

Rețele neuronale și deep learning

Cursul 6: Bazele NLP: reprezentarea textului și reprezentarea vectorială a cuvintelor



Prof. dr. Daniel Traian PELE

Academia de Studii Economice din București (ASE), Departamentul de Statistică și Econometrie
Academia Română, Institutul de Prognoză Economică (IPE)
IDA Institute Digital Assets ▪ AI4EFin ▪ MSCA Digital Finance

Obiectivele cursului

- Să transformați textul financiar în numere pentru prelucrarea limbajului natural (NLP): unități lexicale, vocabular, sac de cuvinte și frecvența termenului–frecvența inversă în documente (TF-IDF).
- Să scrieți funcțiile-obiectiv ale word2vec și GloVe și să le antrenați pe un corpus real de discursuri ale băncii centrale.
- Să construiți clasificatoare de fraze din reprezentări rare și dense, cu modele de referință și teste corecte.
- Să măsurați sentimentul și temele financiare în știri și tweet-uri și să știți cînd un model liniar este suficient.

Abrevieri

AI	inteligență artificială
ASE	Academia de Studii Economice din București
BCE	Banca Centrală Europeană
BERT	reprezentări bidirecționale ale codificatorului transformer (BERT)
BoW	sac de cuvinte (bag of words)
CBOW	sac continuu de cuvinte (varianta word2vec)
CPU	procesor central
ES	pierderea medie așteptată peste VaR (expected shortfall)
ESG	mediu, social și guvernanță
FinBERT	model BERT adaptat textelor financiare
IPE	Institutul de Prognoză Economică, Academia Română
LLM	model lingvistic mare
MLP	perceptron multistrat
NLP	prelucrarea limbajului natural
OOV	în afara vocabularului
PCA	analiza componentelor principale
PMI	informația mutuală punctuală
SGNS	skip-gram cu eșantionare negativă
SVD	descompunerea în valori singulare
TF-IDF	frecvența termenului–frecvența inversă în documente
VaR	valoarea expusă la risc (value-at-risk)

Datele: trei colecții reale de texte

Corpusul	Conținut	Dimensiune	Utilizare aici
Financial PhraseBank	fraze din știri financiare despre companii listate în Finlanda, etichetate pozitiv / neutru / negativ de 5–8 adnotatori	4 838 de fraze	sentiment
Twitter financial news	tweet-uri financiare: pesimist / optimist / neutru și 20 de teme	9 543 + 2 388 (sentiment), 16 990 + 4 117 (teme)	sentiment, teme
Discursurile Băncii Centrale Europene (BCE)	textele integrale ale membrilor Comitetului executiv și ai Consiliului de supraveghere	400 de discursuri, 1 289 896 de unități lexicale	propriii noștri vectori de cuvinte

- Discursuri din perioada 8 ianuarie 2022 – 30 octombrie 2026; PhraseBank de Malo et al. (2014); tweet-urile așa cum sînt publicate pe Hugging Face (seturile de antrenare și de validare).
- Fiecare model este antrenat pe procesorul central (CPU) al unui laptop, cu sămînța 20261108; seturile de test sînt folosite o singură dată.

(data/SOURCES.md)

Cuprins

- Textul ca date
- Reprezentări rare: sacul de cuvinte și TF-IDF
- Reprezentări dense: vectorii de cuvinte (embeddings)
- Clasificatoare de fraze: reprezentări rare și dense
- Studiu de caz: temele tweet-urilor financiare

Împărțirea unei fraze în unități lexicale

- **O frază reală din PhraseBank**
  - "Finnish Aktia Group 's operating profit rose to EUR 17.5 mn in the first quarter of 2010 from EUR 8.2 mn in the first quarter of 2009 ."
  - Eticheta dată de adnotatori: **pozitiv**
- **Tokenizatorul nostru**
  - Litere mici, cuvinte și numere; fiecare număr devine `<num>`; legăturile și mențiunile cu @ sînt eliminate
  - Primele unități: `finnish aktia group s operating profit rose to eur <num> mn in the first` (27 în total)
- **Alegeri care schimbă rezultatele**
  - Trecerea la litere mici unește "Bank" și "bank"; reducerea la rădăcină unește "rose" și "rising"
  - Eliminarea cuvintelor de legătură șterge "not", "down", "up": dăunător pentru sentiment
- **În practică:** documentați preprocesarea și păstrați-o identică în antrenare și în producție
  - Un alt tokenizator înseamnă un alt model.

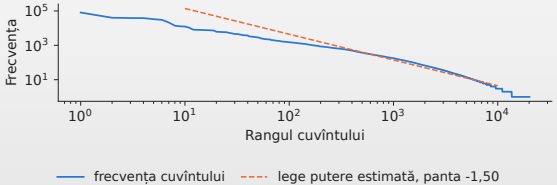
(Malo et al., 2014)

Exercițiu practic: §1

Frecvențele cuvintelor și legea lui Zipf

- **Pe scurt:** în 400 de discursuri ale Băncii Centrale Europene, frecvența scade ca o putere a rangului.

$$f(r) \approx C\, r^{-a} \quad \Longleftrightarrow \quad \ln f(r) = \ln C - a \ln r$$



Sursa: Banca Centrală Europeană, setul de discursuri. Date: 8 ianuarie 2022 - 30 octombrie 2026, 400 de discursuri.

- 1 289 896 de unități lexicale, 20 130 de cuvinte distincte
  - Panta estimată −1,498 (rangurile 10 – 10.000)
  - Cele mai frecvente 10 cuvinte reprezintă 25,0 % din total.
- 32,5 % dintre cuvintele distincte apar o singură dată: niciun model nu le poate învăța numai din acest corpus.

