Topic Modeling for Romanian News Articles: A TF-IDF and LDA Hybrid Approach

Adrian Bogdan STÂNEA
Communications Department
Technical University of Cluj-Napoca, Romania
adrianstanea1@gmail.com

***Abstract*—This paper presents a topic modeling system for Romanian news articles using a hybrid TF-IDF and Latent Dirichlet Allocation (LDA) approach. We employ the MOROCO (Moldavian and Romanian Dialectal Corpus) dataset, comprising 21,719 news articles across six categories: politics, finance, sports, technology, science, and culture. Our methodology incorporates Romanian-specific preprocessing, including diacritic normalization, dialect harmonization, and part-of-speech filtering to extract the semantic skeleton of documents. The TF-IDF vectorization stage selects informative features while filtering noise, followed by LDA topic extraction using bag-of-words counts to preserve probabilistic integrity. Experimental results on a stratified sample of 2,000 documents demonstrate strong topic-category correspondence, particularly for politics (87.1% alignment) and sports (37% alignment), with a mean document assignment confidence of 0.65. We identify limitations including class imbalance effects and propose future improvements such as balanced sampling strategies and alternative embedding-based approaches.**

***Index Terms*—Topic Modeling, LDA, TF-IDF, Romanian NLP, Text Preprocessing, MOROCO**

I. Introduction

Topic modeling represents a fundamental task in natural language processing, enabling the automatic discovery of latent semantic structures within document collections. Since the introduction of Latent Dirichlet Allocation (LDA) by Blei et al. [1], probabilistic topic models have become essential tools for analyzing large text corpora across diverse domains including news analysis, scientific literature, and social media.

A. Related Work

Topic Modeling Foundations. The seminal work on LDA [1] established the generative probabilistic framework that underlies most modern topic models. Subsequent developments have addressed scalability through online learning [8], which enables processing of large-scale corpora by updating topic parameters incrementally. Extensions such as Dynamic Topic Models [2] capture temporal evolution of topics, while Correlated Topic Models [3] model inter-topic correlations through the logistic normal distribution.

Neural Topic Models. Recent advances have integrated deep learning with topic modeling. Neural Variational Document Model (NVDM) [11] uses variational autoencoders for document representation, while ProdLDA [14] replaces the mixture model with a product of experts for improved topic coherence. BERTopic [7] leverages transformer-based embeddings with clustering algorithms, demonstrating superior performance on short texts. However, these neural approaches require substantial computational resources and large training corpora, motivating continued interest in classical LDA for resource-constrained settings.

Topic Modeling for News. News article analysis represents a prominent application domain for topic models. Newman et al. [12] demonstrated LDA's effectiveness for discovering themes in news corpora, while subsequent work has explored supervised variants that incorporate category labels [10]. The combination of TF-IDF feature selection with LDA has been shown to improve topic coherence by filtering noise from high-frequency terms [15].

Romanian NLP and MOROCO. Romanian presents specific challenges for NLP due to its rich morphology, diacritic variations, and dialectal differences between Romanian and Moldavian variants. The MOROCO dataset [4] provides a valuable benchmark for Romanian text classification, containing news articles with dialect labels that enable cross-lingual analysis. Prior work on MOROCO has focused primarily on dialect identification and text classification using transformer models [6], with RoBERT achieving state-of-the-art results on Romanian language understanding tasks. However, topic modeling approaches for Romanian remain underexplored, particularly regarding preprocessing strategies that address orthographic normalization and morphological variation.

Preprocessing for Morphologically Rich Languages. Languages with complex morphology benefit from lemmatization and part-of-speech filtering to reduce vocabulary sparsity [5]. For topic modeling specifically, retaining only content-bearing parts of speech (nouns, adjectives) has been shown to improve topic interpretability by focusing on the "topical skeleton" while filtering stylistic elements [9].

