RAPORT

ETAPA 4 - Decembrie 2024

Titlu Proiect: Soluții informatice pentru analiza impactului rețelelor de social media asupra instrumentelor investiționale cu grad ridicat de risc: cryptomonede și bursă

Domeniul științific: Științe economice

Membri:
prof.univ.dr. Adela Bâra,
drd. Costin Băroiu

În cadrul acestei etape, am diseminat rezultatele cercetării în următoarele lucrări:

Articole Publicate:

1) ”**Big data management and NoSQL databases**” (2023). Oprea, S.-V., Bâra, A., Oprea, N.. Ovidius University Annals, Economic Sciences Series: https://ideas.repec.org/a/ovi/oviste/vxxiiiy2023i1p466-475.html
2) “**Forecasting the Spot Market Electricity Price with a Long Short-Term Memory Model Architecture in a Disruptive Economic and Geopolitical Context**” (2023). Bâra, A., Oprea, S.-V., Băroiu, A.-C.. International Journal of Computational Intelligence Systems, https://link.springer.com/article/10.1007/s44196-023-00309-3
3) Bâra, A., Oprea, S.V, **The Impact of Academic Publications over the Last Decade on Historical Bitcoin Prices using Generative Models**. Journal of Theoretical and Applied Electronic Commerce Research. 2024; 19(1):538-560. https://doi.org/10.3390/jtaer19010029
4) Băroiu, A.-C.; Bâra, A. **A Descriptive-Predictive–Prescriptive Framework for the Social-Media–Cryptocurrencies Relationship**. Electronics 2024, 13, 1277. https://doi.org/10.3390/electronics13071277

Articole în curs de publicare:

5) Băroiu, A.-C.; Bâra, A. LLM-Based Applications in the Financial Sector. Articolul va fi trimis spre evaluare/publicare in cadrul conferintei ICBE
6) Oprea, S.V., Bâra, A., **A LLM-based Recommendation System for Bitcoin Trading Strategy / A data-preprocessing approach for improving machine learning algorithms for classifiying customers**. Articolul este in curs de concepere/redactare si va fi trimis spre evaluare/publicare in cadrul unei conferinte internationale

Proiectul, intitulat "**Soluții informatice pentru analiza impactului rețelelor de social media asupra instrumentelor investiționale cu grad ridicat de risc: cryptomonede și bursă**", are ca obiectiv principal investigarea influenței pe care rețelele de social media o au asupra piețelor financiare volatile, în special criptomonedele și acțiunile bursiere.

În acest context, articolul 1) „**The Impact of Academic Publications over the Last Decade on Historical Bitcoin Prices using Generative Models**”, publicat in revista *Journal of Theoretical and Applied Electronic Commerce Research*. Unul dintre seturile de date a fost extras de pe platforma Web of Science Clarivate pe 30 iulie 2023. Două cuvinte au fost incluse în căutare: Bitcoin și cryptocurrency. Numărul total de lucrări de cercetare care au îndeplinit criteriile de căutare a fost de 9105, iar numărul total de coloane a fost de 72 (precum: Tipul publicației, Abrevierea jurnalului, Editura, Autori, Afiliere, Rezumat, Cuvinte-cheie, Titlul articolului, Numărul referințelor citate, Numărul de citări, WoS Core, Numărul de citări în toate bazele de date, ISSN, eISSN, Data publicării, Volum, Număr, DOI, Link DOI, Numărul de pagini, Categorii WoS, Index Web of Science, Domenii de cercetare, Statut de lucrare foarte citată, UT (ID unic WoS), An, etc.).

Rezumatul a fost analizat și transformat din text în caracteristici. Un total de 357 de articole au fost excluse din listă deoarece nu aveau rezumat. În plus, 3839 de articole nu aveau o dată de publicare care trebuia convertită într-o lună validă. Prin urmare, s-a aplicat o funcție aleatorie între unu și doisprezece. Un total de 2361 de lucrări au fost în domeniul Sistemelor de Informații din Știința Calculatoarelor, 2001 în Teorie și Metode din Știința Calculatoarelor și 1568 în Finanțe de Afaceri. Cercetătorii din SUA au scris 1800 de lucrări, 1647 au fost din China și 818 din Anglia. Un total de 5704 lucrări au fost din categoria articolelor, 2911 au fost lucrări de conferință și 309 au fost în categoria accesului anticipat.

Numărul publicațiilor legate de Bitcoin și alte criptomonede a crescut progresiv în timp (așa cum se arată în Figura 1).

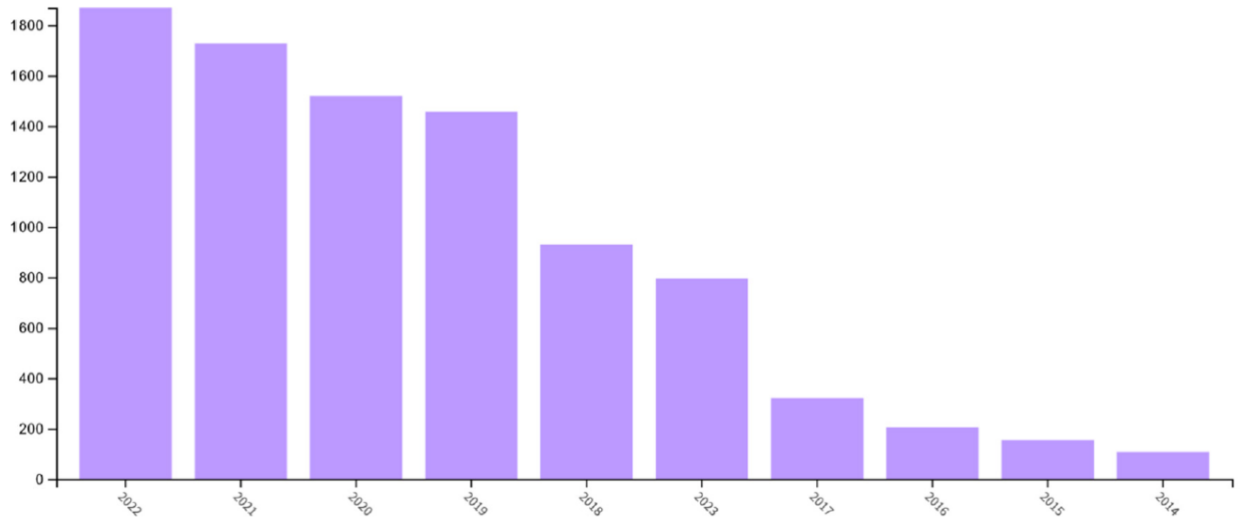


Figura 1. Frecvența publicațiilor din 2014 până în 2022 în ordine descrescătoare. Sursa: Web of Science.

Un total de 2080 de lucrări au fost publicate de editura IEEE, 1740 de Elsevier și 1272 de Springer Nature. Un total de 4058 de publicații au fost incluse în domeniul principal al Științei Calculatoarelor, 2908 în Economie de Afaceri, iar 1571 în Inginerie.

Celălalt set de date utilizat în această cercetare constă în prețurile lunare ale Bitcoin, extrase de pe platforma Investing.com (accesat pe 23 iunie 2023). Prețurile Bitcoin au început să crească în 2017 și apoi din nou în 2019, însă cel mai mare salt a fost înregistrat în 2021. Curba prețurilor în 2021 a prezentat două vârfuri, urmate de o coborâre abruptă în 2022, ajungând la aproape 15.000 USD. În prezent, prețurile sunt la jumătate față de cât erau în 2021, când au atins aproape 70.000 USD (așa cum se arată în Figura 2).

