Topic Modeling for Romanian News Articles

A TF-IDF and LDA Hybrid Approach

Stanea Adrian-Bogdan

Technical University of Cluj-Napoca

NLP Seminar - 2026

Outline

Outline

Project Overview & Objectives

What is Topic Modeling?

- Automatically discover hidden themes in text collections
- Each topic = a group of related words
- Documents can belong to multiple topics

The Challenge:

- Romanian news articles need automatic thematic organization
- Language-specific issues: diacritics, dialects

Our Objectives

1. Romanian-specific preprocessing
2. Hybrid TF-IDF + LDA pipeline
3. Evaluate against known categories
4. Analyze model confidence

MOROCO Dataset at a Glance

Dataset Overview:

- **21,719** news articles
- **6 categories**: politics, finance, sports, tech, science, culture
- Romanian + Moldavian dialects

Key Observation

Class imbalance: Politics & finance = 52% of corpus

Experiment: 2,000 stratified sample

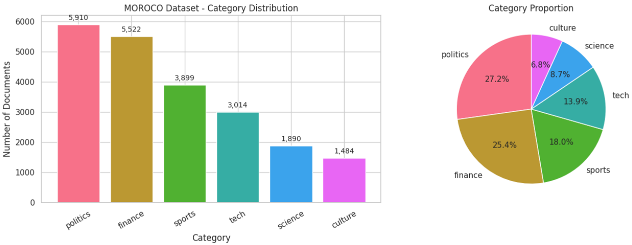


Figure: Category distribution

Outline

Pipeline Architecture: Filter-then-Feed

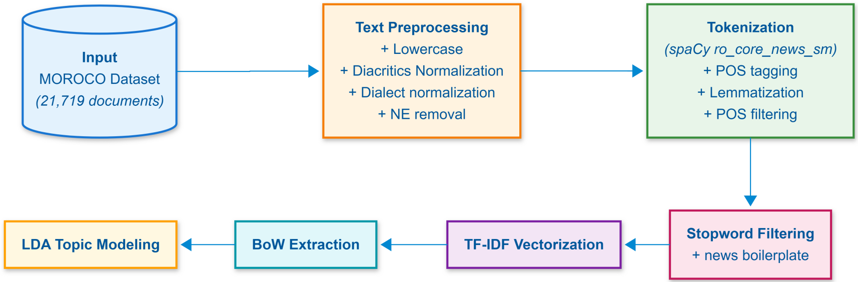


Figure: Complete topic modeling pipeline

Key Insight

- TF-IDF selects informative features
- BoW counts feed LDA (preserves probabilistic integrity)
- LDA uncovers latent topics

Romanian-Specific Text Normalization

Three essential transformations:

1. Diacritics

Legacy Unicode fix

`ş → ș`
`ţ → ț`

cedilla → comma-below

2. Dialect Harmony

Moldovan → Romanian

`sînt → sânt`
`vînt → vânt`

mid-word î → â

3. Placeholders

Remove NE tokens

Anonymized named entities in MOROCO

prevents noise

Result: Unified vocabulary across Romanian & Moldavian text

Tokenization & Feature Selection

POS Filtering Strategy

- **Keep:** Nouns, Proper Nouns, Adjectives
- **Remove:** Verbs, function words

Why?

Nouns/adjectives = **topical content**
Verbs = writing style, not topic

Tool: spaCy `ro_core_news_sm`

TF-IDF Configuration

Parameter	Value
max_df	0.4
min_df	5
n-grams	(1, 2)
sublinear_tf	True
Final vocab	**4,678**

Stopwords: 539 total
(including 81 news-specific terms)

LDA Model Configuration

Latent Dirichlet Allocation

- Documents = mixtures of topics
- Topics = probability over words
- Unsupervised learning

Configuration

Topics	6 (= categories)
Method	Online Variational Bayes
Iterations	750
Random seed	42

Model Output:

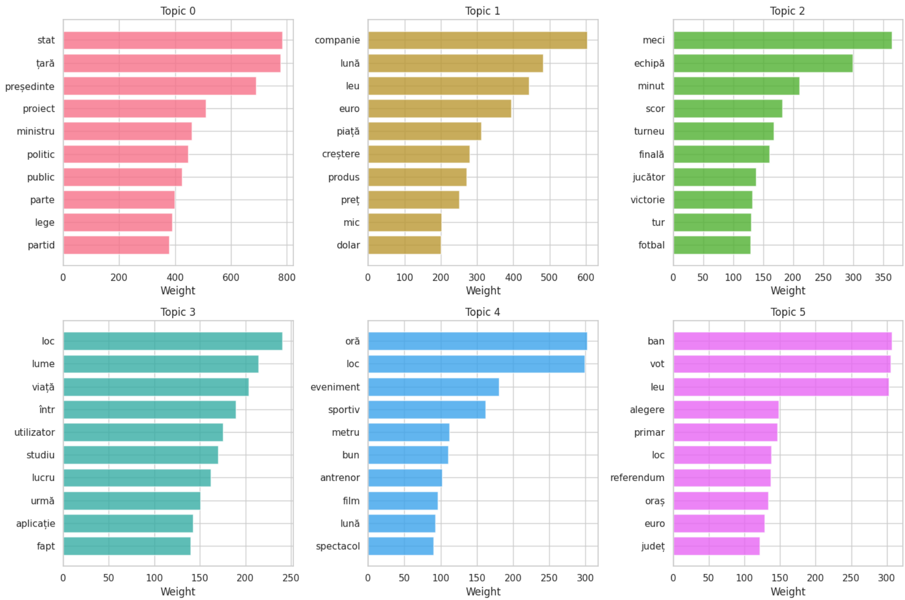
θ (doc-topic)

