibilă cu zgomotul la 5%
  - Raportarea doar a celui mai bun punct ar ascunde asta
- Tipul de dovadă: asociere
  - Niciuna dintre cele 36 de asocieri predictive nu se distinge de zgomot

MFM_ch15_valid_inference

Temă de proiect

Întrebarea și pașii

- Prognozează sentimentul titlurilor, măsurat cu un LLM, randamentele după $d + 1$?
  - Pe știri publicate după lansarea modelului, cu o grilă preînregistrată
- Replicați întîi strategia Qwen din acest capitol
- Apoi aplicați protocolul cu știri de după data-limită din Lopez-Lira & Tang (2026)
- Notați în `AI_USE.md` fiecare prompt care a influențat o decizie
- Notați în `AI_ERRORS.md` fiecare referință inventată, fapt memorat sau cifră greșită depistată
  - Împreună cu testul care a depistat-o
  - Șabloanele: folderul proiectului

Livrabile

- Grila și corecția pentru testarea multiplă, preînregistrate
- Titluri datate de după publicarea unui model deschis
- Testul pe toate variantele grilei
- Un mecanism propus de LLM, formulat înaintea testului împreună cu ce l-ar infirma
- O critică de tip recenzent și răspunsul echipei

Idei principale

- Dicționarele generale citesc greșit textele financiare, iar dicționarul LM rămîne un reper transparent
- Între FinBERT și LLM-urile deschise de 7B–14B nu s-a detectat o diferență semnificativă de acuratețe
  - Pe titluri financiare
  - Ceea ce nu dovedește egalitatea
- Cîteva sute de etichete și un clasificator simplu concurează cu ambele
- LLM-urile își amintesc istoria piețelor: evaluați-le doar după data publicării lor
- Sentimentul titlurilor explică mișcările din aceeași zi
  - Dar, cu o aliniere realistă în timp și cu costuri, nu depășește piața
  - Pentru cele 16 acțiuni mari studiate
- Scorurile sînt măsurători cu eroare
  - Atenuarea, erorile posibil comune ale modelelor, puținele grupuri și numeroasele încercări influențează inferența

Verificarea înțelegerii

- De ce poate o metodă care adună etichetele de dicționar ale cuvintelor izolate să clasifice greșit o expresie financiară?
  - Dați un exemplu în care contextul schimbă semnul.
- O acuratețe mare a clasificării dovedește profitabilitatea tranzacționării?
  - Separați calitatea etichetelor, momentul tranzacției și costurile.
- De ce contează data lansării modelului într-un backtest pe știri istorice?
  - Descrieți alegerea datelor care exclude ca randamentele evaluate să fie memorate în ponderi.
- De ce este scorul de sentiment o măsurătoare, nu un adevăr observat?
  - Explicați ce permite să estimați un eșantion aleator de titluri etichetate de oameni.

Răspunsuri scurte: Anexa.

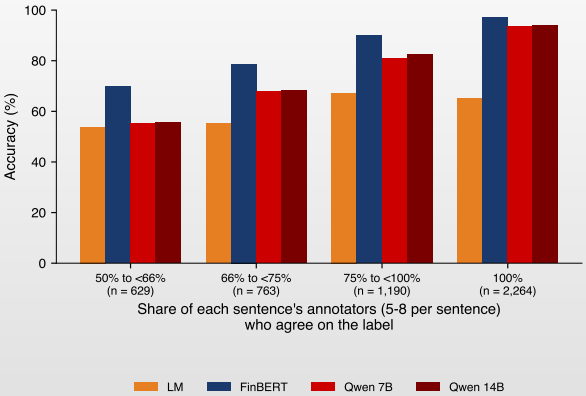
Decizia practică

- **Ce aleg**
  - Un reper rapid și transparent: dicționarul LM, fără antrenare și verificabil
  - Titluri financiare scurte în engleză: FinBERT sau, cu cîteva sute de etichete din domeniu, embeddings + regresie logistică
  - Fără etichete, în alte limbi sau pe documente lungi: un LLM deschis de cel puțin 7B, zero-shot
  - Pentru tranzacționarea zilnică a celor 16 acțiuni mari: niciun semnal de sentiment
- **În ce condiții**
  - Evaluare în afara eșantionului, cu intervale și teste pe perechi
  - Pe Twitter, TF-IDF + regresie logistică are 83,2% acuratețe, față de 74,2% pentru Qwen2.5-7B și 72,5% pentru FinBERT
  - Strategia Qwen2.5-7B are un cost de echilibru de 1,5 bp pe tranzacție, sub costul presupus de 5 bp
- **Ce rezultat m-ar face să renunț**
  - Un test McNemar semnificativ în favoarea altei metode, pe texte din domeniul-țintă
  - Un test de memorie cu AUC peste 0,5 pe perioada backtest-ului (observat aici: 0,69 pentru Qwen2.5-14B, pe lunile dinainte de publicare)
  - Un semnal care rămîne semnificativ după Holm pe titluri de după publicarea modelului, cu un cost de echilibru peste costul real

Exemplu rezolvat: o propoziție, două dicționare

- Propoziție din Financial PhraseBank, etichetată **pozitiv** de toți adnotatorii:
  - *Uponor improved its performance in spite of the decrease in residential construction in the US .* (US: Statele Unite)
- Dicționarul LM găsește un cuvînt pozitiv („improved") și niciun cuvînt negativ
  - $\tau = (1-0)/(1+0) = +1{,}00$: ton pozitiv, deci clasificare corectă
- Dicționarul GI nu găsește niciun cuvînt pozitiv și găsește două cuvinte negative („spite", „decrease")
  - $\tau = (0-2)/(0+2) = -1{,}00$: ton negativ, deci clasificare greșită
  - „Spite" și „decrease" descriu piața, nu firma, iar „improved" lipsește din GI

Dezacordul adnotatorilor și acuratețea modelelor



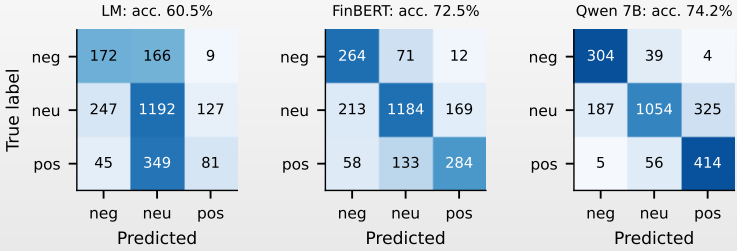
- Financial PhraseBank împărțit după cel mai înalt prag de acord atins de fiecare propoziție
  - Grupul cel mai slab: între 50% și 66%
- Acuratețea la acord deplin: FinBERT 97%, Qwen2.5-14B 94%, LM 65%
  - Acuratețea la un acord de cel puțin 50%, dar sub 66%: FinBERT 70%, Qwen2.5-14B 55%, LM 54%
- Eticheta umană este ea însăși zgomotoasă: limitează superior acuratețea măsurată
  - Modele cu etichete latente: Anexa
  - Un zgomot simetric și independent de clasificatori atenuează diferențele spre 0; unul dependent de clasă le poate mări sau inversa
- Pentru semnale: un ton puternic și clar este mai informativ decît unul la limită