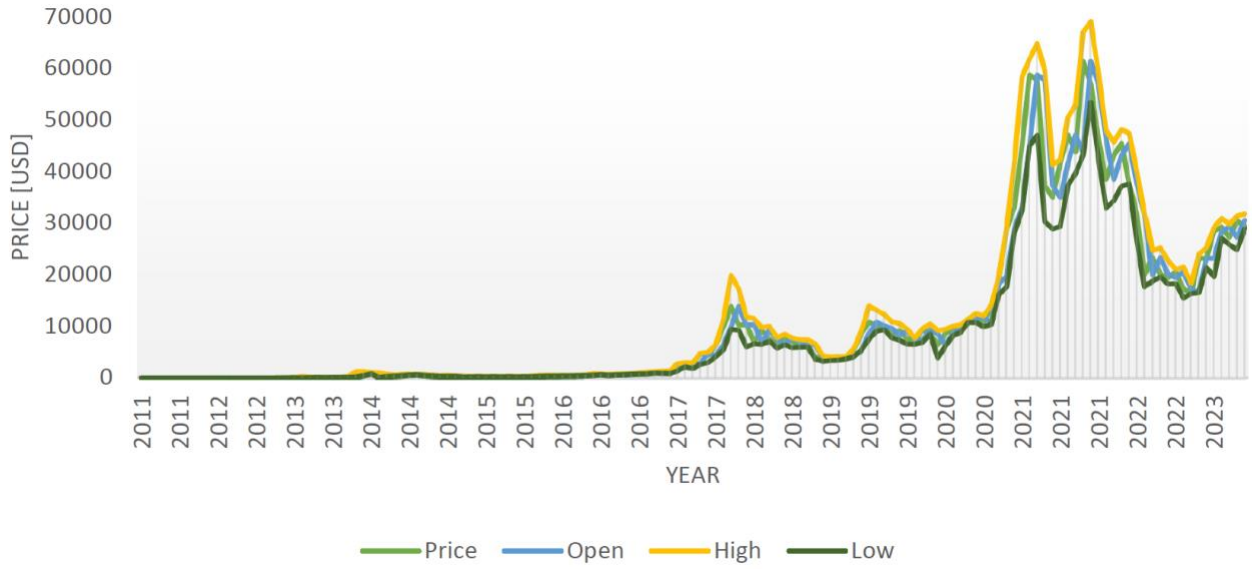


Figura 2. Analiza evoluției prețurilor Bitcoin. Sursa: autorii.

O estimare rapidă a tematicilor incluse în rezumate poate fi obținută utilizând WordCloud. Dimensiunea cuvintelor este direct proporțională cu frecvențele acestora. Cuvântul cu cea mai mare dimensiune a fontului în Figura 3 este blockchain, urmat de Bitcoin, criptomonede, tranzacție, rețea, protocol, sistem, aplicație etc.
Utilizând bibliotecile Python *nltk* și *SpaCy*, s-au aplicat sarcini de preprocesare a textului. În primul rând, s-a eliminat punctuația, literele au fost transformate în litere mici, iar cuvintele de oprire uzuale (*stopwords*), plus câteva cuvinte specifice lucrărilor de cercetare, au fost eliminate.

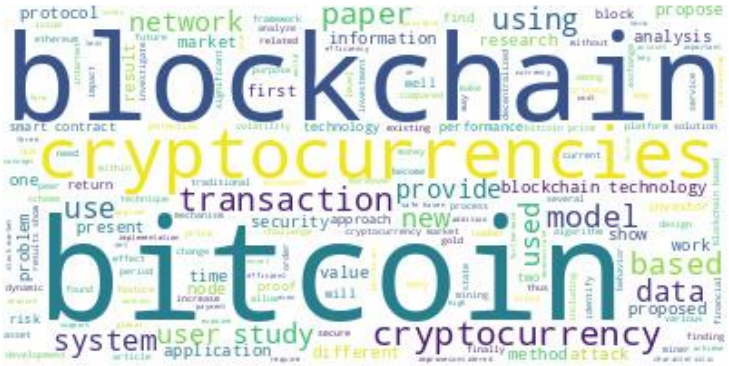


Figura 3. WordCloud aplicat rezumatelor lucrărilor de cercetare indexate în WoS. Sursa: autorii.

Dimensiunea cuvintelor în WordCloud este direct proporțională cu frecvențele acestora. Cuvântul cu cea mai mare dimensiune a fontului în Figura 3 este *blockchain*, urmat de *Bitcoin*, *criptomonede*, *tranzacție*, *rețea*, *protocol*, *sistem*, *aplicație* etc. Utilizând bibliotecile Python *nltk* și *SpaCy*, au fost aplicate sarcini de preprocesare a textului. În primul rând, s-a eliminat punctuația, literele au fost transformate în litere mici, iar cuvintele de oprire (*stopwords*), plus câteva cuvinte specifice lucrărilor de cercetare, au fost eliminate.
Biblioteca *TextBlob* a fost utilizată pentru a evalua polaritatea rezumatelor. Un procent de 86,09% dintre rezumate au fost pozitive, 12,73% au fost negative, iar doar 1,18% au fost neutre. Cu toate acestea, corelația între analiza sentimentelor rezumatelor și prețul Bitcoin (înregistrări lunare descărcate de pe Investing.com: https://www.investing.com/crypto/bitcoin/historical-data, accesat pe 23 iunie 2023) a fost foarte slabă (0,016). Această corelație slabă s-a menținut chiar și atunci când s-a luat în considerare un decalaj de 12 luni între scrierea și publicarea cercetării.
Pentru a realiza predicții, rezumatele au fost vectorizate. Atât *BERT-transformer*, cât și *word2vec* au fost utilizate pentru a transforma textul în vectori numerici, însă predicțiile au fost departe de țintă. Pentru predicție, au fost implementați cinci algoritmi de învățare automată: *random forest*, boost-ul gradient histogramă (*histogram gradient boosting*), boost-ul gradient extrem (*eXtreme Gradient Boosting*), boost-ul gradient luminos (*light gradient boosting*) și regresia liniară, combinați utilizând un regressor prin vot.
Totuși, rezumatele vectorizate nu au putut prezice prețurile Bitcoin. Predicția pentru ultimele 36 de luni este prezentată în Figura 4, evidențiind o acuratețe foarte scăzută și incapacitatea rezumatelor de a fi o variabilă de intrare utilă pentru prezicerea prețurilor Bitcoin. Același nivel scăzut de acuratețe a fost obținut chiar și atunci când s-a utilizat un decalaj de 12 sau 6 luni între scriere și publicare.