(Zipf, 1949; calcul propriu)

NN_L6_zipf

Cuprins

- Textul ca date
- Reprezentări rare: sacul de cuvinte și TF-IDF
- Reprezentări dense: vectorii de cuvinte (embeddings)
- Clasificatoare de fraze: reprezentări rare și dense
- Studiu de caz: temele tweet-urilor financiare

Sacul de cuvinte: un document este un vector de frecvențe

- **Pe scurt:** ignorăm ordinea și păstrăm de cîte ori apare fiecare cuvînt din vocabular.

$$\mathbf{x}_i = (c_{i1}, \ldots, c_{iV})^\top$$

$$c_{ij} = \#\{\text{apariții ale cuvîntului } j \text{ în documentul } i\}$$

- **unde**
  - V dimensiunea vocabularului (cuvintele întîlnite de cel puțin două ori în partea de antrenare)
  - Matricea documente–termeni $\mathbf{C} \in \mathbb{N}^{n \times V}$ este aproape în întregime zero

- Sacul de cuvinte (BoW) + regresia logistică multinomială este modelul de referință standard pentru text.
- n-gramele adaugă ordinea locală: “profit rose” și “profit fell” devin trăsături.
- Vocabularul este construit numai din frazele de antrenare.

- **În practică:** modelele BoW sînt rapide, ieftine și explicabile cuvînt cu cuvînt: auditorilor le plac.

(Salton & Buckley, 1988)

TF-IDF: ponderi mai mici pentru cuvintele care apar în toate documentele

- **Pe scurt:** un cuvînt este informativ dacă este frecvent în acest document și rar în celelalte.

$$w_{ij} = \underbrace{(1 + \ln c_{ij})}_{\text{tf}} \cdot \underbrace{\left(\ln \frac{1+n}{1+n_j} + 1\right)}_{\text{idf}}, \qquad \tilde{\mathbf{x}}_i = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2}$$

- **unde**
  - $c_{ij} > 0$ frecvența termenului j în documentul i ($w_{ij} = 0$ dacă $c_{ij} = 0$)
  - n documente
  - n_j documente care conțin j
  - Tf sublinear și idf netezit, ca în scikit-learn

- Frecvența termenului–frecvența inversă în documente (TF-IDF) pe unigrame și bigrame, apoi regresie logistică cu ponderi echilibrate ale claselor.
- Cuvinte ca “the”, “company”, `<num>` primesc un idf mic
  - “plunged” unul mare.
- Normalizarea ℓ_2 face comparabile frazele lungi și cele scurte.

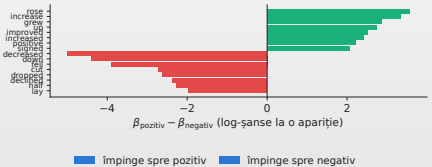
- **În practică:** TF-IDF + regresie logistică este modelul pe care orice furnizor de NLP trebuie să-l depășească pe documentele dumneavoastră.

(Salton & Buckley, 1988)

Modelul liniar învață un dicționar al sentimentului financiar

- **Pe scurt:** diferența dintre coeficienții pozitiv și negativ ai fiecărui cuvînt, regresie logistică pe BoW, PhraseBank.

$$P(y = k \mid \mathbf{x}) = \frac{\exp(\beta_k^\top \mathbf{x} + \alpha_k)}{\sum_l \exp(\beta_l^\top \mathbf{x} + \alpha_l)}$$



Sursa: Financial PhraseBank (Malo et al., 2014). Date: 4.838 de fraze din știri financiare, 5-8 adnotatori pentru fiecare.

- Cele mai pozitive: rose, increase, grew, up, improved
  - Cele mai negative: decreased, down, fell, cut, dropped.
- Dicționarele de limbaj general interpretează greșit cuvintele financiare (“liability”, “tax”): învățați ponderile pe texte financiare.

(Loughran & McDonald, 2011; calcul propriu)

NN_L6_top_words

Cuprins

- Textul ca date
- Reprezentări rare: sacul de cuvinte și TF-IDF
- Reprezentări dense: vectorii de cuvinte (embeddings)
- Clasificatoare de fraze: reprezentări rare și dense
- Studiu de caz: temele tweet-urilor financiare

Ipoteza distribuțională

- **Pe scurt:** cuvintele folosite în contexte asemănătoare primesc vectori asemănători; asemănarea este cosinusul unghiului.

$$\mathbf{e}_w = \mathbf{E}^\top \mathbf{o}_w \in \mathbb{R}^d, \qquad \cos(w, v) = \frac{\mathbf{e}_w^\top \mathbf{e}_v}{\|\mathbf{e}_w\| \, \|\mathbf{e}_v\|}$$

- **unde**
  - $\mathbf{o}_w \in \{0,1\}^V$ vectorul one-hot al cuvîntului w
  - $\mathbf{E} \in \mathbb{R}^{V \times d}$ matricea vectorilor, $d = 100 \ll V = 8\,539$
  - În cod este o căutare, nu un produs

- Vectorii one-hot sînt ortogonali: “inflation” și “prices” au cosinusul 0.
- În vectorii noștri: cos(inflation, prices) = 0,49, cos(inflation, climate) = 0,16, cos(banks, lending) = 0,46.

- **În practică:** vectorii denși permit unui model antrenat pe “profit rose” să generalizeze la “earnings increased”.