MFM_ch15_classification

Acord peste hazard: κ al lui Cohen

- Acuratețea numără și acordul obținut din întîmplare, mare cînd o clasă domină
- κ al lui Cohen (Cohen (1960)): $\kappa = \dfrac{p_o - p_e}{1 - p_e}$
  - p_o: ponderea textelor pe care doi evaluatori (aici modelul și omul) dau aceeași etichetă
  - $p_e = \sum_k p_k^A \, p_k^B$: acordul așteptat din întîmplare, din frecvențele marginale ale etichetelor celor doi evaluatori
  - $\kappa = 1$: acord perfect; $\kappa = 0$: nu mai bun decît hazardul
- Pe setul de validare Twitter: $\kappa = 0{,}19$ (LM), 0,50 (FinBERT), 0,56 (Qwen2.5-7B)
- Reperul „mereu neutru" are $\kappa = 0$, deși acuratețea lui o depășește pe cea a LM
- Pentru mai mulți evaluatori pe text, κ al lui Fleiss (Anexa)

Structura erorilor: modele diferite, greșeli diferite (1/2)



- Setul de validare Twitter, numărul de titluri în fiecare celulă
  - Rîndurile sînt etichetele adevărate, coloanele etichetele prezise

MFM_ch15_classification

Structura erorilor: modele diferite, greșeli diferite (2/2)

- LM clasifică drept neutre majoritatea titlurilor pozitive
  - Lista sa de cuvinte pozitive este scurtă
- FinBERT greșește în ambele direcții în jurul clasei neutre
- Qwen2.5-7B: recall de circa 87% pe pozitive (414/475) și 88% pe negative (304/347)
  - Dar precizie de doar 56% pe pozitive (414/743) și 61% pe negative (304/496)
  - Multe titluri neutre devin pozitive sau negative
- Pentru tranzacționare, semnalele false generează pierderi
  - Contează precizia pe clasele extreme

MFM_ch15_classification

Același model, altă formulare: prompt-ul este un parametru



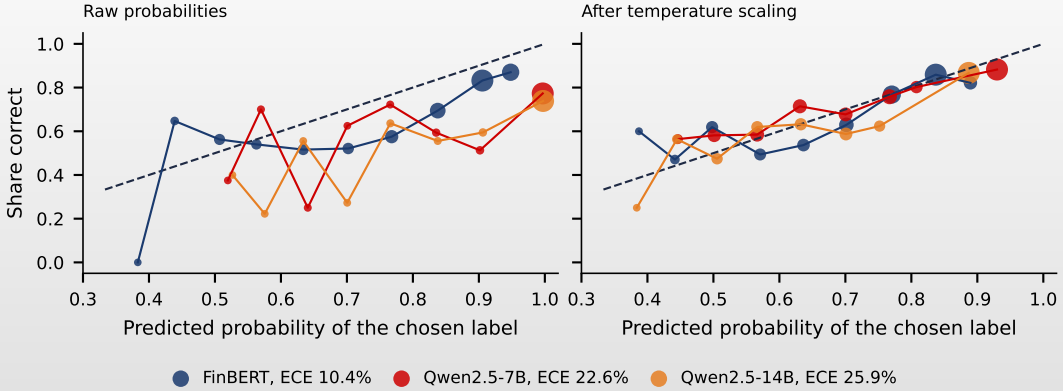
- Trei prompt-uri care pun aceeași întrebare, pe setul de validare Twitter
- Acuratețea variază cu prompt-ul
  - Între 52,7% și 61,7% pentru 1.5B
  - Între 73,7% și 77,8% pentru 7B
  - Între 69,9% și 75,6% pentru 14B
  - Cele trei prompt-uri dau aceeași etichetă doar pentru 67% dintre titluri (7B)
  - Ponderea răspunsurilor „neutral” pentru 7B: de la 48% (P1) la 69% (P3)
- Alegerea prompt-ului după examinarea rezultatelor de test este o formă de căutare repetată în aceleași date (*data snooping*) (Capitolul 13)
  - Fixați prompt-ul dinainte sau alegeți-l pe un set de validare separat

MFM_ch15_llm_zero_shot

Calibrarea probabilităților: ECE și scalarea prin temperatură

- Întrebarea: cînd modelul dă unei clase probabilitatea 0,9, are dreptate în 90% din cazuri?
- ECE $= \sum_b \frac{n_b}{n} |\text{acc}_b - \text{conf}_b|$ pe 10 intervale ale probabilității maxime
  - ECE: expected calibration error, eroarea medie de calibrare
  - n_b și n: numărul de titluri din intervalul b, respectiv numărul total de titluri
  - conf_b: media probabilității maxime în intervalul b
  - acc_b: proporția titlurilor clasificate corect în intervalul b
- Scalarea prin temperatură (Guo et al. (2017)): $p_k(T) \propto p_k^{1/T}$, adică logiții împărțiți la T
  - T estimat prin verosimilitate maximă pe o jumătate aleatoare a setului de validare Twitter
  - ECE măsurat pe cealaltă jumătate
  - Probabilitățile salvate sînt rotunjite la 6 zecimale; zerourile sînt înlocuite numeric cu 10^{-12}
  - Clasa cea mai probabilă nu se schimbă, deci nici acuratețea

Calibrarea probabilităților: LLM-urile sînt supraîncrezătoare (1/3)



MFM_ch15_valid_inference

Calibrarea probabilităților: LLM-urile sînt supraîncrezătoare (2/3)

- Panoul din stînga: probabilitățile brute
- Panoul din dreapta: după scalarea prin temperatură
  - Diagonala înseamnă calibrare perfectă
- ECE pentru probabilitățile brute: FinBERT 10,4%, Qwen2.5-7B 22,6%, Qwen2.5-14B 25,9%
- Qwen2.5-7B atribuie pentru 88% dintre titluri o probabilitate de peste 0,93 (în medie 99,7%)
  - Are dreptate doar la 77% dintre ele: supraîncredere
- Scalarea prin temperatură, $T = 8{,}4$ (7B) și 1,4 (FinBERT): ECE scade la 4,3% și 4,6%
  - T pentru 7B depinde de pragul numeric: 5,6 cu 10^{-6} în locul lui 10^{-12}; supraîncrederea rămîne
- ECE pentru 7B depinde de prompt: 22,6% cu P1, 22,1% cu P2 și 19,8% cu P3

MFM_ch15_valid_inference

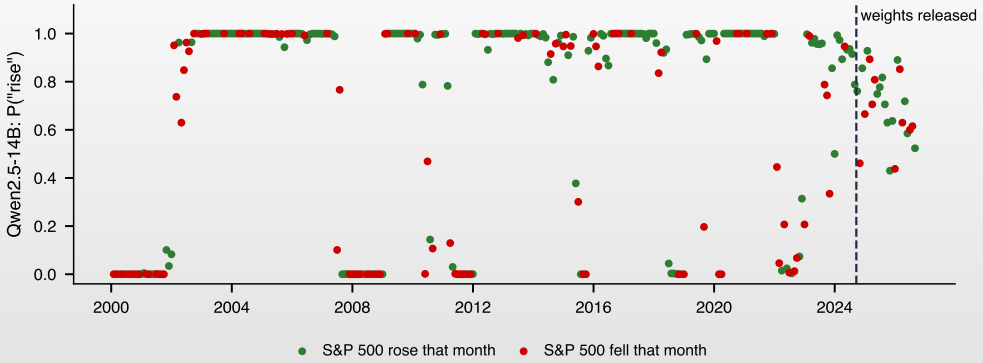
Calibrarea probabilităților: LLM-urile sînt supraîncrezătoare (3/3)

