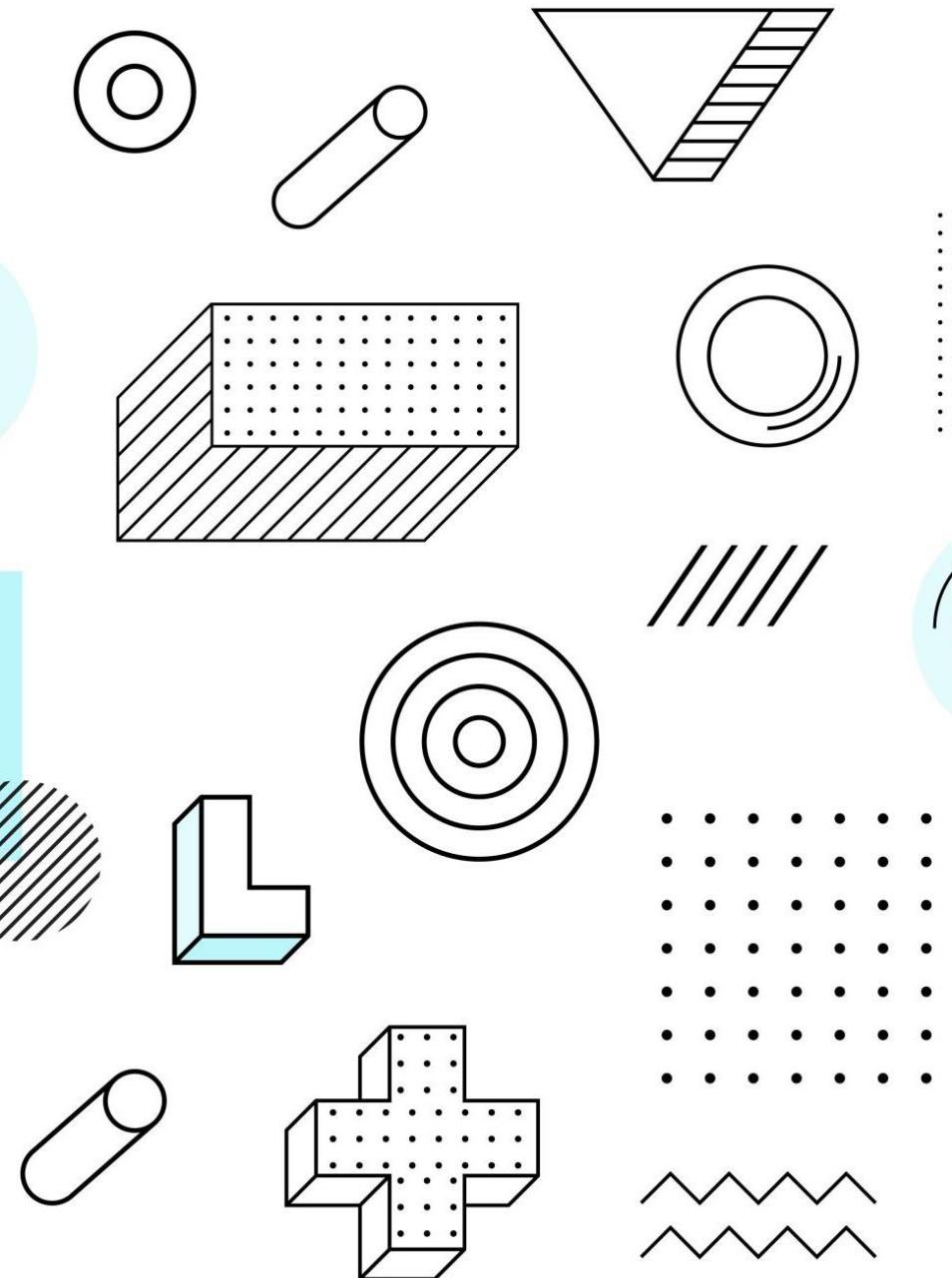


# Analiza euristică a diferențelor lingvistice dintre limba română scrisă în România și cea scrisă în Moldova

Tindeche Alexandru  
Florin Dinu

Proiect coordonat de: Sergiu Nisioi



# Abordare

- ❖ Am facut mai multe analize atat independente cat si dependente de context
- ❖ Analize independente de context:
  - ❖ Analiza statistica a articolelor din Moldova si cele din Romania (media, mediana, deviatia standard pe lungimea articolelor)
  - ❖ Analiza precizata mai sus, dar impartita pe categoriile de articole (politica, economie, social, sport, cultura, stiri, divertisment, tehnologie, sanatate, auto, turism, lifestyle, educatie, locale, justitie, diverse)
- ❖ Antrenarea unor clasificatori binari pentru a vedea daca putem distinge intre cele doua clase
  - ❖ Un clasificator care are ca features direct cuvintele
  - ❖ Un clasificator care are ca features partile de vorbire (aici am folosit spacy si NLPCube de la Adobe)
- ❖ Generarea unor dependency trees si analiza celor mai intalnite structure statistice din fiecare clasa
- ❖ Constructia unui model care generaza propozitii in limba romana din Romana si Moldova, separate

# 1. Colectarea datelor

- ❖ Am ales 10 publicatii din Romania si Moldova si am extras minim 300 de articole per ziari, din categoriile mentionate in slide-ul anterior, folosind un script de automatizare in Python, cu libraria Selenium
- ❖ Pentru fiecare articol am extras atat textul cat si metadata despre articol, precum data publicarii, categoria si titlul
- ❖ Am pus aceste date intr-o baza de date atat SQLite, pentru accesarea mai simpla din cod, cat si in fisiere .txt

```
1 Ce performanță a stabilit Vinicius, după ce a marcat de trei ori în "El Clasico"
2 Sport
3 | Sport / 15 ianuarie 2024, 18:43 / 116
4