(Firth, 1957; Bishop & Bishop, 2024, sect. 12.2)

word2vec skip-gram: prezicem contextul din cuvînt

- **Pe scurt:** fiecare cuvînt din fereastra din jurul cuvîntului central este prezis printr-un softmax peste vocabular.

$$\max_{\mathbf{U},\mathbf{V}} \frac{1}{T} \sum_{t=1}^{T} \sum_{0<|j|\leq c} \log p(w_{t+j} \mid w_t)$$

$$p(o \mid w) = \frac{\exp(\mathbf{u}_o^\top \mathbf{v}_w)}{\sum_{k=1}^{V} \exp(\mathbf{u}_k^\top \mathbf{v}_w)}$$

- **unde**
  - $\mathbf{v}_w$ vectorul de intrare (al cuvîntului central)
  - $\mathbf{u}_o$ vectorul de ieșire (al contextului)
  - c raza ferestrei
  - Numitorul costă $O(V)$ pentru fiecare pereche

- **În practică:** softmax-ul peste 8 539 de cuvinte este prea lent
  - Îl înlocuim prin eșantionare negativă.

(Mikolov et al., 2013)

- Sacul continuu de cuvinte (CBOW) este varianta în oglindă: prezice cuvîntul central din media contextului.
- Rețea liniară cu un strat ascuns
  - Vectorii de cuvinte sînt prima ei matrice de ponderi.
- Corpusul nostru: 3 679 754 de perechi (cuvînt central, context) pe epocă, după subeșantionare.

Eșantionarea negativă: deosebim perechile reale de cele aleatoare

- **Pe scurt:** înlocuim softmax-ul cu $K+1$ regresii logistice binare: un cuvînt real de context și K cuvinte de zgomot.

$$\ell(w, o) = -\log \sigma(\mathbf{u}_o^\top \mathbf{v}_w) - \sum_{k=1}^{K} \log \sigma(-\mathbf{u}_{n_k}^\top \mathbf{v}_w)$$

$$n_k \sim P_n(w) \propto f(w)^{3/4}$$

- **unde**
  - σ funcția logistică
  - $K = 5$
  - $f(w)$ frecvența în corpus
  - Cuvintele frecvente sînt păstrate cu probabilitatea $\min(1, \sqrt{t/f(w)} + t/f(w))$, $t = 10^{-4}$

- Skip-gram cu eșantionare negativă (SGNS), $d = 100$, fereastră 5, cuvinte cu cel puțin 5 apariții, Adam, 5 epoci în 120 de secunde.
- Pierderea pe pereche scade de la 2,622 (epoca 1) la 2,223.
- SGNS factorizează implicit o matrice deplasată de informație mutuală punctuală (PMI) (anexă).

- **În practică:** același truc, contrastul dintre un element real și negative eșantionate, antrenează sisteme de recomandare și reprezentări pentru detectarea fraudei.

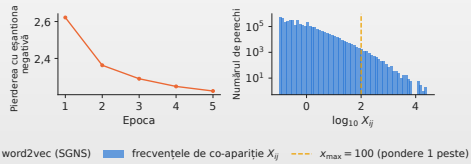
(Mikolov et al., 2013; Levy & Goldberg, 2014)

GloVe: aproximăm logaritmul frecvențelor de co-apariție

- **Pe scurt:** numărăm o singură dată cît de des apar cuvintele împreună, apoi aproximăm logaritmii acestor frecvențe prin produse scalare, cu o pierdere de tip cele mai mici pătrate ponderate.

$$J = \sum_{i,j:\, X_{ij}>0} f(X_{ij}) \left(\mathbf{w}_i^\top \tilde{\mathbf{w}}_j + b_i + \tilde{b}_j - \ln X_{ij}\right)^2$$

$$f(x) = \min\{1, (x/x_{\max})^\alpha\}$$



Sursa: Banca Centrală Europeană, setul de discursuri. Date: 8 ianuarie 2022 - 30 octombrie 2026, 400 de discursuri.

- **unde**
  - X_{ij} co-aparițiile la cel mult 10 cuvinte distanță, ponderate cu $1/\text{distanță}$
  - $x_{\max} = 100$, $\alpha = 0{,}75$
  - Vectorul final este $\mathbf{w}_i + \tilde{\mathbf{w}}_i$

- **În practică:** 3 841 517 de perechi nenule (5,269 % din V^2)
  - 99,81 % sînt sub $x_{\max}$ și primesc o pondere sub 1.

(Pennington et al., 2014; calcul propriu)

NN_L6_emb_training

Vectori de cuvinte din frecvențe: PMI și factorizarea trunchiată a matricei

- **Pe scurt:** înaintea vectorilor neuronali, vectorii denși proveneau din factorizarea unei matrice de asocieri între cuvinte.

$$\mathrm{PMI}(w, c) = \ln \frac{P(w, c)}{P(w)P(c)}$$

$$\mathbf{M}_+ \approx \mathbf{U}_d \boldsymbol{\Sigma}_d \mathbf{V}_d^\top, \qquad \mathbf{E} = \mathbf{U}_d \boldsymbol{\Sigma}_d^{1/2}$$

- **unde**
  - $P(w, c)$ ponderea co-aparițiilor lui w și c
  - $\mathbf{M}_+$ matricea valorilor $\max(\mathrm{PMI}, 0)$
  - $\mathbf{U}_d, \boldsymbol{\Sigma}_d, \mathbf{V}_d$ primii d termeni ai descompunerii în valori singulare (SVD)

- Informația mutuală punctuală (PMI) este pozitivă cînd două cuvinte apar împreună mai des decît la întîmplare.
- Tabelul nostru de co-apariții are 3 841 517 de perechi nenule, 5,269 % din cele V^2 celule: o matrice rară, deci SVD este ieftin.
- word2vec, GloVe și SVD-ul matricei PMI dau vecini asemănători pe același corpus.

- **În practică:** o factorizare a frecvențelor este transparentă și reproductibilă: un prim vector de cuvinte bun pentru un corpus intern.