- Două reguli de scor proprii pentru probabilități, la care o valoare mai mică este mai bună
  - Proper scoring rules
  - log loss: media lui $-\ln p_{iy_i}$, cu p_{iy_i} probabilitatea dată clasei adevărate y_i a titlului i
  - scorul Brier (Brier (1950)): distanța pătratică dintre probabilități și eticheta codificată one-hot (1 pentru clasa adevărată, 0 în rest), însumată pe cele 3 clase
- Pentru 7B, după scalarea prin temperatură:
  - Log loss trece de la 2,61 la 0,63
  - Scorul Brier trece de la 0,468 la 0,366
- Cu probabilitățile brute ale LLM-ului, scorul net folosit într-o regresie
  - Se comportă aproape ca o etichetă cu trei valori

Exemplu rezolvat: LLM față de FinBERT

- Twitter, prompt-ul P1: Qwen2.5-7B 74,2%, FinBERT 72,5%
  - A = FinBERT, B = Qwen; texte pe care doar unul dintre modele are dreptate: $n_{01} = 395$, $n_{10} = 355$
  - Diferența de acuratețe: $(n_{01} - n_{10})/n = 40/2\,388 = +1{,}7$ puncte procentuale, CI 95% $[-0{,}5;\ +3{,}8]$
  - De mînă: $\chi^2 = (|n_{01} - n_{10}| - 1)^2/(n_{01} + n_{10}) = 2{,}03$, deci $p = P(\chi^2_1 > 2{,}03) \approx 0{,}15$; testul exact dă $p = 0{,}154$
- Qwen2.5-14B față de FinBERT: diferența este de $-0{,}1$ puncte procentuale, cu $p = 0{,}942$
- De 70 pînă la 130 de ori mai mulți parametri nu aduc un cîștig demonstrat
  - Nici egalitatea nu este dovedită
- Unde ajută LLM-urile (Anexa)
  - Nu cer etichete și funcționează în orice limbă
  - Pe documente mai lungi și mai complexe

Memoria lui Qwen2.5-14B despre S&P 500 (1/2)



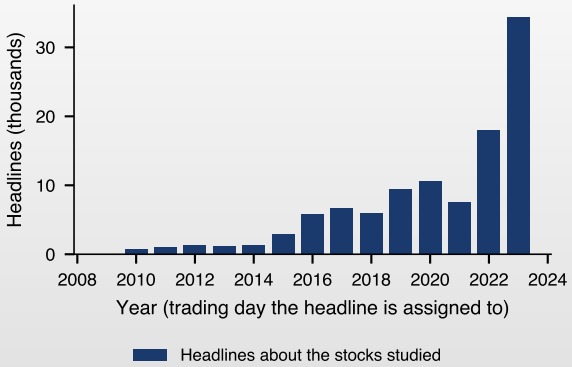
MFM_ch15_lookahead

Memoria lui Qwen2.5-14B despre S&P 500 (2/2)

- Probabilitatea răspunsului „rise” pentru fiecare lună, fără context
  - Culoarea arată ce a făcut de fapt indicele
- Răspunsul este „fall” pentru 2000–2002 și 2008–2009
  - „Rise” pentru cea mai mare parte din 2003–2007 și 2013–2021
  - $P(\text{creștere}) \approx 0$ în lunile cu scăderi mari: octombrie 2008 ($-16{,}9\%$), martie 2020 ($-12{,}5\%$), decembrie 2018 ($-9{,}2\%$)
- Dar septembrie 2002 ($-11{,}0\%$) și februarie 2009 ($-11{,}0\%$) primesc $P(\text{creștere})$ aproape de 1
  - Cunoașterea este aproximativă
  - Modelul reține episoadele memorabile, nu fiecare lună
- După publicare, răspunsurile tind spre „rise”
  - Fără cunoștințe, rămîne convingerea generală că piețele cresc

MFM_ch15_lookahead

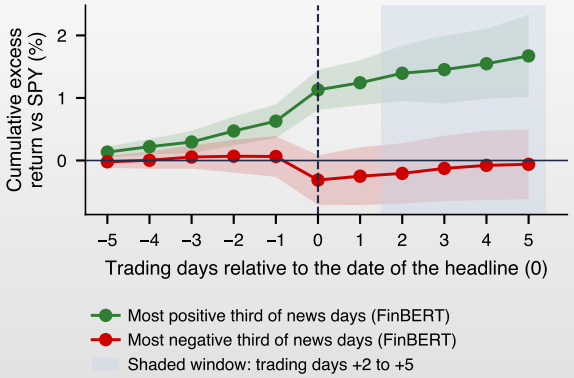
Numărul de titluri în timp: acoperire inegală



- Titluri despre cele 16 acțiuni pe an, după ziua de tranzacționare la care sînt atribuite
- 34 380 de titluri doar în 2023, față de 8 445 în toți anii dinainte de 2016
  - Rezultatele sînt dominate de anii recenți, iar primii ani au puține zile cu știri pe acțiune
- Într-o zi cu știri, o acțiune are în medie 5,8 titluri
- Scorul zilnic: media scorurilor titlurilor pentru acțiunea și ziua respectivă
- Deplasarea de supraviețuire (*survivorship bias*)
  - Cele 16 acțiuni aparțin unor firme mari în prezent, selectate retrospectiv

MFM_ch15_news_data

Studiu de eveniment: mișcarea are loc înainte de știre și în ziua ei (1/2)



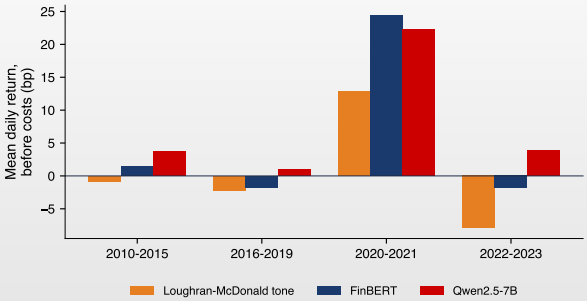
- Randamentul în exces cumulat mediu, de la ziua -5 la ziua $+5$
  - Zilele cu știri împărțite pe tercilele scorului zilnic FinBERT
  - Zona hașurată: zilele tranzacționabile, de la $d+2$ la $d+5$
- Treimea cu scorurile cele mai pozitive: randament în exces cumulat
  - $+0{,}63\%$ în cele 5 zile dinainte
  - $+0{,}50\%$ în ziua d
  - Treimea cu scorurile cele mai negative: $+0{,}06\%$ în cele 5 zile dinainte și $-0{,}38\%$ în ziua d
  - Explicații posibile: scurgeri de informații sau titluri scrise după mișcarea prețului

MFM_ch15_news_signal

Studiu de eveniment: mișcarea are loc înainte de știre și în ziua ei (2/2)