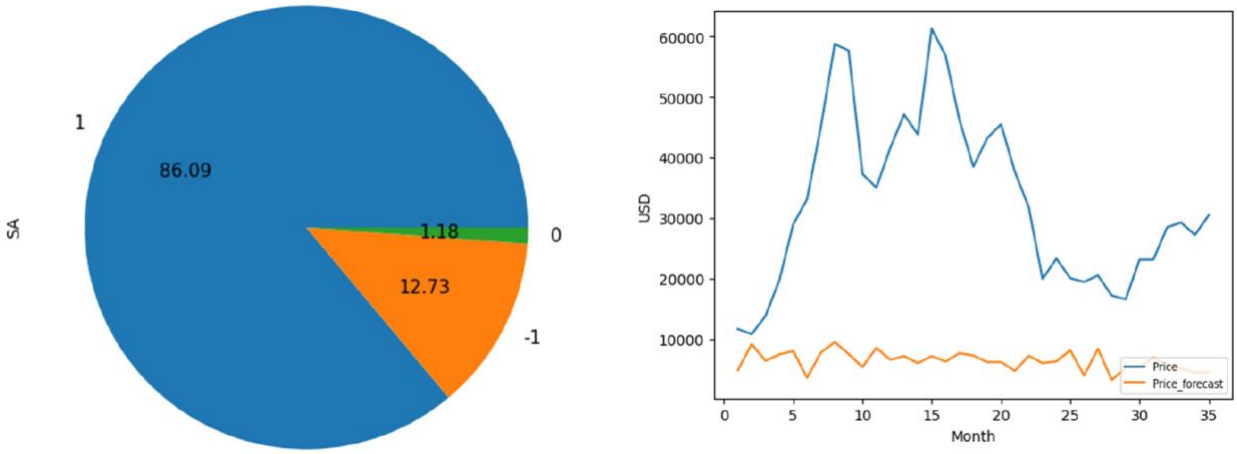
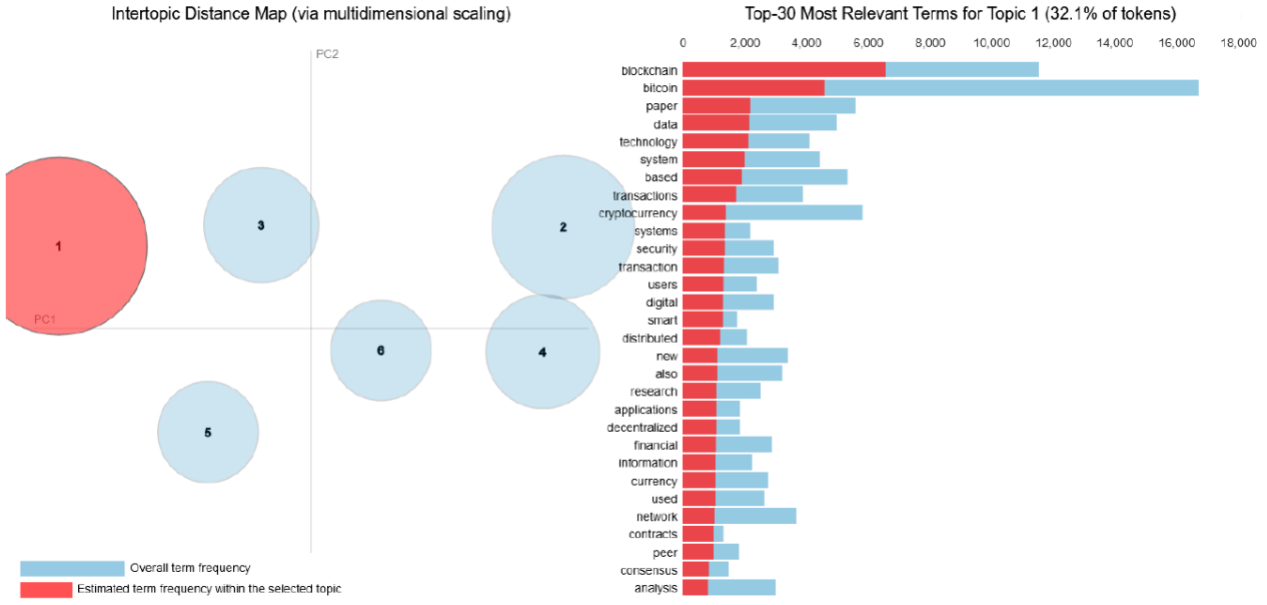
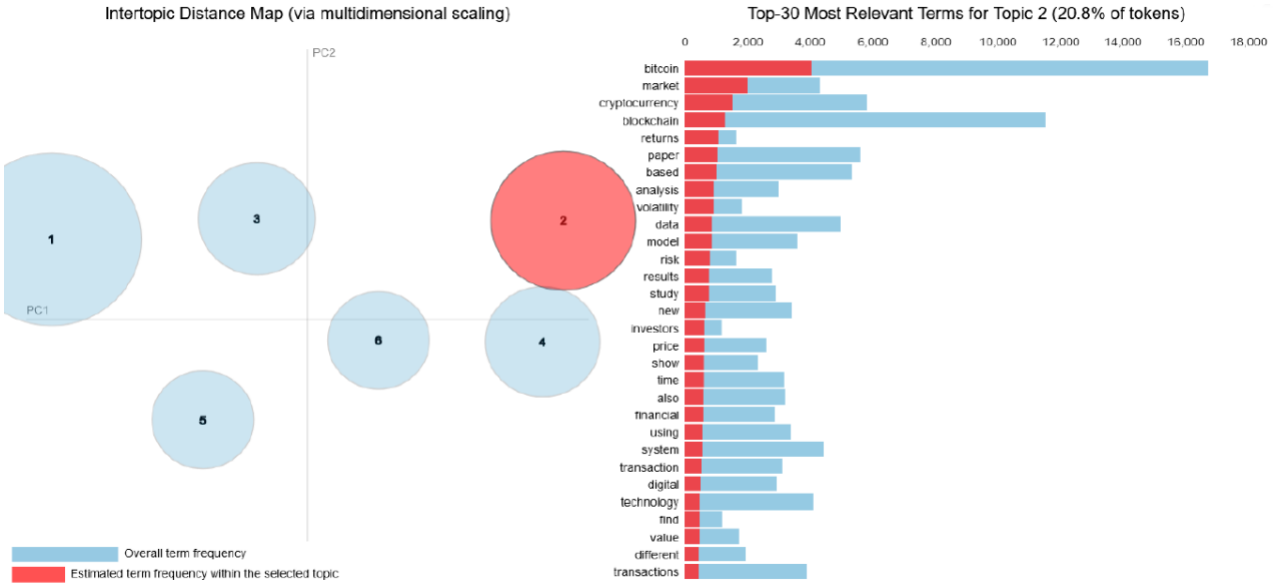


Figura 4. SA și predicția utilizând rezumate vectorizate cu word2vec. Sursa: autorii.

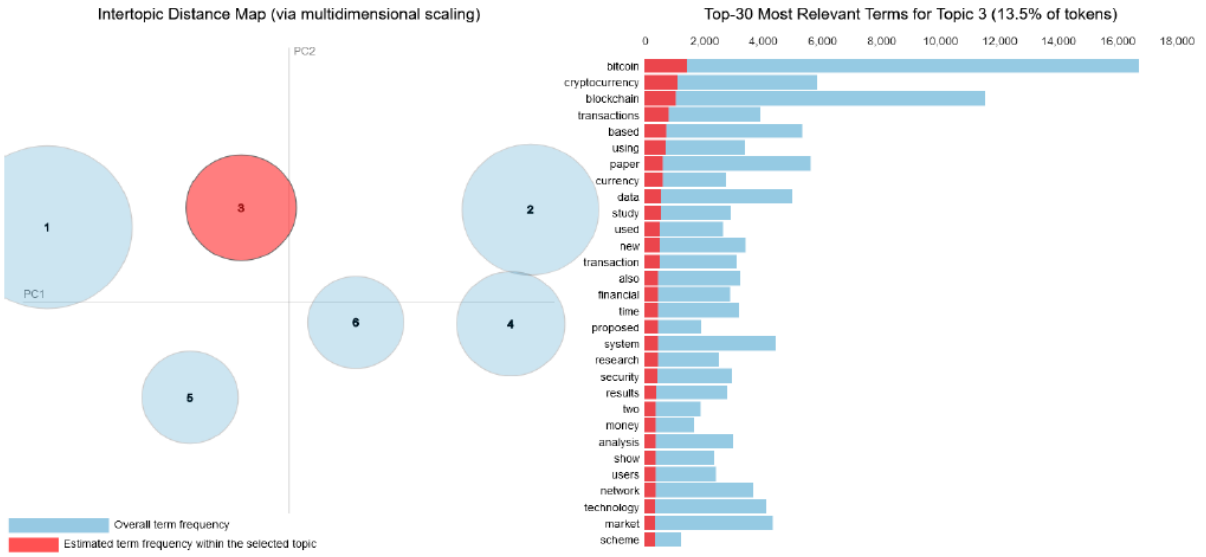
Inițial, au fost setate zece subiecte (K = 10), obținând un scor de coerență a subiectelor = 0,303. *Perplexitatea* este o altă metrică de evaluare care indică cât de surprins este un model de noi observații. Aceasta se calculează ca log-verosimilitatea normalizată a unei noi observații. Totuși, studii recente au arătat că perplexitatea nu este suficient de corelată cu judecata umană [77]. Prin urmare, îmbunătățirea perplexității poate să nu conducă la o interpretare mai bună a subiectelor din perspectivă umană.
Această limitare a scos în evidență o altă metrică menită să modeleze mai bine judecata umană: *coerența subiectelor*, care combină mai mulți factori pentru a evalua coerența între subiecte.
Scorul de coerență de bază pentru modelul LDA implicit este calculat, iar ulterior, s-a creat un test de sensibilitate pentru a determina hiperparametrii optimi ai modelului: numărul de subiecte (*K*), densitatea document-subiect—hiperparametrul Dirichlet alfa (α) și densitatea cuvânt-subiect—hiperparametrul Dirichlet beta (β). Astfel, valorile optime au fost α = 0,01, β = 0,91, cu un K corespunzător de 6 și Cs = 0,385, obținând o îmbunătățire de 27,06%.
Utilizând bibliotecile *gensim* și *pyLDAvis*, au fost identificate șase subiecte bine delimitate cu suprapuneri reduse în diagramele LDA următoare (Figura 5). Primul și cel mai mare subiect, care a inclus 32,1% dintre *tokenuri*, este, fără îndoială, legat de *blockchain*, date, tehnologie distribuită/decentralizată, securitatea sistemului, contracte inteligente și consens (Figura 5a).