(Levy & Goldberg, 2014)

Cei mai apropiați vecini în vectorii noștri, învățați din discursurile băncii centrale

Cuvîntul	word2vec (SGNS)	GloVe
`inflation`	headline, underlying, revert, hovering, medium	expectations, underlying, headline, core, term
`energy`	fertilisers, food, hicpxx, prices, renewable	prices, food, shock, commodity, renewable
`banks`	repaying, supervisors, themselves, obtaining, supervised	central, their, supervisors, they, have
`climate`	change, environmental, degradation, nature, ipcc	change, nature, environmental, related, risks
`digital`	convenient, functionality, kantar, erpb, payments	euro, payments, form, would, cash
`wages`	wage, profits, catch, negotiated, salaries	real, nominal, higher, incomes, rising
`ukraine`	invasion, russia, aggression, russian, brutal	invasion, russia, war, unjustified, russian
`rates`	interest, rate, compressed, hikes, repriced	interest, rate, short, lending, long

- Cele mai apropiate cinci cuvinte după cosinus; ambele modele sînt antrenate pe aceleași 1 289 896 de unități lexicale din discursurile BCE; suprapunerea medie a celor două liste: 1,88 cuvinte.

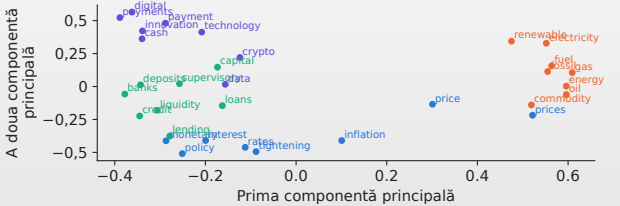
- **În practică:** un analist al politicii monetare poate extinde o căutare după “inflation” cu vecinii învățați, în locul unei liste făcute de mînă.

(calcul propriu)

Exerciții practic: §2

Temele activității băncii centrale formează regiuni în spațiul vectorilor

- **Pe scurt:** proiecția bidimensională prin analiza componentelor principale (PCA) a 32 de cuvinte alese din patru grupuri; grupurile nu au fost date modelului.



Sursa: Banca Centrală Europeană, setul de discursuri. Date: 8 ianuarie 2022 - 30 octombrie 2026, 400 de discursuri.

- Două componente păstrează 31,9 % din varianța acestor 32 de vectori, deci imaginea este doar o umbră aproximativă a spațiului $\mathbb{R}^{100}$.

(calcul propriu)

NN_L6_w2v_pca

Analogii și limitele unui singur vector pentru fiecare cuvînt

$a : b :: c :$?	răspuns
germany : german : : france	fran
italy : italian : : spain	commissioned
banks : bank : : households	household
rise : rising : : fall	buffered
higher : lower : : increase	marginal

- Răspunsul $= \arg\max_v \cos(\mathbf{e}_v, \mathbf{e}_b - \mathbf{e}_a + \mathbf{e}_c)$, excluzînd a, b, c; word2vec-ul nostru, un singur corpus mic.

(Mikolov et al., 2013; Bojanowski et al., 2017)

- **Vectori statici**
  - Un singur vector pentru fiecare cuvînt, indiferent de frază
  - "interest" (dobîndă) și "interest" (atenție) au același vector; la fel "bank"
- **Acoperirea**
  - Vocabularul BCE acoperă 84,7 % din unitățile lexicale din PhraseBank și 76,4 % din cele din tweet-uri
  - Restul sînt în afara vocabularului (OOV); modelele cu subcuvinte (fastText) și transformerele (cursul 7) rezolvă ambele probleme

Vectori din subcuvinte pentru cuvintele nevăzute

- **Pe scurt:** construim vectorul unui cuvînt din vectorii *n*-gramelor sale de caractere, astfel încît un simbol bursier nou sau o greșeală de scriere primește totuși un vector.

$$\mathbf{e}_w = \sum_{g \in \mathcal{G}(w)} \mathbf{z}_g$$

- **unde**
  - $\mathcal{G}(w)$ *n*-gramele de caractere ale lui w (de exemplu de 3–6 caractere, cu simboluri de margine) și cuvîntul însuși
  - $\mathbf{z}_g$ vectori învățați ai *n*-gramelor

- Acoperirea vocabularului nostru BCE la nivel de cuvînt: 84,7 % din unitățile lexicale PhraseBank, 76,4 % din cele din tweet-uri.
- Modelele cu subcuvinte (fastText) împart informația între “profit”, “profits” și “profitability”.
- Transformerele (cursul 7) folosesc aceeași idee, cu unități de subcuvinte învățate.

- **În practică:** pentru tweet-uri și mesaje, cu simboluri bursiere și variante de scriere, folosiți vectori din subcuvinte sau un tokenizator pe subcuvinte.

(Bojanowski et al., 2017)

Cuprins

- Textul ca date
- Reprezentări rare: sacul de cuvinte și TF-IDF
- Reprezentări dense: vectorii de cuvinte (embeddings)
- Clasificatoare de fraze: reprezentări rare și dense
- Studiu de caz: temele tweet-urilor financiare

O rețea embedding bag: media vectorilor de cuvinte și un strat softmax

- **Pe scurt:** cel mai simplu clasificator neuronal de text învață împreună vectorii de cuvinte și clasificatorul.

$$\mathbf{z} = \frac{1}{|S|} \sum_{s \in S} \mathbf{e}_s$$

$$\hat{\mathbf{p}} = \text{softmax}(\mathbf{W}\,\text{dropout}(\mathbf{z}) + \mathbf{b})$$

$$\mathcal{L} = -\frac{1}{n} \sum_i \omega_{y_i} \log \hat{p}_{i,y_i}$$

- Start aleator sau start word2vec: rîndurile lui **E** sînt inițializate cu vectorii noștri BCE, apoi ajustate.
- Comparăm și: media vectorilor word2vec + regresie logistică (vectori înghețați), TF-IDF + un perceptron multistrat (MLP) cu 128 de unități ascunse.