- Efectul sentimentului: diferența dintre treimea pozitivă și cea negativă
  - Estimandul: diferența mediilor, cu ponderi egale pe evenimentele acțiune–zi; CI Driscoll–Kraay (HAC pe zile)
  - În ziua d: $+1{,}00\%$, CI $[+0{,}87;\ +1{,}13]$; în zilele $d+2$–$d+5$: $+0{,}17\%$, CI $[-0{,}11;\ +0{,}46]$
  - Graficul face media întîi în fiecare zi, apoi între zilele fiecărei treimi: $+0{,}88\%$ în ziua d
- Testul robust la varianța indusă de eveniment (Boehmer, Musumeci & Poulsen (1991)), pentru zilele $d+2$–$d+5$: $z = \dfrac{\overline{SCAR}}{s(SCAR)/\sqrt{N}}$
  - $SCAR_e = CAR_e/\hat{\sigma}_e$: randamentul în exces cumulat al evenimentului e, împărțit la abaterea lui standard estimată înaintea evenimentului; $N = 6\,094$ evenimente în fiecare treime
  - $z = +2{,}62$ pentru treimea pozitivă și $+1{,}25$ pentru cea negativă dacă evenimentele sînt independente; cu erori standard Driscoll–Kraay: $+1{,}57$ și $+0{,}66$
- Ambele treimi au randamente cumulate în creștere
  - Poate contribui selecția firmelor care sînt mari astăzi

Stabilitatea în timp: dovezi slabe



- Randamentul zilnic mediu al strategiei înainte de costuri, pe subperioade
- FinBERT: +1,4 bp (2010–2015), −1,8 bp (2016–2019), +24,3 bp (2020–2021), −1,8 bp (2022–2023)
  - Doar subperioada 2020–2021 are valori mari, nesemnificative la 5%: $t = +1{,}89$ pentru FinBERT, iar Qwen are +22,2 bp ($t = +1{,}78$)
  - 2020–2021: perioada acțiunilor-meme (episodul GameStop: Anexa)
  - După costurile presupuse, estimările punctuale rămîn pozitive, dar dovezile statistice sînt slabe
  - Pe întregul eșantion, rezultatul net este negativ
- Explicații: 16 acțiuni foarte urmărite, marcaje de timp doar la nivel de dată, frecvență zilnică
  - Predictibilitatea publicată se bazează pe marcaje de timp precise și pe multe acțiuni mici (studiul de caz)

MFM_ch15_news_signal

Cînd exagerează corecția? Titluri față de zile-acțiune

- Corecția b/ρ se bazează pe trei ipoteze
  - Eroare nediferențială: dat fiind tonul adevărat, eroarea scorului nu prezice randamentele
  - Corelația de pe Twitter se păstrează pentru titlurile FNSPID
  - Aceeași unitate: ρ este măsurat pe titlu, dar b folosește scorul zilnic al unei acțiuni, adică media titlurilor din acea zi (în medie 5,8 titluri)
- Dacă erorile titlurilor sînt independente, iar titlurile aceleiași zile-acțiune au același ton adevărat:
  - Media crește corelația cu adevărul
  - ρ zilnic depășește ρ pe titlu, iar b/ρ corectează prea mult
- Exemplu numeric: un titlu are fiabilitatea $\lambda = \rho^2 = 0{,}25$ ($\rho = 0{,}50$)
  - Media a $k = 4$ astfel de titluri are $\lambda_k = \dfrac{k\lambda}{1 + (k-1)\lambda} = \dfrac{4(0{,}25)}{1 + 3(0{,}25)} = 0{,}57$ (Spearman–Brown)
  - Deci $\rho_k = 0{,}76$
  - O pantă zilnică de 10 bp este corectată atunci la $10/0{,}50 = 20$ bp cu ρ pe titlu, dar la $10/0{,}76 \approx 13$ bp cu ρ zilnic
- Cu erori independente între titluri, valorile b/ρ sînt limite superioare
  - Concluzia despre panta LM depinde de această ipoteză
- Remediu: etichetarea tuturor titlurilor dintr-un eșantion aleator de zile-acțiune
  - Măsoară ρ la nivelul potrivit

Cînd ajută PPI și cum se extinde la regresii

- Intervalul PPI este mai îngust decît cel calculat doar cu etichete cînd
  - $\mathrm{Var}(f)/N + \mathrm{Var}(f - Y)/n < \mathrm{Var}(Y)/n$
  - Aproximativ $\mathrm{Var}(f - Y) < \mathrm{Var}(Y)$ cînd $N \gg n$: modelul trebuie să explice o parte din variația etichetei
- Pentru o regresie, corecția se aplică fiecărei condiții de moment care conține variabila prezisă
  - Și atunci cînd aceasta este regresor (Egami et al. (2023))
- Egami et al. (2023): învățare supervizată bazată pe design
  - O versiune dublu robustă care permite probabilități de etichetare dependente de text
- Cele cîteva sute de documente etichetate nu sînt doar date de antrenare
  - Ele sînt eșantionul de validare care face validă inferența ulterioară

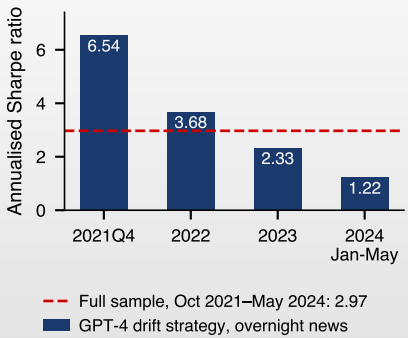
Cum evaluează autorii știrile publicate după data-limită de antrenare a GPT-4

- 159 137 de observații firmă-titlu-dată, octombrie 2021 – mai 2024
  - 4 123 de acțiuni ordinare listate la bursele americane NYSE (*New York Stock Exchange*, Bursa din New York), NASDAQ (*National Association of Securities Dealers Automated Quotations*) și AMEX (*American Stock Exchange*, Bursa Americană de Valori)
  - Eșantionul începe după septembrie 2021, data-limită a datelor de antrenare pentru modelul `gpt-4-0314`
  - Filtre: scorul de relevanță al furnizorului de știri egal cu 100, fără titlurile de tip „stock-gain"/„stock-loss", doar evenimente noi, eliminarea titlurilor aproape identice din aceeași zi
  - Știrile publicate înainte de ora 9 sau după ora 16 reprezintă 82% dintre titluri, iar restul apar în timpul ședinței de tranzacționare
- Prompt-ul: modelul joacă rolul unui expert financiar
  - Răspunde `YES` (veste bună), `NO` (veste rea) sau `UNKNOWN`
  - Despre efectul pe termen scurt asupra prețului acțiunii
  - Temperatura 0
  - Scorul $s_{i,t}$: `YES` $= 1$, `UNKNOWN` $= 0$, `NO` $= -1$