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Overall term frequency
Estimated term frequency within the selected topic
Top-30 Most Relevant Terms for Topic 1 (32.1% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
blockchain
bitcoin
paper
data
technology
system
based
transactions
cryptocurrency
systems
security
transaction
users
digital
smart
distributed
new
also
research
applications
decentralized
financial
information
currency
used
network
contracts
peer
consensus
analysis

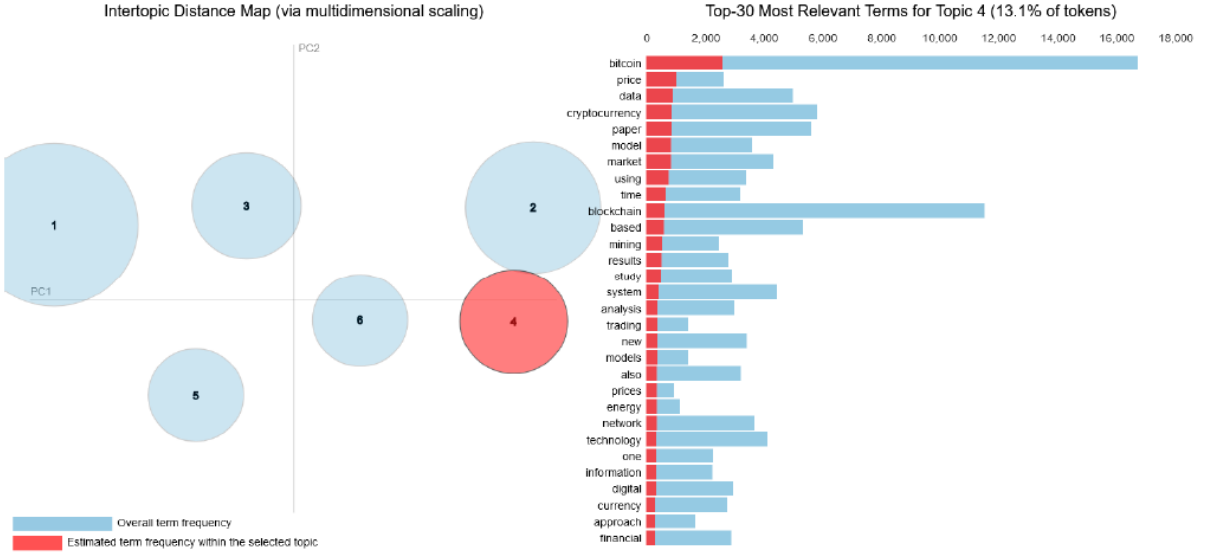
(a)


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Overall term frequency
Estimated term frequency within the selected topic
Top-30 Most Relevant Terms for Topic 2 (20.8% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
bitcoin
market
cryptocurrency
blockchain
returns
paper
based
analysis
volatility
data
model
risk
results
study
new
investors
price
show
time
also
financial
using
system
transaction
digital
technology
find
value
different
transactions

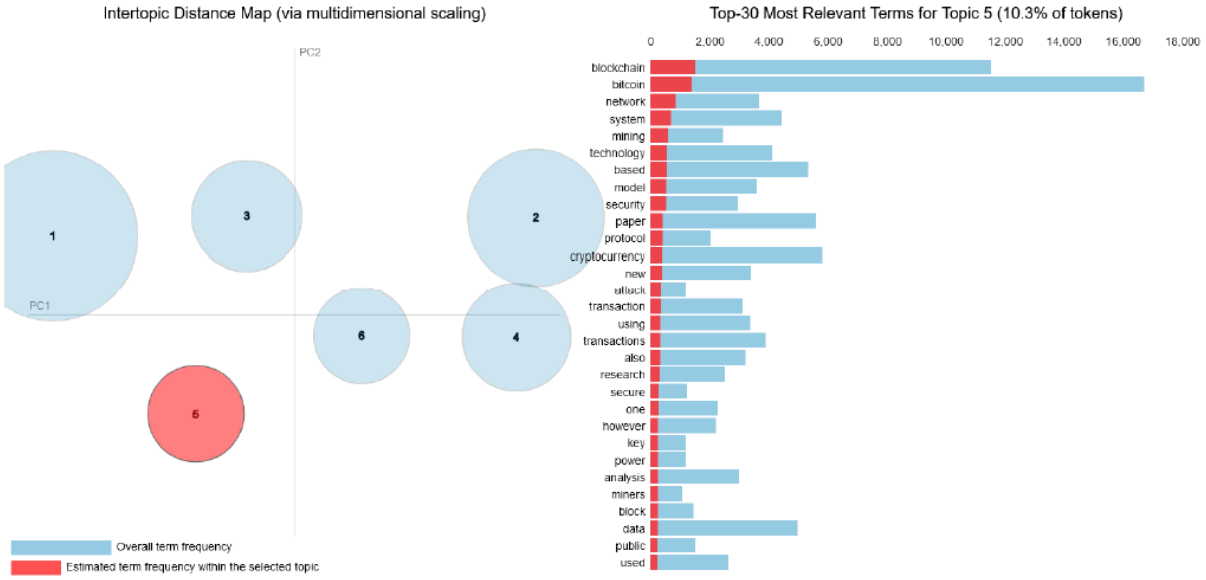
(b)


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Overall term frequency
Estimated term frequency within the selected topic
Top-30 Most Relevant Terms for Topic 3 (13.5% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
bitcoin
cryptocurrency
blockchain
transactions
based
using
paper
currency
data
study
used
new
transaction
also
financial
time
proposed
system
research
security
results
two
money
analysis
show
users
network
technology
market
scheme