- **unde**
  - S unitățile lexicale și bigramele frazei
  - $\mathbf{e}_s$ vectori învățați ($d = 100$)
  - $\omega_k = n/(3n_k)$ ponderi ale claselor, împotriva dominanței clasei "neutru"

- **În practică:** aceasta este arhitectura clasificatoarelor de tip fastText, folosite în producție pentru milioane de documente pe oră.

(Joulin et al., 2017)

Regresia logistică pe text: gradientul

- **Pe scurt:** modelul logistic multinomial este o rețea cu un singur strat; gradientul lui compară probabilitățile prezise cu clasele reale.

$$\mathcal{L}(\mathbf{B}) = -\sum_i \sum_k y_{ik} \ln \hat{p}_{ik} + \frac{1}{2C}\|\mathbf{B}\|_F^2$$
$$\nabla_{\mathbf{B}}\mathcal{L} = \mathbf{X}^\top(\hat{\mathbf{P}} - \mathbf{Y}) + \mathbf{B}/C$$

- **unde**
  - $\mathbf{X}$ matricea TF-IDF ($n \times V$, rară)
  - $\mathbf{Y}$ etichetele one-hot
  - $\hat{\mathbf{P}}$ probabilitățile softmax
  - C inversul penalizării, ales pe validare din $\{0{,}1; 1; 10\}$

- Pierdere convexă: un singur optim, fără sămînță de raportat
  - Perceptronul multistrat (MLP) pe aceleași intrări adaugă un strat ascuns de 128 de unități.
- Macro-F1 pe PhraseBank: logistic 73,49 %, MLP 71,37 %: stratul ascuns nu ajută aici.

- **În practică:** cu cîteva mii de fraze scurte, flexibilitatea suplimentară a unei rețele aduce variantă, nu acuratețe.

(calcul propriu)

Vectorul frazei construit din vectorii cuvintelor

- **Pe scurt:** media vectorilor de cuvinte este cea mai simplă reprezentare a unei fraze și pierde exact ce contează pentru sentiment.

$$\mathbf{s}_i = \frac{1}{|S_i|} \sum_{w \in S_i} \mathbf{e}_w, \qquad \mathbf{s}_i^{\mathrm{w}} = \frac{\sum_w a_w \mathbf{e}_w}{\sum_w a_w}, \quad a_w = \frac{\alpha}{\alpha + f(w)}$$

- **unde**
  - S_i unitățile lexicale din vocabular ale frazei i
  - a_w o pondere care micșorează influența cuvintelor frecvente
  - α o constantă mică

- Media vectorilor BCE înghețați + regresie logistică: macro-F1 57,30 % pe PhraseBank și 53,71 % pe tweet-uri.
- "rose" și "fell" sînt apropiate într-un spațiu de co-apariții (aceleași contexte), deci media estompează semnul.
- Ajustarea vectorilor pe date etichetate (embedding bag) recuperează cea mai mare parte a diferenței.

- **În practică:** vectorii generali sînt buni pentru căutare și similaritate, nu în mod automat pentru sentiment.

(calcul propriu)

Protocolul și metrica: macro-F1 pentru o țintă dezechilibrată

- **Pe scurt:** clasa "neutru" reprezintă 59,3 % din PhraseBank, deci acuratețea răsplătește un model care spune mereu "neutru".

$$\mathrm{F1}_k = \frac{2P_kR_k}{P_k + R_k}, \qquad \text{macro-F1} = \frac{1}{3}\sum_{k=1}^{3} \mathrm{F1}_k$$

- **unde**
  - P_k precizia și R_k sensibilitatea clasei k
  - Fiecare clasă contează la fel, indiferent de mărime

- PhraseBank: 5 împărțiri aleatoare stratificate 70/15/15
  - Tweet-uri: setul de validare publicat, folosit ca test, 3 semințe.
- Penalizarea C, oprirea timpurie și vocabularul sînt alese numai pe partea de validare.

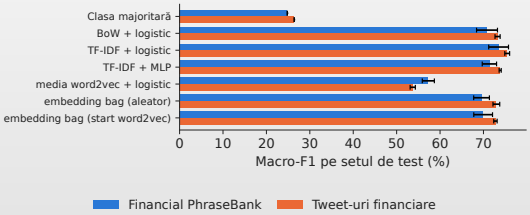
- **În practică:** pentru un sistem de alertă, raportați sensibilitatea clasei negative alături de macro-F1: aceasta este clasa care costă bani.

(calcul propriu)

Exerciții practic: §3

Pe frazele financiare scurte, TF-IDF cu regresie logistică are cel mai mare macro-F1

- **Pe scurt:** macro-F1 pe test pentru șapte reprezentări și clasificatoare, PhraseBank și tweet-uri financiare (bare: media, mustăți: abaterea standard).



Sursa: Financial PhraseBank (Malo et al., 2014); Twitter financial news (Hugging Face, zeroshot). Date: PhraseBank: 5 împărțiri aleatoare 70/15/15; tweet-uri: setul de validare publicat, folosit ca test, 3 semințe.

- Pe PhraseBank: 73,49 (TF-IDF + logistic), față de 69,93 (embedding bag, start word2vec)
  - Clasa majoritară 24,81.