Portofoliile și regresia din Lopez-Lira și Tang

- Portofolii zilnice cu ponderi egale
  - Poziții long pentru scoruri pozitive și short pentru scoruri negative
  - Fiecare componentă (long sau short) se formează doar dacă are cel puțin două firme
  - Pentru știrile dintre ședințe, reacția inițială se măsoară de la închiderea anterioară la deschidere, iar drift-ul, de la deschidere la închidere
  - Rata de succes: ponderea zilelor cu randament pozitiv al portofoliului
  - Raportul Sharpe: anualizat din randamentele zilnice ale portofoliului, ca pentru strategia noastră
- Regresia: $r_{i,t} = a_i + b_t + \gamma\, s_{i,t} + \varepsilon_{i,t}$
  - Efecte fixe de firmă și de dată a_i, b_t
  - Erori standard grupate pe date și pe firme
  - γ este randamentul pe o unitate de scor
  - Douăsprezece modele de limbaj de mărimi crescătoare, evaluate cu aceeași metodologie, de la GPT-1 la GPT-4

Drift-ul scade pe măsură ce LLM-urile se răspîndesc



- Strategia long-short GPT-4 pe drift-ul știrilor dintre ședințe (sursa: Lopez-Lira & Tang (2026))
  - Raportul Sharpe anualizat pe subperioade
- De la 6,54 în trimestrul IV din 2021 la 1,22 în ianuarie–mai 2024, o scădere de peste cinci ori
  - Valoarea de 2,97 pe tot eșantionul combină un început puternic cu un final slab: nu este o prognoză a rezultatelor actuale
- Concordant cu erodarea prezisă de modelul autorilor, iar momentul coincide cu răspîndirea ChatGPT
- Autorii menționează și schimbarea condițiilor de piață
  - Subperioade scurte: intervalele de 95% sînt largi, iar cel pentru 2024 îl conține pe 0
  - Știrile din timpul ședinței: drift-ul de la închidere la închidere nu mai este semnificativ în 2024

MFM_ch15_llt_case

Dezacordul adnotatorilor: acord și etichete latente

- κ **al lui Fleiss** (Fleiss (1971)): N texte, m evaluatori, n_{ij} evaluatori care dau textului i eticheta j
  - $\bar{P} = \frac{1}{N}\sum_i \sum_j \frac{n_{ij}(n_{ij}-1)}{m(m-1)}$, $\bar{P}_e = \sum_j p_j^2$, $\kappa = \dfrac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$, cu p_j ponderea tuturor voturilor acordate etichetei j
  - Formula presupune același m pentru fiecare text
  - PhraseBank are 5–8 evaluatori pe propoziție și publică doar nivelurile de acord, nu voturile: κ nu poate fi calculat din datele publicate
- **Modelul Dawid–Skene** (Dawid & Skene (1979))
  - Evaluatorul a are o matrice de confuzie $\pi^{(a)}_{jl} = P(\text{spune } l \mid \text{adevărat } j)$
  - Clasa adevărată este latentă (neobservată)
  - Algoritmul EM (expectation–maximization) estimează $\pi^{(a)}$ și dă fiecărui text o etichetă probabilistă $P(\text{adevărat} = j \mid \text{voturi})$
- Evaluați modelele cu o regulă de scor proprie, nu prin acurateță față de eticheta majoritară
  - De exemplu, pierderea logaritmică față de etichetele probabiliste
  - Plafonul din graficul acordului este un plafon al zgomotului etichetelor, nu un eșec al modelului

Înapoi

Exemplu rezolvat: FinBERT față de LM

- Setul de validare Twitter, 2 388 de titluri: LM 60,5%, FinBERT 72,5%
  - LM greșește și FinBERT are dreptate: $n_{01} = 539$
  - LM are dreptate și FinBERT greșește: $n_{10} = 252$
  - Testul McNemar exact: $p < 0{,}001$
- Diferența: +12,0 puncte procentuale, CI 95% bootstrap [+9,8; +14,2]
  - Avantaj clar al modelului cu fine-tuning, chiar și în afara datelor lui de antrenare

Înapoi

Timpul de procesare al modelelor Qwen2.5

- Timpul de procesare pe titlu, pe procesorul grafic al unui laptop, inclusiv prompt-ul
  - Mediana unei treceri, pentru fiecare prompt
  - Timpul median pe titlu: 4 ms (0.5B), 12 ms (1.5B), 20 ms (3B), 92 ms (7B), 105 ms (14B)
  - O singură trecere este zgomotoasă: 7B a durat între 51 și 94 ms, iar 14B între 88 și 112 ms
  - O afirmație despre viteză cere rulări repetate

Înapoi

Dincolo de titluri: documente lungi

- Rapoarte anuale: Loughran & McDonald (2011) și-au construit listele din formulare 10-K
  - Tonul și incertitudinea anunță randamentele și volatilitatea în jurul publicării
- Conferințe despre rezultate și rapoarte ale analiștilor (Huang, Wang & Yang (2023))
  - FinBERT extrage informație pe care dicționarele nu o surprind
- Kim, Muhn & Nikolaev (2023): rezumatele făcute de LLM ale raportărilor sînt mult mai scurte
  - Dar explică mai bine reacțiile pieței decît originalele
  - Ei măsoară lungimea excesivă a raportărilor: textul lung și puțin informativ este asociat cu o eficiență mai mică a prețurilor
- Documentele lungi cer modele cu ferestre de context mari
  - Sau segmentarea textului și agregarea rezultatelor

Înapoi

Regresii pe text cu fundamentare statistică

- Taddy (2013): regresia inversă multinomială
  - Modelăm frecvențele cuvintelor condiționat de rezultat, $c_d \sim \text{MN}(q(y_d), m_d)$
  - MN: distribuția multinomială, c_d: frecvențele, $m_d = \sum_j c_{dj}$: lungimea
  - Model log-liniar $q_j(y) \propto e^{\alpha_j + \varphi_j^\top y}$, încărcările $\Phi = (\varphi_1, \ldots, \varphi_p)^\top$: dat fiind m_d și acest model, $z_d = \Phi^\top c_d / m_d$ este suficient pentru y_d
- Ke, Kelly & Xiu (2019) (SESTM, sentiment extraction via screening and topic modelling)
  - Selectăm cuvintele corelate cu randamentele
  - Estimăm un model cu două teme (cuvinte pozitive și negative)
  - Atribuim scoruri articolelor noi prin verosimilitate penalizată
  - Lucrarea dă garanții statistice pentru estimator, iar semnalul este cel mai puternic în ziua știrii și în ziua următoare
- Bybee et al. (2024): atenția acordată temelor în 800 000 de articole Wall Street Journal
  - Urmărește activitatea economică și prognozează randamentele agregate ale pieței
  - Vezi și Jegadeesh & Wu (2013) (ponderi ale cuvintelor estimate din reacțiile pieței la 10-K)
  - Regula comună: estimarea se face doar pe date anterioare perioadei de prognoză