5 Real Madrid a castigat Supercupa Spaniei, dupa ce in finala a invins-o pe Barcelona, scor 4-1.Madrilenii au avut un start de meci excelent, cu doua goluri marcate in doar trei minute de Vinicius. Brazilianul a reusit hat-trick-ul in minutul 39, dupa ce a transformat un penalty. Partida din Supercupa Spaniei este cel de-al 15-lea "El Clasico" pentru Vinicius. Brazilianul a reusit o performanta importanta pentru cariera sa. Starul lui Real Madrid este cel de-al 16-lea jucator din istorie care a marcat de trei ori intr-un "El Clasico". Pe aceasta lista se mai afla: Jaime Lazcano, Joan Ramon i Pera, Ventora, Jesus Narro, Cesar, Evaristo de Macedo, Amancio, Ferenc Puskas, Ivan Zamorano, Fernando Sanudo, Gary Lineker, Romario, Luis Suarez, Karim Benzema si Lionel Messi. Este important de precizat ca Messi este singurul fotbalist dintre cei enumerati care a reusit aceasta performanta de doua ori in cariera, potrivit Fanatik. Partida din Supercupa Spaniei este cel de-al 15-lea "El Clasico" pentru Vinicius. Brazilianul a reusit o performanta importanta pentru cariera sa. Starul lui Real Madrid este cel de-al 16-lea jucator din istorie care a marcat de trei ori intr-un "El Clasico". Pe aceasta lista se mai afla: Jaime Lazcano, Joan Ramon i Pera, Ventora, Jesus Narro, Cesar, Evaristo de Macedo, Amancio, Ferenc Puskas, Ivan Zamorano, Fernando Sanudo, Gary Lineker, Romario, Luis Suarez, Karim Benzema si Lionel Messi. Este important de precizat ca Messi este singurul fotbalist dintre cei enumerati care a reusit aceasta performanta de doua ori in cariera, potrivit Fanatik. Pe aceasta lista se mai afla: Jaime Lazcano, Joan Ramon i Pera, Ventora, Jesus Narro, Cesar, Evaristo de Macedo, Amancio, Ferenc Puskas, Ivan Zamorano, Fernando Sanudo, Gary Lineker, Romario, Luis Suarez, Karim Benzema si Lionel Messi. Este important de precizat ca Messi este singurul fotbalist dintre cei enumerati care a reusit aceasta performanta de doua ori in cariera, potrivit Fanatik. Este important de precizat ca Messi este singurul fotbalist dintre cei enumerati care a reusit aceasta performanta de doua ori in cariera, potrivit Fanatik.
```