(calcul propriu)

NN_L6_sentiment_models

Rezultatele pentru sentiment, cu incertitudine

Modelul	PhraseBank macro-F1	acur.	Tweet-uri macro-F1	acur.
Clasa majoritară	$24{,}81 \pm 0{,}00$	59,28	$26{,}40 \pm 0{,}00$	65,58
BoW + logistic	$70{,}82 \pm 2{,}41$	74,94	$73{,}25 \pm 0{,}59$	80,00
TF-IDF + logistic	$73{,}49 \pm 2{,}28$	77,69	$75{,}47 \pm 0{,}56$	81,77
TF-IDF + MLP	$71{,}37 \pm 1{,}65$	76,23	$73{,}85 \pm 0{,}24$	80,56
media word2vec + logistic	$57{,}30 \pm 1{,}39$	62,01	$53{,}71 \pm 0{,}60$	61,59
embedding bag, start aleator	$69{,}56 \pm 1{,}81$	74,11	$72{,}97 \pm 0{,}78$	79,77
embedding bag, start word2vec	$69{,}93 \pm 2{,}16$	74,99	$72{,}76 \pm 0{,}39$	79,51

- Pereche, pe cele 5 împărțiri PhraseBank: TF-IDF + logistic minus embedding bag (start word2vec) $= +3{,}55$ puncte, $t = 9{,}88$ (4 grade de libertate).
- Pornirea de la vectorii noștri word2vec BCE schimbă rezultatul embedding bag cu $+0{,}37$ puncte pe PhraseBank și cu $-0{,}21$ pe tweet-uri.

- **În practică:** modelul cel mai ieftin aflat la cel mult o abatere standard de cel mai bun este cel care trebuie folosit și explicat.

(calcul propriu)

Alegerea dimensiunii vectorului de cuvinte

Dimensiunea d	Macro-F1 (%)
8	67,3
16	66,5
32	67,2
64	67,8
128	68,8
256	67,3

- Embedding bag, start aleator, împărțirea 1 PhraseBank, media pe 3 semințe.

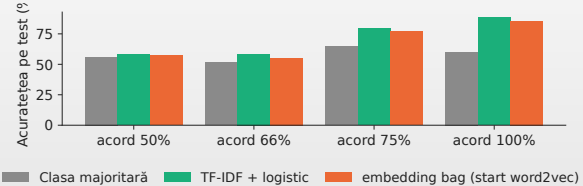
- Parametrii stratului de embedding: $V \times d$
  - Cu bigrame, V are cîteva mii de elemente, deci d determină mărimea modelului.
- Aici scorul variază puțin cu d: de la 67,3 la $d = 8$ la 68,8 la $d = 128$.
- O sarcină cu trei clase pe fraze scurte are nevoie de puține dimensiuni
  - Vectorii de uz general folosesc 100–300 pentru că servesc multor sarcini.

- **În practică:** alegeți d pe datele de validare
  - Un model mai mic este mai ieftin de rulat și de explicat.

(calcul propriu)

Erorile de clasificare și acordul adnotatorilor

- **Pe scurt:** acuratețea pe test în funcție de ponderea adnotatorilor care au fost de acord asupra etichetei frazei.



Sursa: Financial PhraseBank (Malo et al., 2014). Date: frazele de test din 5 împărțiri aleatoare, după ponderea adnotatorilor care sînt de acord.

- TF-IDF + logistic: 88,4 % pe frazele cu acord unanim, 58,1 % la un acord de 50 %
  - 2 259 din 4 838 de fraze au acord unanim.

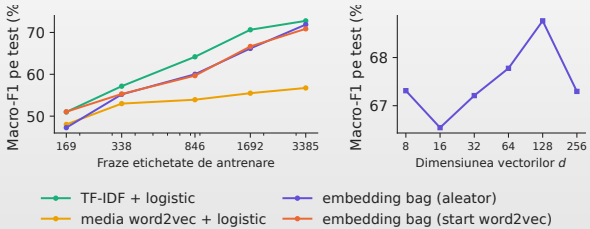
- **În practică:** zgomotul din etichete limitează orice model: înainte de a cumpăra un model mai mare, verificați cît de mult sînt de acord propriii analiști.

(calcul propriu)

NN_L6_agreement

Curbele de învățare cu puține etichete: inițializarea word2vec și TF-IDF

- **Pe scurt:** stînga: macro-F1 pe test în funcție de numărul de fraze etichetate; dreapta: dimensiunea vectorilor din embedding bag.



Sursa: Financial PhraseBank (Malo et al., 2014). Date: stînga: 3 împărțiri aleatoare; dreapta: embedding bag (start aleator), împărțirea 1, 3 semințe.

- Cu 169 de fraze etichetate: TF-IDF 51,0, media word2vec 48,0, embedding bag aleator 47,2, start word2vec 51,0
  - Cu toate cele 3 385: 72,8, 56,7, 71,9, 70,9.

(calcul propriu)

NN_L6_learning_curve

Principalul tip de eroare: știri negative clasificate ca neutre

Real \ prezis	neg.	neu.	poz.
negativ	–	107	48
pozitiv	50	263	–

- TF-IDF + logistic, cele 5 seturi de test PhraseBank reunite (3 635 de fraze).

- Sensibilitatea: negativ 65,9 %, neutru 84,1 %, pozitiv 69,5 %.
- Erorile de semn opus sînt rare
  - Cele mai multe erori mută fraza în clasa neutră.
- Eroare tipică: o scădere descrisă cu cuvinte neutre și cifre (“sales were EUR 10 mn, down from 12”).

- **În practică:** în finanțe, cifrele poartă sentiment
  - Înlocuirea lor cu o singură unitate pierde comparațiile de tip “down from” pe care un model secvențial le poate citi.