Înapoi

Dispariția memoriei: un test formal și puterea lui

- H_0 : $\text{AUC}_{\textbf{înainte}} = \text{AUC}_{\textbf{după}}$, două eșantioane independente, cu varianța DeLong pentru fiecare AUC (DeLong et al. (1988))
  - $z = (\widehat{\text{AUC}}_1 - \widehat{\text{AUC}}_2)/\sqrt{\hat{V}_1 + \hat{V}_2}$: pentru 14B, $z = 0{,}73$ ($p = 0{,}47$); pentru 7B, $z = 1{,}01$ ($p = 0{,}31$)
- Sub H_0 (AUC $= 0{,}5$, fără valori egale): $SE_0 = \sqrt{\dfrac{n_1 + n_0 + 1}{12\, n_1 n_0}}$
  - n_1, n_0: numărul lunilor de creștere, respectiv de scădere
  - După publicare: $n_1 = 13$, $n_0 = 10$, deci $SE_0 = \sqrt{24/1\,560} = 0{,}124$
- Cea mai mică scădere detectabilă (test bilateral la 5%, putere 80%):
  - $2{,}80\sqrt{SE^2_{\textbf{înainte}} + SE_0^2} = 2{,}80\sqrt{0{,}03227^2 + 0{,}124^2} = 0{,}36$
  - $2{,}80 = 1{,}96 + 0{,}84$, cuantilele normale pentru test și pentru putere; $SE_{\textbf{înainte}}$: eroarea standard DeLong pentru cele 296 de luni dinainte de publicare
- O cădere de la 0,69 la 0,5 (0,19) este sub acest prag
  - Nerespingerea lui H_0 nu arată că memoria persistă după publicare
- Lunile necesare: cel mai mic m pentru care scăderea detectabilă coboară la 0,19
  - Cu ponderea lunilor de creștere de după publicare păstrată la 13/23
  - $m = 101$ (7B: 129)

înapoi

Etichete discrete: clasificarea greșită nu este eroare clasică

- Variabila binară „știre negativă" $s^* \in \{0, 1\}$ și eticheta prezisă $\hat{s}$
  - $\alpha_0 = P(\hat{s} = 1 \mid s^* = 0)$ și $\alpha_1 = P(\hat{s} = 0 \mid s^* = 1)$: ratele fals-pozitivelor și fals-negativelor
  - Eroarea $u = \hat{s} - s^*$ este corelată negativ cu s^*: formula clasică nu se aplică (Aigner (1973))
  - plim $\hat{b} = \beta \dfrac{\pi(1-\pi)(1-\alpha_0-\alpha_1)}{p(1-p)}$, $\pi = P(s^* = 1)$, $p = P(\hat{s} = 1)$
- Pe setul de validare Twitter, din matricile de confuzie: $\pi = 14{,}5\%$
  - Factorul de atenuare a pantei: 0,45 pentru FinBERT ($\alpha_0 = 13{,}3\%$, $\alpha_1 = 23{,}9\%$), 0,59 pentru Qwen2.5-7B și 0,28 pentru LM ($\alpha_1 = 50{,}4\%$)
- Cît timp $\alpha_0 + \alpha_1 < 1$ semnul se păstrează
  - Dar panta unei variabile binare se poate reduce la o treime sau mai puțin

Înapoi

Două măsuri zgomotoase: modelul factorial și variabilele instrumentale

- Un factor: $z_j = \ell_j f + u_j$, $\mathrm{Var}(f) = 1$, u_j necorelate cu f și între ele
  - $\rho_{jk} = \ell_j \ell_k$, deci $\ell_j^2 = \rho_{jk}\rho_{jl}/\rho_{kl}$
  - Încărcările ℓ_j estimate pe scorurile zilnice FNSPID: 0,53 (LM), 0,85 (FinBERT), 0,78 (Qwen)
- IV (instrumental variables, variabile instrumentale), cu $r = a + \beta f + \varepsilon$
  - Instrumentăm z_Q cu z_F: $\hat{b}_{IV} = \widehat{\mathrm{Cov}}(r, z_F)/\widehat{\mathrm{Cov}}(z_Q, z_F) \to \beta/\ell_Q$
  - Condiții: $\ell_F \neq 0$, f este necorelat cu ε, iar u_F este necorelată cu u_Q și cu ε
  - Panta Qwen în ziua d: 59 bp [53; 66] cu FinBERT ca instrument și 37 bp [28; 46] cu LM ca instrument (intervale bootstrap pe zile)
- Dacă ambele instrumente ar fi valide, estimările ar coincide
  - Diferența este de 22 bp, cu $z = 5{,}4$, deci egalitatea este respinsă
  - Testul respinge restricțiile comune, nu un instrument anume; cauze plauzibile: erori FinBERT–Qwen corelate prin texte de antrenare care se suprapun sau o eroare a LM legată de randamente
  - În acest caz, și încărcările ℓ ale modelului factorial devin îndoielnice

Înapoi

Rețele sociale și comportament de turmă: GameStop, 2021

- Bollen, Mao & Zeng (2011): starea de spirit colectivă de pe Twitter a anticipat Dow Jones în 2008
  - Rezultatul este sensibil la eșantion și la metodă
- Ianuarie 2021: investitori individuali s-au coordonat pe rețelele sociale
  - Au ridicat GameStop de peste zece ori în cîteva săptămîni
  - Atenția, sentimentul și tranzacțiile au crescut împreună
  - Acest episod singur nu poate separa predicția de cauzalitate: fiecare cere propriul test
- Lecții pentru semnalele textuale
  - Volumul atenției poate conta mai mult decît tonul
  - Semnalele urmate de prea mulți investitori (crowding) se inversează și creează risc pentru vînzătorii în lipsă (Capitolul 17: bule)



Un magazin GameStop, Vallejo, California, februarie 2021

Foto: Will Buckner (2021); CC BY 2.0; Wikimedia Commons

Înapoi

Bucla descoperirii, prompt-uri reutilizabile (1/2)

- **Literatura**:
  - `"Găsește studii din 2023 încoace care prognozează randamentele acțiunilor cu sentimentul din LLM-uri și testează după data-limită a cunoștințelor modelului: DOI sau id arXiv, data-limită, fereastra de test, estimarea principală."`
  - Căutați cu Semantic Scholar, Google Scholar sau Elicit și verificați fiecare identificator (DOI sau id arXiv); Lopez-Lira & Tang (2026) este designul de referință
- **Ipoteze**:
  - `"Propune mecanisme prin care sentimentul titlurilor ar putea prognoza randamentele după d+1 și, pentru fiecare, tiparul transversal care l-ar infirma."`
  - Reacția insuficientă prezice un drift concentrat în acțiunile mici și puțin urmărite
  - Memorarea prezice un avantaj care dispare după data publicării
- **Date și cod**:
  - `"Din scorurile zilnice ale știrilor din acest capitol și închiderile zilnice ale cursului, reprodu strategia zilnică Qwen din acest capitol (intrare la închiderea din d+1, acoperire cu SPY) și t-ul ei Newey-West."`
  - Scorurile sînt salvate, deci niciun model nu trebuie rulat din nou

Înapoi

Bucla descoperirii, prompt-uri reutilizabile (2/2)

- **Robustețea**, fixată dinainte:
  - `"Înainte de orice rulare nouă, scrie grila completă de variante pe care le-ai testa și corecția pentru testarea multiplă; toate vor fi raportate."`
  - Verificări: placebo (scoruri amestecate între acțiunile din aceeași zi), subperioade și costuri de 5 bp pe tranzacție
- **Critica**:
  - `"Scrie un raport de recenzent: toate felurile în care strategia ar putea reflecta deplasarea de anticipare, memorarea, selecția sau marcajele de timp, cu testul pentru fiecare."`
  - Apoi verificați direct fiecare test propus: și o critică poate conține rezultate inventate
- **Raportul**: toată grila, varianta aleasă înaintea rezultatelor
  - Versiunile modelelor și datele publicării