## 2. Statistici independente de context

- ❖ Am incarcat toate articolele din Moldova si Romania (bineintele, separat)
- ❖ Am incarcat toate cuvintele in doua array-uri si am facut urmatoarele statistici: mean, median si standard deviation
- ❖ Apoi am grupat articolele pe categorii si repetam aceleasi statistici, pe care le salvam intr-un fisier .CSV

```
Mean for Moldova: 5.4549551186625935
Median for Moldova: 5.0
Standard deviation for Moldova: 3.522459223910178
Mean for Romania: 5.386629254503347
Median for Romania: 5.0
Standard deviation for Romania: 3.5738912536493164
```

```
Statistics for Romania-----  
Statistics for Stiri: [5.320666800098534, 5.0, 3.339209102945649]  
Statistics for Locale: [5.141808730219402, 5.0, 3.108392048018299]  
Statistics for Lifestyle: [5.252813852813853, 5.0, 3.18398337895951]  
Statistics for Diverse: [5.403568371698442, 5.0, 3.2341992568984796]  
Statistics for Sport: [5.17363428442878, 5.0, 3.167829978822963]  
Statistics for Economie: [5.459404719971821, 5.0, 3.420038199649928]  
Statistics for Politica: [5.441450239570299, 5.0, 3.520399602827184]  
Statistics for Social: [5.5965387724052675, 5.0, 3.5990526074396705]  
Statistics for Cultura: [5.613301473516527, 5.0, 3.412180709262932]  
Statistics for Moldova-----  
Statistics for Diverse: [5.389331785646136, 5.0, 3.337188140048604]  
Statistics for Politica: [5.671197219396015, 5.0, 3.573285884289556]  
Statistics for Economie: [5.448890059862628, 5.0, 3.4213425775147344]  
Statistics for Social: [5.4417094514277045, 5.0, 3.4452988390847317]  
Statistics for Stiri: [5.432707643673019, 5.0, 3.3304712125911515]  
Statistics for Tehnologie: [5.118650979263363, 5.0, 3.1068769784475414]  
Statistics for Locale: [5.600110685296708, 5.0, 3.6287914821587193]  
Statistics for Divertisment: [5.525937628304682, 5.0, 3.4653434666004523]  
Statistics for Cultura: [5.452595040845193, 5.0, 5.6640341942141]
```

# Concluzii

- ❖ Nu exista mari diferente intre mediile si deviatiile standard pe cele 2 limbi

## 2. Clasificatori binari: cu feature-uri cuvintele in sine:

1. Am facut un clasificator binar de tip Logistic Regression cu l2, un vectorizer cu urmatoarele setari:  
`TfidfVectorizer(min_df=3, max_df=0.7, max_features=10000, stop_words=get_stop_words('ro'))`  
si un grid search pentru automatizarea hyperparameter tuning
2. Datele sunt stratified split
3. Pentru a nu face modelul sa fie biased, am ales sa ii dam cate 916 articole din fiecare clasa (Romania – Moldova)

```
sss = StratifiedShuffleSplit(n_splits=10, test_size=0.1, random_state=11)
text_clf = Pipeline(steps=[
    ('tfidf', TfidfVectorizer(min_df=3, max_df=0.7, max_features=10000, stop_words=get_stop_words('ro'))),
    ('clf', LogisticRegression(penalty='l2'))
], verbose=True)
parameters = {
    'tfidf__ngram_range': [(1, 4)],
    'tfidf__use_idf': (True, False),
    'clf__C': (0.1, 1, 10),
}
gs_clf = GridSearchCV(text_clf, parameters, cv=sss, n_jobs=-1, verbose=1)
```

# Concluzii

- ❖ Modelul are o acuratete de 99% atat pe romana scrisa in Romania cat si pe cea in Moldova
- ❖ Acest lucru poate inseamna ca cele doua limbi difera indeajuns de mult pentru ca un clasificator binar sa le differentieze bine

```
Mean score: 0.9901960784313726
Mean grid search score: 0.9970588235294118
Best parameters: {'clf_c': 1, 'tfidf_ngram_range': (1, 4), 'tfidf_use_idf': True}
Best score: 0.9989130434782609
Classification report:
precision    recall  f1-score   support
moldova      0.99     1.00     1.00     102
romana       1.00     0.99     1.00     102
accuracy          1.00     1.00     1.00     204
macro avg       1.00     1.00     1.00     204
weighted avg    1.00     1.00     1.00     204
```