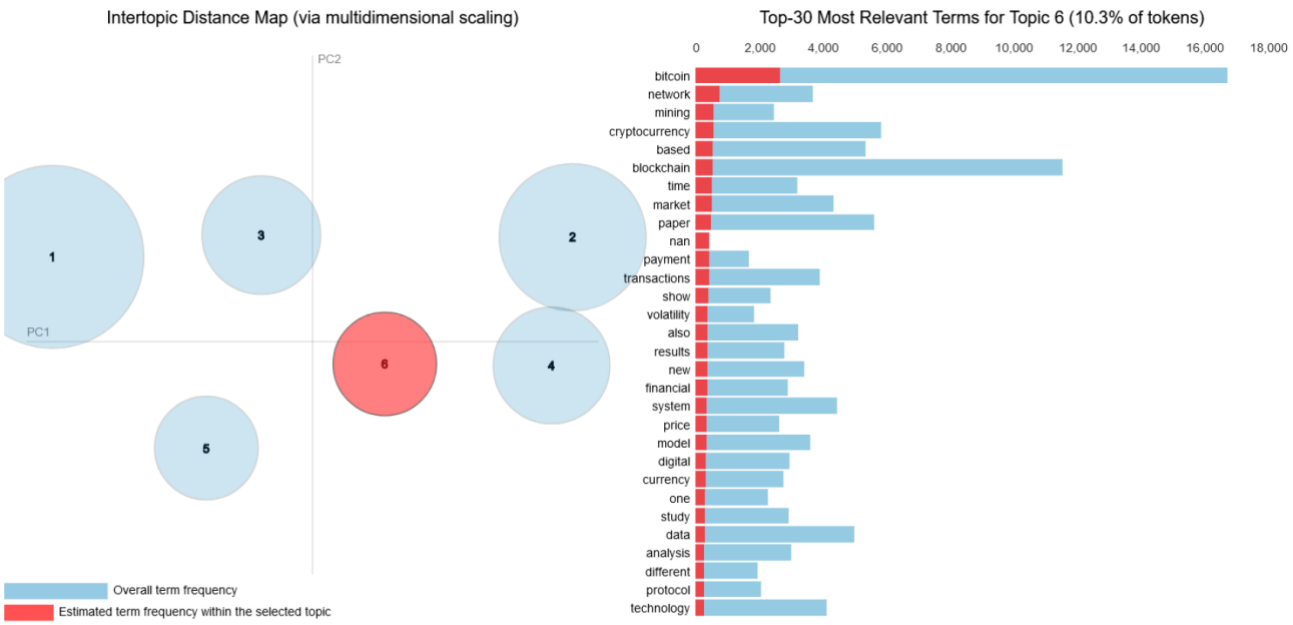
(c)


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Overall term frequency
Estimated term frequency within the selected topic
Top-30 Most Relevant Terms for Topic 4 (13.1% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
bitcoin
price
data
cryptocurrency
paper
model
market
using
time
blockchain
based
mining
results
study
system
analysis
trading
new
models
also
prices
energy
network
technology
one
information
digital
currency
approach
financial

(d)


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Overall term frequency
Estimated term frequency within the selected topic
Top-30 Most Relevant Terms for Topic 5 (10.3% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
blockchain
bitcoin
network
system
mining
technology
based
model
security
paper
protocol
cryptocurrency
new
attack
transaction
using
transactions
also
research
secure
one
however
key
power
analysis
miners
block
data
public
used

(e)



(f)

Figura 5. Diagrame LDA.

Al doilea cel mai mare subiect, care a inclus 20,8% dintre *tokenuri*, se concentrează pe piața criptomonedelor, randamente, volatilitate, investitori în risc, aspecte financiare, preț și valoare (Figura 5b). A treia categorie de subiecte, care a inclus 13,5% dintre *tokenuri*, este legată de tranzacțiile cu criptomonede, date, securitate financiară și a sistemului (Figura 5c).
Al patrulea subiect, care a inclus 13,1% dintre *tokenuri*, se concentrează pe prețul Bitcoin, modele de stabilire a prețurilor, piață, minare și energie (Figura 5d).
Al cincilea subiect, care a inclus 10,3% dintre *tokenuri*, este legat de securitatea rețelei *blockchain*, tehnologia de minare, securitatea sistemului, atacuri, protocoale și mineri (Figura 5e).
Al șaselea subiect este echivalent cu al cincilea în ceea ce privește *tokenurile* și se concentrează pe procesarea minării, piață, plăți, tranzacții, volatilitate și probleme financiare (Figura 5f).

Vizualizarea modelării subiectelor

Modelarea subiectelor este un tip de model statistic utilizat pentru a descoperi „subiectele" abstracte care apar într-o colecție de documente, fiind frecvent utilizată în extragerea de text. În paragrafele următoare, se explică ce reprezintă în general aceste vizualizări:

1. **Harta distanței inter-subiecte** (*inter-topic distance map*, folosind scalare multidimensională):
    - Harta arată diferite subiecte reprezentate ca cercuri numerotate.
    - Dimensiunea fiecărui cerc corespunde prevalenței subiectului în setul de date.
    - Distanța dintre cercuri reprezintă similitudinea între subiecte; cercurile mai apropiate sugerează o similitudine mai mare.
    - Poziționarea pe axe indică modul în care subiectele sunt distribuite.
2. **Top 30 termeni cei mai relevanți pentru subiecte**:
    - Fiecare diagramă cu bare corespunde unui subiect și listează primii 30 de termeni relevanți care caracterizează acel subiect.
    - Lungimea barelor reprezintă frecvența termenului în cadrul subiectului.
    - Barele roșii indică termeni mai specifici subiectului, în timp ce barele albastre indică termeni relevanți, dar nu unici pentru acel subiect.
    - Procentajul *tokenurilor* indică greutatea subiectului în întregul set de date.

Analizarea și interpretarea vizualizărilor modelării subiectelor

Pentru a compara, analiza și interpreta vizualizările modelării subiectelor, se caută modele în prevalența subiectelor, unicitatea termenilor pentru subiecte specifice și relațiile dintre subiecte pe baza proximității lor pe harta distanței inter-subiecte.

Se analizează următoarele aspecte pentru a interpreta diagramele din Figura 5:

(a) **Compararea prevalenței subiectelor**: Se verifică dimensiunea cercurilor pe harta distanțelor pentru a identifica care subiecte sunt mai prevalente. Cercurile mai mari indică o frecvență mai mare a termenilor în setul de date.

(b) **Analizarea relevanței termenilor**: Pentru a înțelege fiecare subiect, se analizează termenii listați în diagramele cu bare. Termenii mai spre dreapta sunt mai relevanți pentru subiect.

(c) **Identificarea termenilor unici**: Barele roșii reprezintă termeni care nu doar că sunt frecvenți, dar sunt și mai specifici subiectului, oferind o mai bună înțelegere a ceea ce distinge un subiect de altul.

(d) **Interpretarea relațiilor dintre subiecte**: Se examinează care subiecte sunt apropiate pe hartă. Subiectele mai apropiate pot împărtăși termeni sau teme comune.

(e) **Contribuția subiectelor**: Se analizează ponderea fiecărui subiect în cadrul setului de date.

Procentul de *tokenuri* indicat lângă fiecare titlu al diagramelor cu bare arată cât de mult din setul de date text a fost atribuit fiecărui subiect, oferind o idee despre importanța sau dominația subiectului în cadrul datasetului. Urmând acești pași, se construiește o înțelegere detaliată a subiectelor și a relațiilor dintre acestea.

LDA (Latent Dirichlet Allocation) este o formă de învățare nesupravegheată, potrivită în mod special pentru analiza și clasificarea volumelor mari de text fără etichete. LDA presupune că fiecare document dintr-un corpus poate fi descris printr-o distribuție de subiecte, iar fiecare subiect poate fi descris printr-o distribuție de cuvinte.