B. Problem Statement and Contributions

The proliferation of digital news content in Romanian necessitates automated methods for content organization and thematic analysis. While neural approaches dominate current benchmarks, classical LDA remains attractive for its interpretability, computational efficiency, and theoretical foundations. This paper addresses the specific challenge of extracting coherent topics from Romanian news articles, where the inherent category structure provides a natural evaluation framework.

We define key terminology used throughout this work. A *document* refers to a text sample from the MOROCO dataset

representing a news article. A *topic* is a probability distribution over vocabulary words representing a semantic theme—for instance, a politics topic would assign high probability to words such as "president," "government," and "law." A *corpus* denotes the complete collection of documents under analysis.

This work pursues four primary objectives:

1) Implement Romanian-specific text preprocessing, including diacritic normalization and cross-dialect harmonization between Romanian and Moldavian variants
2) Apply a hybrid TF-IDF and LDA pipeline for topic extraction, leveraging TF-IDF for feature selection while preserving LDA's probabilistic foundations
3) Evaluate topic coherence against known news categories, using the ground-truth labels as an extrinsic validation mechanism
4) Analyze model confidence and topic interpretability through document assignment distributions

The remainder of this paper is organized as follows. Section II provides theoretical background on TF-IDF weighting and LDA topic models. Section III details our dataset, preprocessing pipeline, and modeling approach. Section IV presents experimental results with visualizations. Section V summarizes findings and proposes future directions.

II. Theoretical Overview

This section establishes the theoretical foundations for our topic modeling approach, covering TF-IDF weighting, Latent Dirichlet Allocation, and considerations specific to Romanian natural language processing.

A. TF-IDF Weighting

Term Frequency-Inverse Document Frequency (TF-IDF) provides a statistical measure of term importance within a document relative to a corpus [13]. The weighting scheme combines two components: term frequency (TF), measuring local importance within a document, and inverse document frequency (IDF), measuring global importance across the corpus.

The TF-IDF weight for term t in document d is computed as:

$$\text{tfidf}(t, d) = \text{tf}(t, d) \times \text{idf}(t) \tag{1}$$

where the term frequency with sublinear scaling is:

$$\text{tf}(t, d) = 1 + \log(\text{count}(t, d)) \tag{2}$$

and the inverse document frequency is:

$$\text{idf}(t) = \log\left(\frac{N}{\text{df}(t)}\right) \tag{3}$$

Here, $\text{count}(t, d)$ denotes the raw frequency of term t in document d, N represents the total number of documents in the corpus, and $\text{df}(t)$ is the document frequency—the number of documents containing term t.

The sublinear term frequency scaling (Equation 2) mitigates the impact of term repetition, preventing very frequent terms within a single document from dominating the representation. The IDF component down-weights terms appearing in many documents, as such terms provide less discriminative power for distinguishing document content.

In our pipeline, TF-IDF serves a feature selection role rather than providing the final document representation. We use TF-IDF scores to identify the most informative vocabulary terms, which are then represented as bag-of-words counts for LDA training.

B. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for discovering latent topics in document collections [1]. LDA assumes that documents are mixtures of topics, where each topic is characterized by a probability distribution over words.

The generative process for LDA proceeds as follows:

1) For each topic $k \in \{1, \ldots, K\}$:
   - Draw topic-word distribution $\phi_k \sim \text{Dirichlet}(\beta)$
2) For each document d:
   - Draw document-topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$
   - For each word position n in document d:
     - Draw topic assignment $z_n \sim \text{Multinomial}(\theta_d)$
     - Draw word $w_n \sim \text{Multinomial}(\phi_{z_n})$

The hyperparameters α and β control the sparsity of document-topic and topic-word distributions, respectively. Lower values encourage sparse distributions where documents focus on fewer topics and topics concentrate on fewer words.

Exact inference in LDA is intractable due to the coupling between latent variables. We employ online variational Bayes inference [8], which processes documents in mini-batches and updates global topic parameters incrementally. This approach scales efficiently to large corpora while maintaining approximation quality.