Înapoi

Agenți AI în tranzacționare: utilizări și riscuri

- Un agent este un LLM într-o buclă: citește intrările, decide, apelează instrumente, observă, repetă
  - Instrumente: date, cod, căutarea știrilor, ordine printr-un API (application programming interface)
  - Folosit azi pentru rezumate de cercetare, programare și monitorizare
  - Tranzacționarea complet autonomă este încă rară și experimentală
- Manipulare: un text citit de agent poate conține instrucțiuni (prompt injection)
  - De exemplu, un comunicat de presă fals
- Risc sistemic (Fondul Monetar Internațional, IMF (2024))
  - Multe firme cu modele asemănătoare pot tranzacționa la fel
  - Crescînd viteza și corelația mișcărilor în perioade de stres
  - Controale: limite de poziție, aprobarea umană a ordinelor, jurnalizarea fiecărui pas, teste după fiecare actualizare a modelului

Înapoi

Lista erorilor frecvente

- **Suprapunerea cu datele de antrenare**: un model pare mai bun decît este
  - Evaluat pe propriile date de antrenare (FinBERT pe PhraseBank)
- **Deplasarea de anticipare**: un LLM își poate aminti piața de după știre
  - Testați după data publicării lui
- **Alinierea în timp**: marcajele de timp conțin doar data
  - Tranzacționați cel mai devreme la închiderea de după ce știrea este sigur publică
- **Căutarea repetată în aceleași date**: prompt-uri, praguri și modele alese după examinarea rezultatelor de test
- **Costuri și crowding**: cîștigurile zilnice mici dispar
  - Odată ce scădem costuri de cîteva puncte de bază
- **Regresori generați**: scorurile sînt estimări
  - Corectați atenuarea și validați pe un eșantion aleator etichetat de oameni (PPI)

Înapoi

Răspunsuri scurte la întrebările de verificare (1/2)

- **Dicționarul**: o metodă care adună etichetele cuvintelor izolate dă fiecărui cuvînt o singură etichetă
  - În „lower costs" aplică eticheta negativă a lui „lower" fără să interpreteze expresia
  - În contextul obișnuit al profitabilității firmei, costurile mai mici pot fi favorabile
  - „Capital" sau „board" sînt substantive neutre din limbajul de afaceri, pe care GI le numără ca negative
- **Acuratețe și profit**: acuratețea se măsoară față de etichete umane zgomotoase, pe alte texte
  - O tranzacție cere primul randament sigur ($d + 2$)
  - Trebuie să acopere $4c$ pe zi și să reziste corecției pentru numeroasele încercări
  - Aici toate cele trei scoruri explică mișcările din aceeași zi, dar nicio strategie nu aduce profit net după costuri

Înapoi

Răspunsuri scurte la întrebările de verificare (2/2)

- **Data publicării**: înainte de ea, randamentele evaluate pot fi deja memorate în ponderi
  - Știrile și randamentele de după publicarea ponderilor nu puteau fi memorate în ele
  - Condiția: fără acces la informații viitoare prin instrumente
  - Celelalte surse de deplasare de anticipare din backtest rămîn de controlat
- **Testul de memorie fără context**: poate semnala memorarea
  - Qwen2.5-14B: AUC 0,69 pe lunile dinainte de publicare
  - Un AUC după publicare nesemnificativ (0,59 pe 23 de luni) nu dovedește absența contaminării
- **Scorul ca măsurătoare**: $\hat{s} = s^* + u$ estimează tonul adevărat
  - Chiar o eroare clasică, de medie zero, poate atenua panta (în modelul clasic, panta pe z = efectul $\times$ ρ)
  - Un eșantion aleator etichetat de oameni permite estimarea distorsiunii medii $f - Y$ și a fiabilității scorului
  - Estimările sînt valabile pentru unitatea de analiză la care a fost măsurat eșantionul
  - PPI corectează media scorurilor cu media diferenței $f - Y$
  - Corecția b/ρ cere ipotezele modelului și un ρ măsurat la aceeași unitate de agregare ca regresia
  - Un ρ măsurat pe titluri nu justifică automat corectarea unei pante pe medii zilnice

Înapoi

Analiza textului: lecturi și aplicații

- Tetlock (2007): pesimismul din presă prezice randamentele pieței
- Loughran & McDonald (2011): un dicționar de sentiment specific finanțelor
- Gentzkow, Kelly & Taddy (2019): textul ca date în economie
- Aplicații pe criptoactive:
  - Clasificarea criptoactivelor Pele et al. (2023)
  - Bulele Dogecoin și sentimentul Pele (2025)
  - Quantlet de sentiment Pele (2024)
- Studiu suplimentar: Quantinar: Natural Language Processing, Quantinar: DELTA, Quantinar: LDA Extensions
- Manualul Franke, Härdle & Hafner (2019) pentru statistica din spatele testelor

Înapoi

Acronime folosite în acest material (1/3)

- **AI**: Artificial Intelligence — *inteligență artificială*
- **AIC**: Akaike Information Criterion — *criteriul informațional Akaike*
- **AMEX**: American Stock Exchange — *Bursa Americană de Valori*
- **API**: Application Programming Interface — *interfață de programare a aplicațiilor*
- **AUC**: Area Under the (ROC) Curve — *aria de sub curba ROC*
- **BART**: Bidirectional and Auto-Regressive Transformer — *model de limbaj codificator–decodor*
- **BERT**: Bidirectional Encoder Representations from Transformers — *reprezentări bidirecționale produse de codificatorul unui Transformer*
- **BH**: Benjamini–Hochberg (procedure) — *procedura Benjamini–Hochberg*
- **BY-NC**: Creative Commons Attribution-NonCommercial — *licența Creative Commons: atribuire, utilizare necomercială*
- **CI**: Confidence Interval — *interval de încredere*
- **CLS**: CLaSsification token (BERT-type models) — *tokenul special de clasificare al modelelor de tip BERT*
- **DOI**: Digital Object Identifier — *identificatorul digital al unei lucrări*
- **ECE**: Expected Calibration Error — *eroarea medie de calibrare*
- **EM**: Expectation-Maximisation (algorithm) — *algoritmul expectation–maximization: pasul E (media condiționată) și pasul M (maximizarea)*

Acronime folosite în acest material (2/3)

- **FDR**: False Discovery Rate — *rata descoperirilor false*
- **FNSPID**: Financial News and Stock Price Integration Dataset — *setul de date care integrează știri financiare și prețuri de acțiuni*
- **FWER**: Family-Wise Error Rate — *probabilitatea de cel puțin o eroare de tip I în familia de teste*
- **GI**: (Harvard) General Inquirer, dictionary IV-4 — *dicționarul Harvard General Inquirer IV-4*
- **GPT**: Generative Pre-trained Transformer — *Transformer generativ pre-antrenat*
- **HARKing**: Hypothesising After the Results are Known — *formularea ipotezelor după ce se cunosc rezultatele*
- **IMF**: International Monetary Fund — *Fondul Monetar Internațional*
- **IV**: Instrumental Variables — *variabile instrumentale*
- **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*
- **MIT**: Massachusetts Institute of Technology — *Institutul de Tehnologie din Massachusetts*
- **ML**: Machine Learning — *învățare automată*
- **MN**: MultiNomial (distribution) — *distribuția multinomială*
- **MNIR**: MultiNomial Inverse Regression — *regresia inversă multinomială*