Fiecare subplot din Figura 5 reprezintă o vizualizare derivată din modelarea subiectelor utilizând LDA. Pentru a analiza fiecare subplot, sunt luate în considerare următoarele:

1. **Harta distanțelor dintre subiecte**:
    - Fiecare cerc (*bubble*) reprezintă un subiect diferit descoperit în dataset.
    - În hărțile sau subploatele (a–f), subiectele sunt numerotate de la 1 la 6.
    - Dimensiunea unui cerc reflectă prevalența subiectului. De exemplu, Subiectele 1 și 2 sunt cele mai mari cercuri, sugerând că sunt subiecte proeminente în dataset.
    - Distanța dintre cercuri indică diferențele dintre subiecte; subiectele mai apropiate pot împărtăși termeni comuni (de exemplu, Subiectul 2—piață și Subiectul 4—prețul Bitcoin), în timp ce cele mai îndepărtate sunt mai distincte (de exemplu, Subiectul 2 și Subiectul 5—securitatea rețelei blockchain).
2. **Top 30 cei mai relevanți termeni**:
    - Pentru fiecare subiect, există o diagramă cu bare corespunzătoare care arată primii 30 de termeni relevanți pe baza frecvenței și specificității lor.
    - Axa x indică frecvența termenilor, în timp ce codificarea culorilor (de la roșu la albastru) arată cât de exclusivi sunt termenii pentru subiect (roșul indică o specificitate mai mare).

Din subploatele din Figura 5, se pot face următoarele observații:

1. **Dimensiunea și extinderea subiectelor**:
    - Subiectele mai mari (precum Subiectele 1 și 2) prezintă o răspândire mai mare a termenilor, indicând o discuție mai largă în cadrul datasetului.
    - Subiectele mai mici pot reprezenta domenii de discuție mai nișate.
2. **Specificitatea termenilor**:

- Barele roșii reprezintă termeni care nu doar că sunt frecvenți, dar sunt și foarte specifici subiectului, oferind informații despre unicitatea fiecărui subiect.

3. **Termeni dominanți**:
   - Termenii comuni între diferite subiecte pot reprezenta teme generale în dataset, în timp ce termenii unici conferă un caracter specific unui subiect.
4. **Proximitatea subiectelor**:
   - Subiectele apropiate pe harta distanțelor pot avea conținut care se suprapune sau pot fi legate tematic (doar Subiectele 2 și 4 indică o suprapunere mică, restul subiectelor fiind clar distincte).

După cum se observă în Figura 5, rezultatele LDA arată că datasetul conține o varietate de discuții legate de criptomonede, unele subiecte fiind axate pe aspecte tehnice, altele pe aplicații sau perspective economice și de tranzacționare.

Peste 9100 de lucrări de cercetare despre Bitcoin și criptomonede au fost publicate și indexate pe platforma Web of Science Clarivate până la sfârșitul lunii iulie 2023. În acest studiu, am investigat efectul pe care publicațiile îl pot avea asupra prețurilor Bitcoin sau relația dintre opinia comunității academice și prețurile Bitcoin. De asemenea, studiul a avut ca scop identificarea principalelor subiecte legate de criptomonede în general și de Bitcoin în mod special.

Modelul LDA a fost aplicat pentru a identifica subiectele relevante din datele nestructurate ale rezumatelor lucrărilor de cercetare. Hiperparametrii modelului au fost ajustați pentru a determina numărul optim de subiecte.

Revenind la întrebările de cercetare expuse în introducere, după fuzionarea și analiza celor două seturi de date, am reușit să oferim răspunsuri bazate pe date:

Subiectele predominante legate de criptomonede sunt: blockchain, piață, tranzacții, preț, securitatea rețelei și procesul de minare. Perplexitatea și coerența subiectelor au fost calculate pentru a stabili numărul optim de subiecte.

Ajustarea numărului de subiecte și a parametrilor metodei LDA s-a dovedit a fi un proces consumator de timp (aproximativ 12 ore) pentru procesarea și testarea diferitelor scenarii. Prin ajustarea modelului LDA, am obținut numărul optim de subiecte care oferă cea mai bună coerență și cele mai bune scoruri de perplexitate.

Rezumatele au fost analizate pentru a evalua potențialul lor de a prezice prețurile Bitcoin. Analiza sentimentului (*Sentiment Analysis - SA*) a fost realizată, iar relația dintre SA și prețurile Bitcoin a fost examinată. Investigațiile privind polaritatea au arătat că majoritatea rezumatelor exprimă un sentiment pozitiv (86%), în timp ce 12% au fost negative, iar mai puțin de 2% au fost neutre.

Totuși, concluziile sugerează că publicațiile academice au un impact minim asupra prețurilor Bitcoin, demonstrând o legătură slabă. Acest lucru indică faptul că fluctuațiile prețurilor Bitcoin nu sunt influențate de cercetările academice din acest domeniu specific.

Pe baza corelației Pearson, relația dintre analiza sentimentului extrasă din rezumate și prețurile Bitcoin este foarte slabă sau aproape inexistentă. De asemenea, potențialul scrierilor academice de a influența mișcările prețurilor Bitcoin este limitat.

Cele șase subiecte identificate în acest studiu reprezintă principalele direcții de cercetare sau teme relevante legate de Bitcoin și criptomonede. Totuși, acestea evidențiază și unele lacune în literatura de specialitate:

- **Procesul de minare**, care influențează consumul de energie și aspectele de mediu, necesită cercetări suplimentare. Deoarece Bitcoin reprezintă aproximativ jumătate din cererea de energie a tuturor criptomonedelor, este nevoie de mecanisme de consens mai eficiente din punct de vedere energetic.

- Măsuri pentru **combaterea spălării banilor**, comportamente neadecvate și evenimente legate de **criminalitatea criptomonedelor** ar trebui să fie studiate mai aprofundat.

LLM-Based Applications in the Financial Sector

A. Progrese în Modelarea Riscului de Credit

În cadrul acestei etape a proiectului de cercetare, activitatea s-a concentrat pe dezvoltarea și implementarea unui sistem avansat de modelare a riscului de credit utilizând Modele Mari de Limbaj (Large Language Models - LLMs). Această abordare reprezintă o evoluție semnificativă în domeniul evaluării riscului de credit, combinând tehnici tradiționale cu capacitățile avansate ale inteligenței artificiale.

Obiective și Metodologie

Principalele obiective urmărite în dezvoltarea acestui sistem au fost:

- Îmbunătățirea preciziei scorurilor de credit prin integrarea datelor nestructurate
- Dezvoltarea unui model predictiv robust pentru evaluarea riscurilor
- Implementarea unei soluții transparente și explicabile
- Optimizarea procesului decizional în acordarea creditelor

Metodologia adoptată s-a bazat pe o abordare sistematică, incluzând:

1. Colectarea și Preprocesarea Datelor:
   - Utilizarea datelor istorice despre împrumuturi și comportamentul clienților
   - Integrarea informațiilor din rapoartele de credit
   - Colectarea și analiza datelor din social media
   - Normalizarea și standardizarea seturilor de date
2. Dezvoltarea Modelului:
   - Implementarea unui pipeline de procesare a datelor
   - Antrenarea modelelor predictive folosind LLM-uri
   - Integrarea tehnicilor de învățare automată pentru extragerea caracteristicilor
   - Validarea și optimizarea performanței modelului

Primele rezultate ale implementării acestui sistem indică o îmbunătățire semnificativă în acuratețea predicțiilor și o reducere substanțială a ratei de false negative în identificarea riscurilor de credit. În cadrul dezvoltării modelului de risc de credit, am implementat o serie de analize și experimente pentru validarea eficacității sistemului. Această abordare metodologică a permis o înțelegere aprofundată a factorilor care influențează riscul de credit și a modului în care aceștia interacționează.