## 2. Clasificatori binari: cu feature-uri parti de vorbire

1. Am incercat construirea a 2 modele care au ca input feature POS dependente de context dar si independente de context
2. Pentru cele dependente de context am folosit biblioteca spaCy
3. Am extras partile de vorbire si am retinut primele 10 cele mai frecvente parti de vorbire
4. Ca model este cel folosit mai devreme, Logistic Regression cu L2
5. Pentru cele independente de context am incercat folosirea bibliotecii NLP-Cube de la Adobe, dar din pacate ar necesita prea multe resurse computationale, pe care noi nu le avem la dispozitie

# Concluzii

- ❖ Modelul dependent de context are scor f1 de 1 pe Moldova si 0.98 pe Romania
- ❖ Acest lucru poate insemna ca cele doua limbi difera si contextual destul de mult pentru a putea fi diferențiate “lejer”

```
Mean score: 0.9980039920159681
Mean grid search score: 0.9928057553956835
Best parameters: {'clf__C': 10, 'tfidf__ngram_range': (1, 4), 'tfidf__use_idf': False}
Best score: 0.9970005546311702
Classification report:
      precision    recall  f1-score   support

  moldova       0.99     1.00     1.00      631
  romana        1.00     0.97     0.98      203

  accuracy          -         -     0.99      834
  macro avg       1.00     0.99     0.99      834
  weighted avg    0.99     0.99     0.99      834
```

### 3. Dependency trees

1. Am incarcat articolele si am folosit libraria spaCy pentru a calcula dependency trees pentru fiecare dintre cele 2 clase
2. Am parcurs fiecare fraza din cele doua clase si pentru fraze care au mai mult de 3 cuvinte si numaram cate fraze apar cu exact aceleasi parti de vorbire astfel:
  1. sortam partile de vorbire si le concatenam intr-un string (doua string-uri nu pot fi identice decat daca sunt propozitii care au exact aceleasi parti de vorbire)
  2. memoram intr-un dictionar numarul de aparitii al frazelor
3. La sfarsit, sortam listele corespunzatoare celor doua clase si vedem ce dependinte exista intra partile de vorbire din frazele care apar cel mai des si sunt compuse din aceste parti de vorbire
4. De asemenea am facut un tip de dependency trees in care conteaza topica partilor de vorbire, adica am eliminat sortarea

# Concluzii (sortat)

- ❖ Există diferențe majore între cele două limbi

## Moldova

ADP ADP ADV AUX NOUN NOUN PRON PUNCT 121  
ADP ADP ADV AUX NOUN NOUN NOUN NOUN PRON PUNCT PUNCT 115  
ADJ ADP ADP ADP ADP ADV ADV AUX AUX DET DET NOUN NOUN NOUN NUM PRON PROPN PUNCT PUNCT VERB 114

## Romana

ADJ ADP ADP ADP ADP AUX DET NOUN NOUN NOUN PRON PRON PUNCT PUNCT PUNCT SPACE VERB VERB VERB 517  
ADJ ADP ADP ADV ADV AUX DET NOUN PROPN PUNCT SPACE 166  
ADP ADV ADV NOUN NOUN PRON PUNCT PUNCT SPACE VERB 21

Ro-Md common: 714  
Ro common percentage: 3.58%  
Md common percentage: 1.73%

## RO ---> MD

ADJ ADP ADP ADP ADP AUX DET NOUN NOUN NOUN PRON PRON PUNCT PUNCT PUNCT SPACE VERB VERB VERB - RO:517 - MD:0  
ADJ ADP ADP ADV ADV AUX DET NOUN PROPN PUNCT SPACE - RO:166 - MD:0  
ADP ADV ADV NOUN NOUN PRON PUNCT PUNCT SPACE VERB - RO:21 - MD:0