C. Filter-then-Feed Architecture

Our pipeline employs a "filter-then-feed" architecture that combines TF-IDF feature selection with LDA topic modeling. This design addresses a fundamental tension: TF-IDF excels at identifying discriminative features but produces weighted representations incompatible with LDA's generative assumptions, while pure bag-of-words input to LDA includes substantial noise from high-frequency, low-information terms.

The architecture proceeds in two stages:

1) **Feature Selection**: Apply TF-IDF vectorization with frequency thresholds to identify the top-K most informative vocabulary terms
2) **Topic Modeling**: Extract raw word counts for the selected vocabulary and train LDA on these bag-of-words representations

This approach preserves LDA's probabilistic integrity—the model receives actual word counts from which it can infer topic proportions—while benefiting from TF-IDF's noise reduction capabilities. Terms that are either too common

(appearing in more than 40% of documents) or too rare (appearing in fewer than 5 documents) are pruned during the feature selection stage.

D. Romanian NLP Considerations

Romanian presents specific challenges for text processing that our pipeline addresses:

Diacritic Variation: Romanian uses diacritical marks (ă, â, î, ș, ț) that historically had multiple Unicode representations. Legacy text often uses cedilla characters (ş, ţ) instead of the standard comma-below forms (ș, ț). Our normalization maps all variants to the standard Unicode representation.

Dialect Harmonization: The MOROCO corpus includes both Romanian and Moldavian text, which exhibit orthographic differences. Most notably, Moldavian text uses "î" in word-medial positions where Romanian uses "â" (e.g., "sînt" vs. "sânt" for "am"). Our preprocessing applies mid-word î→â conversion to unify the vocabulary across dialects.

Semantic Compression: Romanian's rich morphology produces many inflected forms that obscure semantic relationships. We apply lemmatization and part-of-speech filtering to extract the "topical skeleton" of documents—nouns, proper nouns, and adjectives that carry semantic content—while excluding verbs (which reflect writing style rather than topic) and function words.

III. Methodology

This section describes our experimental methodology, including the dataset, preprocessing pipeline, and topic modeling configuration.

A. Dataset Description

We employ the MOROCO (Moldavian and Romanian Dialectal Corpus) dataset [4], a collection of Romanian news articles available through the Hugging Face datasets repository. The corpus comprises 21,719 documents in the training split, distributed across six categories as shown in Table I.

TABLE I
MOROCO Dataset Category Distribution

Category	Documents	Percentage
Politics	5,910	27.2%
Finance	5,522	25.4%
Sports	3,899	18.0%
Tech	3,014	13.9%
Science	1,890	8.7%
Culture	1,484	6.8%
Total	**21,719**	**100.0%**

The category distribution exhibits notable class imbalance, with politics and finance together comprising over half the corpus while culture represents less than 7%. This imbalance influences topic modeling outcomes, as discussed in Section IV.

Document length varies substantially across the corpus, with a mean of 309 words (median: 238 words) and maximum of 15,988 words. The long tail of document lengths reflects the presence of extended news articles and compilations. Figure 1 illustrates the category distribution.

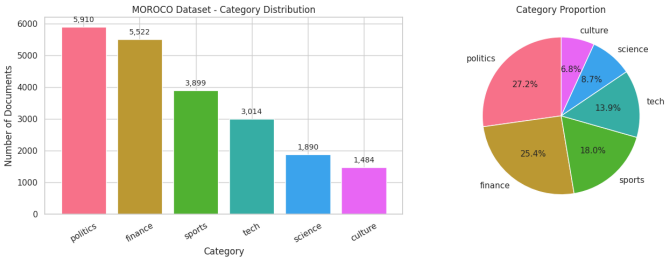

Fig. 1. Distribution of documents across the six MOROCO categories. Politics and finance dominate the corpus, while culture and science are underrepresented.