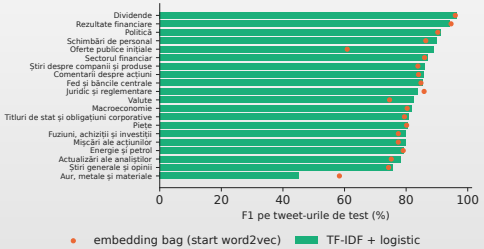
(calcul propriu)

Cuprins

- Textul ca date
- Reprezentări rare: sacul de cuvinte și TF-IDF
- Reprezentări dense: vectorii de cuvinte (embeddings)
- Clasificatoare de fraze: reprezentări rare și dense
- Studiu de caz: temele tweet-urilor financiare

Clasificarea pe teme: F1 este mare pentru rezultate și Fed, mic pentru știrile generale

- **Pe scurt:** F1 pe fiecare temă, pentru cele 4 117 de tweet-uri de test (setul de validare publicat).



Sursa: Seturile Twitter financial news (Hugging Face, zeroshot). Date: 4.117 de tweet-uri de test (setul de validare publicat). 20 de teme.

- Macro-F1: TF-IDF + logistic 82,62 %, embedding bag (start word2vec) 80,45 %
  - Acuratețe 84,48 %, față de 20,69 % pentru tema majoritară.

(calcul propriu)

NN_L6_topic_f1

Aplicații ale modelelor de text în finanțe

- **Triajul știrilor și al tweet-urilor**
  - Direcționarea pe teme (cea mai bună: Dividende, F1 96,4 %) și semnalarea celor negative pentru analist
- **Comunicarea băncii centrale**
  - Urmărirea cuvintelor din jurul lui “inflation” sau “energy”, discurs cu discurs
- **Conformitate și analiza de mediu, social și guvernanță (ESG)**
  - Căutarea în rapoarte cu sinonime învățate, nu cu liste fixe de cuvinte-cheie
- **Limitele modelelor de azi**
  - Ordinea cuvintelor dincolo de bigrame se pierde
  - Un singur vector pentru fiecare cuvînt ignoră contextul
  - Tema cea mai slabă: Aur, metale și materiale, F1 45,2 %
- **Urmează: atenția**
  - Vectori contextuali din transformere, de exemplu reprezentări bidirecționale ale codificatorului transformer (BERT) și un model BERT adaptat textelor financiare (FinBERT), în cursul 7, comparați corect cu aceste modele de referință

- **În practică:** începeți orice proiect de text cu TF-IDF + regresie logistică
  - Treceți la vectori de cuvinte sau transformere doar cînd cîștigul este demonstrat pe datele dumneavoastră.

(calcul propriu)

Test 1: idf

- Un cuvînt apare în fiecare dintre cele n fraze de antrenare
  - Care este idf-ul său netezit?

Răspuns

- $\ln \frac{1+n}{1+n} + 1 = 1$: valoarea minimă, deci aproape nu modifică vectorul normalizat.

Test 2: eșantionarea negativă

- De ce extrage word2vec cuvintele negative din $f(w)^{3/4}$ și nu din $f(w)$?

Răspuns

- Puterea 3/4 aplatizează distribuția: cuvintele rare sînt extrase mai des decît frecvența lor, iar cele frecvente, ca “the”, mai rar, ceea ce dă contraste mai utile.

Test 3: acuratețea

- Un furnizor raportează o acuratețe de 60 % pe sentimentul din PhraseBank
  - Este un rezultat bun?

Răspuns

- Nu: prognoza "mereu neutru" dă 59,28 %
  - Cereți macro-F1 (TF-IDF + logistic, la noi: 73,49 %).

Test 4: transferul

- Aveți doar 169 de fraze etichetate
  - Ce reprezentare ajută cel mai mult?

Răspuns

- Comparați pe datele de validare: aici media word2vec dă 48,0, TF-IDF 51,0, embedding bag cu start word2vec 51,0 macro-F1.

Test 5: polisemia

- De ce nu poate un vector static să deosebească “interest rate” de “of interest to investors”?

Răspuns

- Are un singur vector pentru fiecare cuvînt, media tuturor utilizărilor lui
  - Modelele contextuale (cursul 7) dau un vector diferit în fiecare frază.

Recapitulare

- Textul ca date
- Reprezentări rare: sacul de cuvinte și TF-IDF
- Reprezentări dense: vectorii de cuvinte (embeddings)
- Clasificatoare de fraze: reprezentări rare și dense
- Studiu de caz: temele tweet-urilor financiare

- Textul devine numere prin unități lexicale și frecvențe (BoW, TF-IDF) sau prin vectori denși (word2vec, GloVe) învățați din co-apariții.
- Pe fraze financiare reale, TF-IDF + regresie logistică este modelul de depășit
  - Startul word2vec ajută rețeaua doar cînd etichetele sînt puține
  - Urmează atenția și transformerele (cursul 7).

Cercetări ale grupului nostru: modele de limbaj în finanțe

- **Modelele lingvistice mari (LLM) prognozează valoarea expusă la risc (VaR) și pierderea medie peste VaR (ES)** (Pele, Bolovăneanu, Lin, Ren, Ginavar, Spilak, Andrei, Toma, Lessmann & Härdle, 2026, Expert Systems with Applications)
  - Lucrarea: "In the beginning was the Word: LLM-VaR and LLM-ES"; prognozele LLM sînt comparate cu modelele clasice de risc prin teste retrospective formale
  - Modelele de text trec de la scoruri de sentiment la cifrele de risc propriu-zise (cursul 12)
- **BitMood: analiza cu inteligență artificială (AI) a tendințelor Bitcoin prin emoțiile de pe Facebook** (Conda, Găman, Bâg, Mazurencu-Marinescu-Pele, Pele & Härdle, 2026, Digital Finance)
  - 120.000 de postări de pe Facebook, 2015–2023, clasificate cu FinBERT în scoruri compuse de sentiment, împreună cu reacțiile cititorilor
  - Sentimentul prognozează volumul tranzacțiilor cu Bitcoin; strategiile bazate pe sentiment depășesc strategia cumpără-și-păstrează; modelarea temelor urmărește subiectele discuției
- Rezultatele de pe acest diapozitiv sînt cele raportate în lucrări; celelalte cifre din curs provin din calculele noastre.