Each document gets a probability distribution over 6 topics

ϕ (topic-word)

Each topic gets a probability distribution over 4,678 words

Discovered Topics: Top Words



Outline

Document Distribution Across Topics

Observations:

- Topic 0 dominates (39.8%)
- Topic 5 smallest (4.9%)
- Reflects source data imbalance

Topic Labels

0	Politics
1	Economy
2	Sports
3	Science/Life
4	Events
5	Elections

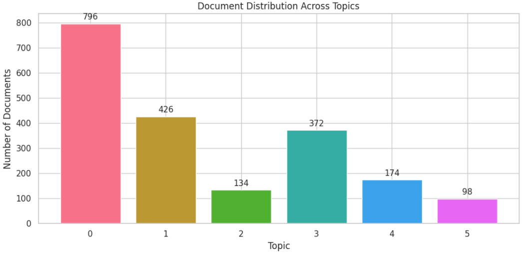


Figure: Documents assigned per topic

Topic-Category Correspondence

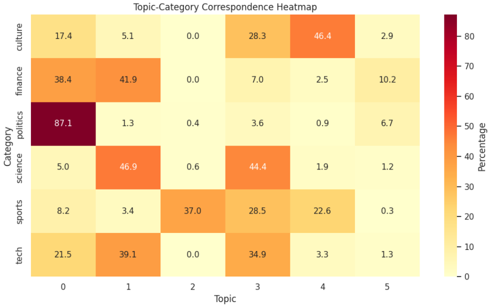


Figure: Percentage of each category assigned to each topic

✓ Strong

Politics → T0

87.1%

✓ Moderate

Finance → T1: 42%

Sports → T2: 37%

× Confusion

Science/Tech overlap in T1 & T3

Model Confidence Analysis

- Most documents have **moderate-to-high** topic affinity
- Low confidence ($\sim$0.17) = multi-topic or ambiguous content.

Confidence Statistics:

Mean	0.65
Median	0.63
Min	0.17
Max	0.99

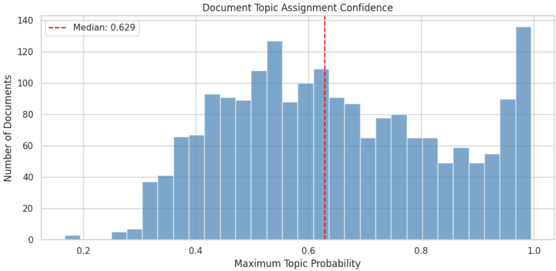


Figure: Distribution of max topic probability

Example Topic Assignments

How the model classifies real documents:

Example 1: Politics (Confidence: 0.89)

"Ministrul a declarat că proiectul de lege va fi votat în parlament..."
→ **Topic 0** (stat, țară, președinte, ministru)

Example 2: Sports (Confidence: 0.72)

"Echipa a câștigat meciul cu scorul de 3-1 în minutul 90..."
→ **Topic 2** (meci, echipă, scor, jucător)

Example 3: Finance (Confidence: 0.68)

"Compania a raportat o creștere de 15% pe piața europeană..."
→ **Topic 1** (companie, euro, piață, creștere)

Outline

Key Takeaways & Future Directions

✓ **What We Achieved:**

- Romanian preprocessing pipeline (diacritics, dialect harmony)
- 6 interpretable topics extracted
- Strong alignment for politics (**87%**)
- Moderate for sports (37%) and finance (42%)
- Mean confidence: **0.65**

→ **Future Improvements:**

- Balanced sampling to reduce category bias
- More topics for better science/tech separation
- Hyperparameter tuning (α, β, K)
- Neural approaches (BERTopic) for comparison
- Full dataset training

Bottom Line

Classical LDA with proper preprocessing remains effective for interpretable topic discovery in Romanian text

Thank You!

Questions?

Stanea Adrian-Bogdan

stanea.adrian@student.utcluj.ro