Analiza Univariată și Multivariată

1. Analiza Distribuțiilor:
   - Evaluarea comportamentului scorurilor de credit pentru diferite segmente de clienți
   - Studiul tiparelor în istoricul de plăți și comportamentul financiar
   - Analiza tendințelor în utilizarea creditelor și capacitatea de rambursare
   - Identificarea factorilor determinanți în comportamentul de plată
2. Analiza Corelațiilor:
   - Analiza matricei de corelație
   - Generarea vizualizărilor pentru relațiile între variabile
   - Studiul interdependențelor între factorii de risc
   - Evaluarea impactului variabilelor socio-economice

Dezvoltarea Modelului Predictiv

Implementarea a fost realizată în mai multe etape:

1. Modelul de Bază:

- Implementarea regresiei logistice ca punct de referință
- Evaluarea performanței pe setul de date de test
- Identificarea limitărilor și ariilor de îmbunătățire
- Stabilirea unui baseline pentru comparații ulterioare

2. Modelul Îmbunătățit:
- Dezvoltarea unei suite de modele (DT, RF, SVM, kNN, XGBoost)
- Implementarea tehnicilor avansate de regularizare
- Optimizarea parametrilor pentru performanță cât mai bună
- Validarea rezultatelor prin multiple iterații

Optimizări Implementate

Pentru îmbunătățirea performanței generale, am implementat:

- Tehnici avansate de selecție a parametrilor
- Metode de validare încrucișată pentru robustețe
- Strategii de gestionare a dezechilibrului din date
- Mecanisme de ajustare dinamică a parametrilor

Rezultatele obținute indică o îmbunătățire semnificativă în capacitatea modelului de a prezice riscul de credit, demonstrând potențialul acestei abordări în contextul evaluării riscului financiar. Implementarea sistemului de modelare a riscului de credit în industrie a reprezentat o etapă crucială în validarea eficacității soluției dezvoltate. Procesul de implementare a necesitat o abordare graduală, începând cu testarea în medii controlate și progresând către integrarea completă în fluxurile operaționale existente.

Procesul de validare în industrie s-a desfășurat prin evaluarea comparativă a scorurilor de credit generate de noul sistem în raport cu metodele tradiționale. Această analiză s-a extins pe multiple segmente de clienți și în diverse condiții de piață, permițând o evaluare comprehensivă a robusteții modelului. Monitorizarea performanței în timp real a oferit perspective pentru optimizarea continuă a sistemului și adaptarea la schimbările din comportamentul clienților. Integrarea cu sistemele bancare existente a reprezentat o provocare semnificativă, necesitând dezvoltarea unor interfețe specializate și implementarea unor protocoale de securitate. Un accent deosebit a fost pus pe asigurarea scalabilității soluției și optimizarea timpilor de procesare, aspecte esențiale pentru utilizarea în mediul operațional bancar.

În contextul actual al reglementărilor stricte din domeniul bancar, am acordat o atenție specială asigurării conformității cu cerințele GDPR și alte reglementări relevante (EU AI Act). Sistemul a fost proiectat cu mecanisme incorporate de audit și trasabilitate, oferind transparență completă asupra procesului decizional. Documentarea detaliată a proceselor și procedurilor a fost realizată pentru a facilita atât conformitatea continuă, cât și potențiale audituri viitoare. Managementul datelor sensibile a reprezentat o prioritate majoră în dezvoltarea sistemului. Au fost implementate protocoale avansate de criptare și un sistem granular de control al accesului, asigurând protecția informațiilor confidențiale ale clienților. Monitorizarea continuă și logarea activităților oferă un nivel suplimentar de securitate și control.

Implementarea sistemului a generat beneficii tangibile pentru instituțiile financiare care l-au adoptat. Timpul necesar pentru procesarea cererilor de credit a fost redus semnificativ, permițând o creștere a eficienței operaționale în departamentele de credit. Consistența îmbunătățită în evaluarea riscurilor a condus la decizii mai bine fundamentate și la o reducere a expunerii la risc. În plus, automatizarea proceselor a rezultat într-o diminuare semnificativă a costurilor asociate cu evaluarea manuală a cererilor de credit.

B. Dezvoltarea și Implementarea Cadrului RAG pentru Auditul Bancar

În cadrul acestei etape a proiectului, am dezvoltat și implementat un cadru inovativ de tip RAG (Retrieval-Augmented Generation) destinat optimizării proceselor de audit bancar. Această soluție reprezintă un răspuns direct la provocările actuale din domeniul auditului bancar, unde volumul mare de date și complexitatea informațiilor necesită abordări automatizate și inteligente.

Cadrul RAG dezvoltat se bazează pe o arhitectură modulară, proiectată să faciliteze procesarea eficientă a documentelor și extragerea informațiilor relevante. Arhitectura sistemului este structurată pe patru componente principale, fiecare cu roluri și funcționalități specifice în procesul de audit. Frontend-ul sistemului reprezintă interfața principală prin care auditorii interacționează cu platforma, oferind acces la funcționalitățile de procesare a documentelor și vizualizare a rezultatelor. Această componentă a fost dezvoltată cu accent pe ușurința în utilizare și eficiența în navigare, permițând auditorilor să își desfășoare activitatea într-un mod intuitiv și eficient.

Stratul de procesare încorporează algoritmi avansați pentru analiza întrebărilor și extragerea cuvintelor cheie. Această componentă este responsabilă pentru înțelegerea și interpretarea corectă a solicitărilor auditorilor, asigurând că sistemul poate furniza răspunsuri precise și relevante. Procesul include analiza semantică a întrebărilor și identificarea conceptelor cheie care ghidează căutarea informațiilor. Componenta de date gestionează stocarea și organizarea eficientă a documentelor și informațiilor financiare. Aceasta include baze de date specializate pentru conturi și monografii contabile, precum și sisteme de indexare avansată care facilitează recuperarea rapidă a informațiilor relevante. Structura de date a fost proiectată să suporte volume mari de informații, menținând în același timp performanța și acuratețea sistemului.

Stratul GPT, integrat în arhitectura sistemului, este responsabil pentru generarea răspunsurilor și interpretarea contextului. Această componentă utilizează modele avansate de limbaj pentru a procesa și genera răspunsuri coerente și precise la întrebările auditorilor, asigurând că informațiile furnizate sunt atât precise, cât și ușor de înțeles. Procesul de implementare a cadrului RAG pentru auditul bancar a necesitat o abordare metodică și structurată, cu accent pe integrarea eficientă a tuturor componentelor. Această secțiune detaliază aspectele tehnice și operaționale ale implementării, precum și modul în care diferitele module ale sistemului interacționează pentru a oferi o soluție completă. Arhitectura sistemului este prezentată în Figura 6.