(Academia de Studii Economice din București; Academia Română, Institutul de Prognoză Economică)

Explorați vizual

- **Embedding Atlas (Apple)**: instrument open-source care desenează embeddings de text în două dimensiuni, cu etichete automate ale grupurilor, căutare și vecini apropiați
  - Rulați-l pe propozițiile PhraseBank din acest curs și colorați punctele după sentiment.
- **Transformer Explainer**: primul pas al unui model lingvistic pe propoziția voastră: tokenii, embeddings-urile lor și embeddings-urile de poziție adunate la ele.
- **3Blue1Brown, capitolul 5**: direcții în spațiul embeddings, de exemplu diferența dintre vectorii lui „man” și „woman”.

Lecturi și resurse suplimentare

- **Mikolov et al. (2013)** și **Pennington et al. (2014)**: cele două articole despre vectorii de cuvinte, ambele scurte.
- **Levy & Goldberg (2014)**: de ce word2vec este o factorizare de matrice.
- **Loughran & McDonald (2011)**: dicționarul de sentiment financiar și de ce eșuează dicționarele generale.
- **Acest curs**: cod, date, notebook-uri Jupyter — github.com/danpele/NEURAL_BIZ.
- **Exercițiu practic**: notebook-ul reproduce graficele acestui curs — notebooks/.
- **Seminarul 6**: un indice de sentiment al discursurilor BCE, construit cu propriul clasificator.

Bibliografie (1/2)

- Bishop, C. M. & Bishop, H. (2024). *Deep Learning: Foundations and Concepts*, ch. 12. Springer. doi:10.1007/978-3-031-45468-4
- Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. *TACL*, 5, 135–146. doi:10.1162/tacl_a_00051
- Conda, A. I., Găman, Ș., Bâg, R. C., Mazurencu-Marinescu-Pele, M., Pele, D. T. & Härdle, W. K. (2026). BitMood: AI analysis of Bitcoin trends via Facebook emotions. *Digital Finance*, 8(3), 42. doi:10.1007/s42521-026-00205-3
- Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, 1–32. Blackwell.
- Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2017). Bag of tricks for efficient text classification. *EACL*, 427–431. doi:10.18653/v1/E17-2068
- Levy, O. & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *NIPS*, 27, 2177–2185. papers.nips.cc
- 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. doi:10.1111/j.1540-6261.2010.01625.x

Bibliografie (2/2)

- Malo, P., Sinha, A., Korhonen, P., Wallenius, J. & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *JASIST*, 65(4), 782–796. doi:10.1002/asi.23062
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NIPS*, 26, 3111–3119. arXiv:1310.4546
- Pele, D. T., Bolovăneanu, V., Lin, M.-B., Ren, R., Ginavar, A. T., Spilak, B., Andrei, A.-V., Toma, F.-M., Lessmann, S. & Härdle, W. K. (2026). In the beginning was the Word: LLM-VaR and LLM-ES. *Expert Systems with Applications*, 295, 128676. doi:10.1016/j.eswa.2025.128676
- Pennington, J., Socher, R. & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP*, 1532–1543. doi:10.3115/v1/D14-1162
- Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. doi:10.1016/0306-4573(88)90021-0
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Anexă: gradientul eșantionării negative

- Pentru o pereche, $\ell = -\log \sigma(s_o) - \sum_k \log \sigma(-s_k)$, cu $s_o = \mathbf{u}_o^\top \mathbf{v}_w$, $s_k = \mathbf{u}_{n_k}^\top \mathbf{v}_w$.
- Folosind $\frac{d}{ds} \log \sigma(s) = 1 - \sigma(s)$: $\partial \ell / \partial \mathbf{v}_w = -(1 - \sigma(s_o))\mathbf{u}_o + \sum_k \sigma(s_k)\mathbf{u}_{n_k}$.
- $\partial \ell / \partial \mathbf{u}_o = -(1 - \sigma(s_o))\mathbf{v}_w$ și $\partial \ell / \partial \mathbf{u}_{n_k} = \sigma(s_k)\mathbf{v}_w$: contextele reale sînt atrase, cuvintele de zgomot sînt respinse.
- La optim, $\mathbf{u}_o^\top \mathbf{v}_w = \mathrm{PMI}(w, o) - \ln K$, cu $\mathrm{PMI}(w, o) = \ln \frac{P(w,o)}{P(w)P(o)}$ (Levy & Goldberg, 2014).

Anexă: modelele în cod

```
tf = TfidfVectorizer(ngram_range=(1, 2), min_df=2, sublinear_tf=True).fit(train_text)
clf = LogisticRegression(C=C, class_weight='balanced', max_iter=3000).fit(tf.transform(train_text), y)

class EmbBag(nn.Module):                          # fastText-style classifier
    def __init__(self, n_vocab, dim, n_cls):
        super().__init__()
        self.emb = nn.EmbeddingBag(n_vocab, dim, mode='mean')
        self.out = nn.Linear(dim, n_cls)
    def forward(self, ids, offsets):
        return self.out(F.dropout(self.emb(ids, offsets), 0.3, self.training))
# SGNS step:  loss = -(logsigmoid((v*u_pos).sum(1)) + logsigmoid(-(U_neg @
v[...,None]).squeeze()).sum(1)).mean()
```

- Codul complet: lectures/L6/analysis_L6.py și notebook-ul acestui curs.