B. Text Preprocessing

Our preprocessing pipeline applies a sequence of normalization steps tailored to Romanian text characteristics. Each step addresses specific data quality issues present in the MOROCO corpus.

Lowercasing: All text is converted to lowercase to ensure consistent vocabulary matching regardless of sentence position or stylistic capitalization.

Diacritic Normalization: Legacy cedilla characters are mapped to standard Unicode comma-below forms. For example:

- "Şcoala ţării" → "Școala țării"
- "România şi Moldova" → "România și Moldova"

Dialect Normalization: Mid-word "î" characters are converted to "â" to harmonize Romanian and Moldavian orthography:

- "sînt" → "sânt" (I am)
- "pămînt" → "pământ" (earth)
- "vînt" → "vânt" (wind)

This transformation preserves word-initial and word-final "î", which is standard in both variants.

Named Entity Placeholder Removal: The MOROCO corpus uses "NE" placeholders to mark anonymized named entities. These placeholders are removed to prevent them from appearing as vocabulary terms.

Whitespace Normalization: Multiple consecutive whitespace characters are collapsed to single spaces, and leading/trailing whitespace is removed.

C. Tokenization and POS Filtering

Tokenization employs the spaCy `ro_core_news_sm` model, a Romanian language model providing tokenization, lemmatization, and part-of-speech tagging capabilities.

The tokenization process extracts lemmatized tokens—base dictionary forms that abstract over morphological variation. For example, "economiei" (of the economy), "economia" (the economy), and "economic" (economic) all reduce to the lemma "economie" or related forms.

We apply part-of-speech filtering to retain only tokens tagged as NOUN, PROPN (proper noun), or ADJ (adjective). This filtering strategy implements semantic compression: nouns and adjectives carry topical content ("economy," "political," "championship"), while verbs often reflect journalistic writing conventions ("declared," "announced," "confirmed") rather than document topics.

Table II illustrates the filtering process for a sample sentence.

TABLE II
POS FILTERING EXAMPLE

Token	POS	Retained
ministrul (minister)	NOUN	✓
economiei (economy)	NOUN	✓
a (has)	AUX	–
declarat (declared)	VERB	–
că (that)	SCONJ	–
piața (market)	NOUN	✓
financiară (financial)	ADJ	✓
crește (grows)	VERB	–

D. Stopword Handling

Stopword filtering combines multiple sources to ensure comprehensive coverage:

- **Standard Lists**: NLTK Romanian stopwords, stop-words package, and stopwordsiso provide baseline coverage of function words and common terms
- **News Boilerplate**: 81 additional terms specific to journalistic text, including verbal noise ("declara" [to declare], "anunța" [to announce]) and web artifacts ("foto" [photo], "video")

The combined stopword list comprises 539 unique terms. Additionally, tokens shorter than 3 characters are excluded to remove residual noise.

E. Topic Modeling Pipeline

The topic modeling pipeline implements the filter-then-feed architecture described in Section II-C. Figure 2 illustrates the complete processing flow.

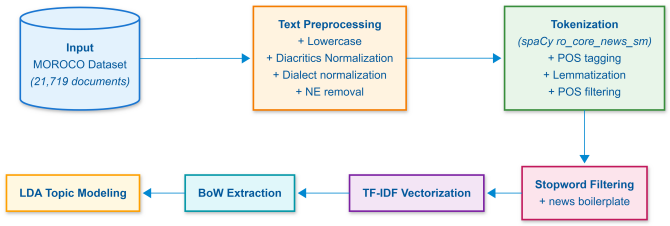

Fig. 2. Topic modeling pipeline architecture. TF-IDF identifies informative vocabulary, then bag-of-words counts for those terms feed the LDA model.