## MD ---> RO

ADP ADP ADV AUX NOUN NOUN PRON PUNCT - RO:0 - MD:121  
ADP ADP AUX NOUN NOUN NOUN NOUN PRON PUNCT PUNCT - RO:0 - MD:115  
ADJ ADP ADP ADP ADP ADP ADV ADV AUX AUX DET DET NOUN NOUN NOUN NUM PRON PROPN PUNCT PUNCT VERB - RO:0 - MD:114

# Concluzii (nesortat, cu topica)

- ❖ Atunci cand luam in considerare si topica partilor de vorbire putem observa diferente si mai mari intre cele doua limbi

## Moldova

NOUN PRON AUX ADP NOUN PUNCT NOUN ADP NOUN PUNCT 114

PRON AUX ADV DET ADV ADJ NOUN AUX VERB ADP PROPN PUNCT ADP DET NOUN ADP NUM ADP NOUN ADP NOUN PUNCT 114

PRON ADV AUX ADP NOUN ADP NOUN PUNCT 113

## Romana

NOUN ADP NOUN ADP PRON PRON AUX VERB ADP NOUN ADJ PUNCT VERB ADP PUNCT SPACE DET VERB PUNCT 517

ADV AUX DET NOUN ADV ADP ADJ ADP PROPN PUNCT SPACE 166

NOUN PRON VERB ADV ADV PUNCT ADP NOUN PUNCT SPACE 21

Ro-Md common: 221

Ro common percentage: 1.09%

Md common percentage: 0.52%

RO ---> MD

NOUN ADP NOUN ADP PRON PRON AUX VERB ADP NOUN ADJ PUNCT VERB ADP PUNCT SPACE DET VERB PUNCT - RO:517 - MD:0

ADV AUX DET NOUN ADV ADP ADJ ADP PROPN PUNCT SPACE - RO:166 - MD:0

NOUN PRON VERB ADV ADV PUNCT ADP NOUN PUNCT SPACE - RO:21 - MD:0

MD ---> RO

NOUN PRON AUX ADP NOUN PUNCT NOUN ADP NOUN PUNCT - RO:0 - MD:114

PRON AUX ADV DET ADV ADJ NOUN AUX VERB ADP PROPN PUNCT ADP DET NOUN ADP NUM ADP NOUN ADP NOUN PUNCT - RO:0 - MD:114

PRON ADV AUX ADP NOUN ADP NOUN PUNCT - RO:0 - MD:113

### 3. Generarea de text

1. Incarcam toate textele din cele doua limbi si le concatenam in doua stringuri
2. Cele doua stringuri sunt trim-uite la 1 mil de caractere pentru a le putea pasa catre spaCy
3. Impartim textele in grupuri de 5 cuvinte si intr-un dictionar care are ca si cheie 4-gram-ul format din primele 4 cuvinte contorizam aparitile celui de-al 5-lea cuvant
4. Generam astfel:
  1. Alegem un 4-gram random si completam al 5-lea cuvant cu cel mai frecvent pentru acel 4-gram
  2. Alegem ultimele 4 cuvinte din fraza generate si repetam procesul pana se ajunge la un semn de punctuatie dintre . ? ! sau pana cand nu se mai gaseste un cuvant potrivit
  3. Conditia de oprire este semn de punctuatie de oprire sau lipsa unui cuvant de ales

Text generat automat Romania: PENTRU MAI MULTE REGIUNI DIN ROMÂNIA ) Un ciclon polar lovește România Așa cum spuneam , testul de astăzi are o dificultate ridicată și necesită ceva mai multă atenție și concentrare .

Text generat automat Moldova: Aici , Danemarca si Germania vor avea din start cate patru puncte , Romania si Polonia vor avea cate doua , iar Japonia si Serbia - nici cate unul .

# Concluzii generale

1. Conform datelor prezентate anterior putem concluziona ca exista diferențe vizibile între limba romana scrisa in Romania si cea scrisa in Moldova, atat la nivel de vocabular cat si la nivel de legaturi de parti de vorbire si topica
2. Ce putem imbunatatiti:
  1. Colectarea unui numar mai mare de articole din cele doua tari
  2. Implementarea unui model state of the art care sa analizeze cele doua limbi
  3. Folosirea altor modele de generare de text automat precum GPT-Neo