Acronime folosite în acest material (3/3)

- **NASDAQ**: National Association of Securities Dealers Automated Quotations — *bursa americană NASDAQ (cotații automate ale Asociației Naționale a Dealerilor de Valori Mobiliare)*
- **NYSE**: New York Stock Exchange — *Bursa din New York*
- **OLS**: Ordinary Least Squares — *metoda celor mai mici pătrate*
- **PCA**: Principal Component Analysis — *analiza componentelor principale*
- **PPI**: Prediction-Powered Inference — *inferență susținută de predicții*
- **RLHF**: Reinforcement Learning from Human Feedback — *învățare prin recompensă din evaluări umane*
- **ROC**: Receiver Operating Characteristic (curve) — *curba caracteristică de operare a receptorului*
- **SE**: Standard Error — *eroare standard*
- **SESTM**: Sentiment Extraction via Screening and Topic Modelling — *extragerea sentimentului prin selecție și modelarea temelor*
- **TF-IDF**: Term Frequency - Inverse Document Frequency — *frecvența termenului înmulțită cu inversul frecvenței în documente*
- **US**: United States — *Statele Unite ale Americii*
- **UTC**: Coordinated Universal Time (Temps Universel Coordonné) — *timpul universal coordonat*
- **VADER**: Valence Aware Dictionary and sEntiment Reasoner — *dicționar și reguli pentru valența și sentimentul textelor*

Bibliografie (1/7)

- 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.
- Antweiler, W., Frank, M.Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, 59(3), 1259–1294.
- Araci, D. (2019). FinBERT: financial sentiment analysis with pre-trained language models. arXiv:1908.10063.
- Baker, M., Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4), 1645–1680.
- Battaglia, L., Christensen, T., Hansen, S., Sacher, S. (2024). Inference for regression with variables generated by AI or machine learning. arXiv:2402.15585.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1), 289–300.
- Boehmer, E., Musumeci, J., Poulsen, A.B. (1991). Event-study methodology under conditions of event-induced variance. *Journal of Financial Economics*, 30(2), 253–272.
- Bollen, J., Mao, H., Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.

Bibliografie (2/7)

- Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language models are few-shot learners. *NeurIPS*; arXiv:2005.14165.
- Bybee, L., Kelly, B., Manela, A., Xiu, D. (2024). Business news and business cycles. *Journal of Finance*, 79(5), 3105–3147.
- 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.
- Cameron, A.C., Gelbach, J.B., Miller, D.L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2), 238–249.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- 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.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: pre-training of deep bidirectional Transformers for language understanding. *NAACL*; arXiv:1810.04805.
- 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.

Bibliografie (3/7)

- Drinkall, F., Rahimikia, E., Pierrehumbert, J.B., Zohren, S. (2024). Time Machine GPT. arXiv:2404.18543.
- Driscoll, J.C., Kraay, A.C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, 80(4), 549–560.
- Egami, N., Hinck, M., Stewart, B.M., Wei, H. (2023). Using imperfect surrogates for downstream inference: design-based supervised learning for social science applications of large language models. arXiv:2306.04746.
- Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Franke, J., Härdle, W.K., Hafner, C.M. (2019). Statistics of Financial Markets: An Introduction, 5th ed. Springer.
- Garcia, D., Hu, X., Rohrer, M. (2023). The colour of finance words. *Journal of Financial Economics*, 147(3), 525–549.
- Gentzkow, M., Kelly, B., Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574.
- 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.

Bibliografie (4/7)

- Harvey, C.R., Liu, Y., Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.
- He, S., Lv, L., Manela, A., Wu, J. (2025). Chronologically consistent large language models. arXiv:2502.21206.
- Huang, A.H., Wang, H., Yang, Y. (2023). FinBERT: a large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2), 806–841.
- Hutto, C., Gilbert, E. (2014). VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225.
- IMF (2024). Advances in artificial intelligence: implications for capital market activities. *Global Financial Stability Report*, October 2024, Chapter 3.
- Jegadeesh, N., Wu, D. (2013). Word power: a new approach for content analysis. *Journal of Financial Economics*, 110(3), 712–729.
- Ke, Z.T., Kelly, B., Xiu, D. (2019). Predicting returns with text data. NBER Working Paper 26186.
- Kearney, C., Liu, S. (2014). Textual sentiment in finance: a survey of methods and models. *International Review of Financial Analysis*, 33, 171–185.
- Kim, A., Muhn, M., Nikolaev, V. (2023). Bloated disclosures: can ChatGPT help investors process information? arXiv:2306.10224.

Bibliografie (5/7)

- Lopez-Lira, A., Tang, Y. (2026). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Journal of Financial Economics*, 184, 104335.
- Lopez-Lira, A., Tang, Y., Zhu, M. (2025). The memorization problem: can we trust LLMs' economic forecasts? arXiv:2504.14765.
- 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.
- Loughran, T., McDonald, B. (2016). Textual analysis in accounting and finance: a survey. *Journal of Accounting Research*, 54(4), 1187–1230.
- Lu, C., Lu, C., Lange, R.T., Foerster, J., Clune, J., Ha, D. (2024). The AI Scientist: towards fully automated open-ended scientific discovery. arXiv:2408.06292.
- Ludwig, J., Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *Quarterly Journal of Economics*, 139(2), 751–827.
- Ludwig, J., Mullainathan, S., Rambachan, A. (2024). Large language models: an applied econometric framework. arXiv:2412.07031.
- 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.
- Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Bibliografie (6/7)

- Ouyang, L., Wu, J., Jiang, X., Almeida, D., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*; arXiv:2203.02155.
- Pele, D.T. (2024). Sentiment Analysis (Quantlet, BitMood project). IDA Quantlets Repository, Bucharest University of Economic Studies.
- Pele, D.T. (2025). Dogecoin and speculative bubbles: a study using LPPL and sentiment analysis. Quantinar course, Bucharest University of Economic Studies.
- Pele, D.T., Wesselhöfft, N., Härdle, W.K., Kolossiatis, M., Yatracos, Y.G. (2023). Are cryptos becoming alternative assets? *The European Journal of Finance*, 29(10), 1064–1105.
- 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.
- Reimers, N., Gurevych, I. (2019). Sentence-BERT: sentence embeddings using Siamese BERT-networks. *EMNLP*; arXiv:1908.10084.
- Sarkar, S.K., Vafa, K. (2024). Lookahead bias in pretrained language models. SSRN Working Paper 4754678.
- Shannon, C.E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1), 50–64.

Bibliografie (7/7)

- Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489), 218–235.
- Shiller, R.J. (2017). Narrative economics. *American Economic Review*, 107(4), 967–1004.
- Simonsohn, U., Simmons, J.P., Nelson, L.D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214.
- Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503), 755–770.
- Tetlock, P.C. (2007). Giving content to investor sentiment: the role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168.
- Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S. (2008). More than words: quantifying language to measure firms' fundamentals. *Journal of Finance*, 63(3), 1437–1467.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention is all you need. *NeurIPS*; arXiv:1706.03762.
- 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.
- Wu, S., Irsoy, O., Lu, S., Dabravolski, V., et al. (2023). BloombergGPT: a large language model for finance. arXiv:2303.17564.