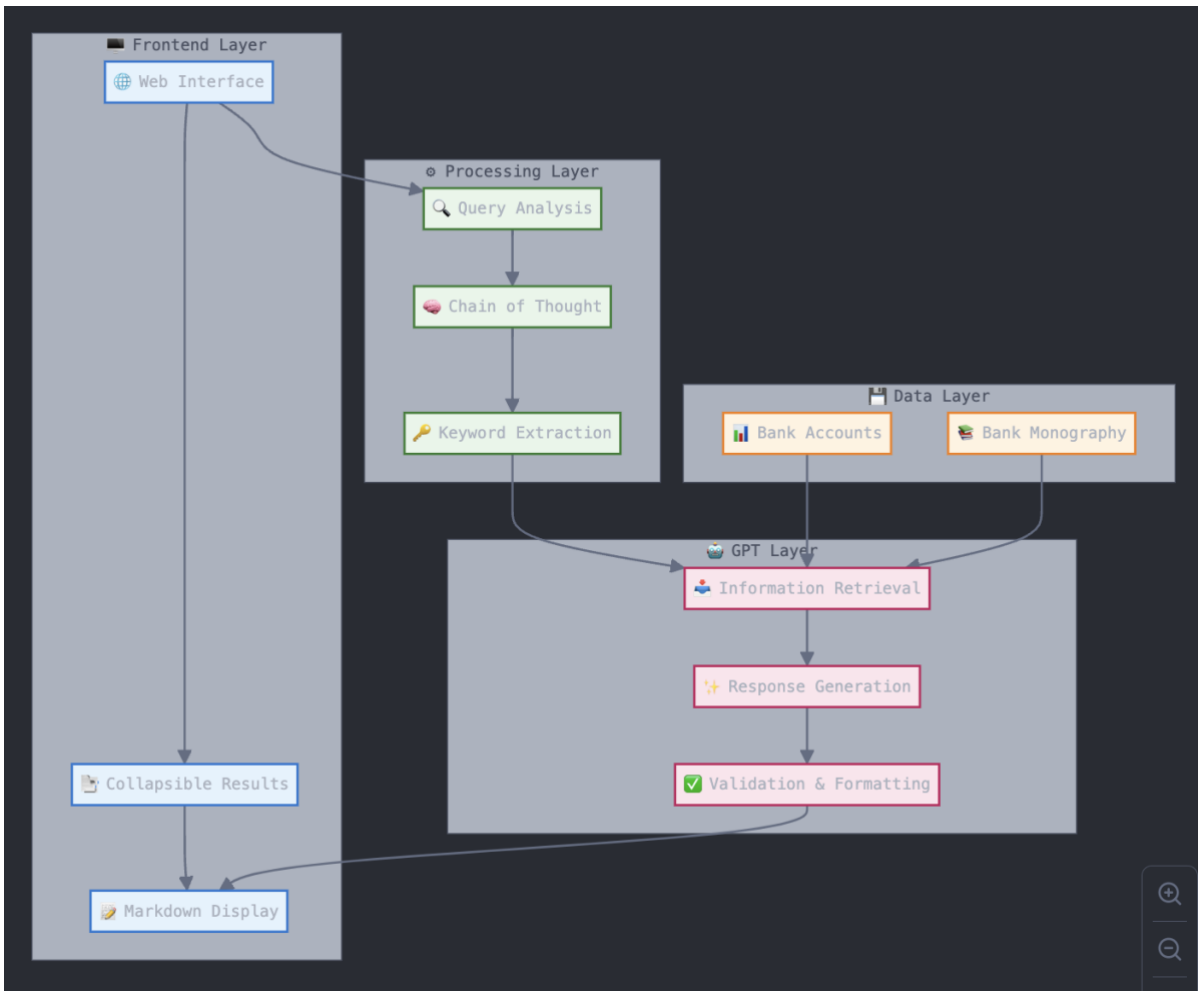


Figura 6. Arhitectura Sistemului RAG

Pipeline-ul de date reprezintă prima componentă a sistemului, fiind responsabil pentru procesarea și transformarea documentelor în formate utilizabile. Implementarea acestuia a inclus dezvoltarea unui sistem robust de OCR (Optical Character Recognition) care permite extragerea precisă a informațiilor din diverse tipuri de documente bancare. Sistemul a fost optimizat pentru a gestiona atât documente structurate (precum rapoarte financiare standardizate), cât și documente semi-structurate (precum notele explicative sau documentația suport). Pentru asigurarea acurateței datelor extrase, am implementat multiple niveluri de validare și verificare. Acestea includ verificări automate ale consistenței datelor, sisteme de detectare a anomaliilor și mecanisme de reconciliere cu bazele de date existente. Procesul de validare este esențial pentru menținerea integrității informațiilor și asigurarea fiabilității rezultatelor auditului.

Un aspect crucial al implementării a fost integrarea cu sistemele bancare existente. Aceasta a necesitat dezvoltarea unor interfețe specializate și protocoale de comunicare sigure care să permită accesul la datele necesare fără a compromite securitatea sistemului. Am implementat mecanisme de sincronizare în timp real care asigură că informațiile procesate sunt întotdeauna actuale și relevante pentru procesul de audit. Interfața cu PowerBI a fost dezvoltată pentru a oferi vizualizări interactive și dashboarduri comprehensive care permit auditorilor să analizeze și să interpreteze datele într-un mod eficient. Această componentă facilitează identificarea rapidă a tiparelor și anomaliilor în datele financiare, contribuind la eficientizarea procesului de audit.

Pentru a asigura performanța a sistemului, am implementat diverse strategii de optimizare, inclusiv tehnici de caching pentru interogările frecvente și sisteme de indexare avansată pentru accesul rapid

la date. Aceste optimizări au condus la reduceri semnificative ale timpilor de răspuns și la o experiență mai fluidă pentru utilizatorii finali. Implementarea cadrului RAG în contextul auditului bancar a generat rezultate semnificative, demonstrând potențialul acestei tehnologii în transformarea proceselor tradiționale de audit.

Evaluarea sistemului în condiții reale de operare a evidențiat o serie de îmbunătățiri substanțiale în eficiența și acuratețea proceselor de audit. Timpul necesar pentru analiza documentelor a fost redus considerabil, permițând auditorilor să se concentreze pe aspectele care necesită expertiză umană. Sistemul a demonstrat o capacitate remarcabilă de a procesa și analiza volume mari de date într-un timp semnificativ mai scurt comparativ cu metodele tradiționale. De asemenea, am observat o creștere sesizabilă în calitatea și consistența analizelor efectuate. Automatizarea proceselor repetitive și standardizarea metodologiilor de analiză au condus la reducerea erorilor umane și la o mai bună trasabilitate a procesului de audit. Capacitatea sistemului de a păstra și organiza documentația în mod structurat facilitează revizuiri ulterioare și asigură conformitatea cu cerințele de reglementare.

Rezultatele obținute în această etapă a proiectului demonstrează potențialul semnificativ al tehnologiilor bazate pe LLM-uri și cadre RAG în modernizarea sectorului financiar-bancar. Atât modelarea riscului de credit, cât și implementarea sistemului RAG pentru audit au generat îmbunătățiri măsurabile în eficiență și acuratețe. Aceste realizări pun bazele pentru dezvoltări viitoare și demonstrează potențialul transformativ al inteligenței artificiale în sectorul financiar.

Rezultatele obținute în cadrul acestui proiect de cercetare demonstrează potențialul transformativ al tehnologiilor bazate pe inteligență artificială în sectorul financiar-bancar. Prin dezvoltarea și implementarea unor soluții inovatoare pentru modelarea riscului de credit și optimizarea proceselor de audit, am contribuit la avansarea cunoașterii în domeniul aplicațiilor practice ale Modelelor Mari de Limbă în sectorul financiar. Sistemele dezvoltate oferă beneficii tangibile, de la îmbunătățirea preciziei în evaluarea riscurilor până la eficientizarea semnificativă a proceselor de audit. Aceste realizări deschid noi perspective pentru cercetări viitoare și implementări practice în domeniul serviciilor financiare, contribuind la modernizarea și digitalizarea sectorului bancar din România.

Această cercetare a fost realizată cu sprijinul financiar al Academiei Oamenilor de Știință din România (AOSR). Aducem mulțumiri AOSR pentru suportul acordat, care a făcut posibilă dezvoltarea acestor soluții inovatoare și obținerea rezultatelor prezentate în acest proiect. Sprijinul AOSR a fost esențial în avansarea cunoașterii în domeniul aplicațiilor inteligenței artificiale în sectorul financiar și în formarea unei noi generații de cercetători în acest domeniu.