TF-IDF Configuration: The vectorization stage applies the following parameters:

- `max_df=0.4`: Prune terms appearing in more than 40% of documents
- `min_df=5`: Prune terms appearing in fewer than 5 documents
- `max_features=5000`: Limit vocabulary to top 5,000 features by TF-IDF score
- `ngram_range=(1, 2)`: Include unigrams and bigrams
- `sublinear_tf=True`: Apply logarithmic term frequency scaling

These parameters produced a vocabulary of 4,678 features after filtering, providing sufficient expressiveness while excluding corpus-wide common terms and rare terms that may represent noise or OCR errors.

LDA Configuration: The topic model employs the following settings:

- `n_topics=6`: Match the number of known categories
- `learning_method='online'`: Online variational Bayes for efficiency
- `max_iter=750`: Maximum training iterations
- `random_state=42`: Fixed seed for reproducibility

The choice of 6 topics aligns with the known category structure, enabling direct evaluation of topic-category correspondence.

IV. EVALUATION AND RESULTS

This section presents experimental results, analyzing topic quality through multiple lenses: document distribution, topic-category correspondence, topic word interpretability, and assignment confidence.

A. Experimental Setup

Due to computational constraints in the demonstration context, we train on a stratified random sample of 2,000 documents that preserves category proportions. The sample distribution comprises: politics (552), finance (489), sports (354), tech (307), science (160), and culture (138). All experiments use Python 3.12 with scikit-learn for LDA and spaCy for preprocessing.

B. Document Distribution Across Topics

Table III presents the distribution of documents across discovered topics, with each document assigned to its highest-probability topic.

TABLE III
DOCUMENT DISTRIBUTION ACROSS TOPICS

Topic	Documents	Percentage
Topic 0	796	39.8%
Topic 1	426	21.3%
Topic 2	134	6.7%
Topic 3	372	18.6%
Topic 4	174	8.7%
Topic 5	98	4.9%

Topic 0 dominates with nearly 40% of documents, while Topic 5 contains only 4.9%. This imbalance partially reflects the source category distribution but also indicates that certain topics capture broader semantic content than others. Figure 3 visualizes this distribution.

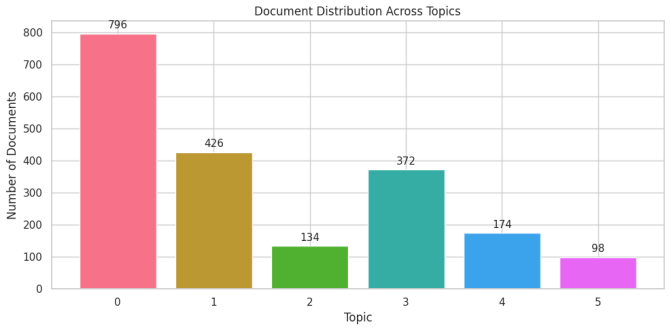

Fig. 3. Distribution of documents across the six discovered topics. Topic 0 captures the largest share, corresponding primarily to political content.

C. Topic-Category Correspondence

A key evaluation criterion is the alignment between discovered topics and known document categories. Table IV presents a confusion matrix showing the percentage of each category's documents assigned to each topic.

TABLE IV
TOPIC-CATEGORY CORRESPONDENCE (%)

Category	T0	T1	T2	T3	T4	T5
Culture	17.4	5.1	0.0	28.3	46.4	2.9
Finance	38.4	41.9	0.0	7.0	2.5	10.2
Politics	87.1	1.3	0.4	3.6	0.9	6.7
Science	5.0	46.9	0.6	44.4	1.9	1.2
Sports	8.2	3.4	37.0	28.5	22.6	0.3
Tech	21.5	39.1	0.0	34.9	3.3	1.3

Several patterns emerge from this analysis:

Strong Alignment: Politics exhibits the clearest topic correspondence, with 87.1% of political documents assigned to Topic 0. This reflects the distinctive vocabulary of political discourse (government, parliament, law, minister).

Moderate Alignment: Sports shows 37% concentration in Topic 2, with the remainder split across Topics 3 and 4. Finance concentrates in Topics 0 (38.4%) and 1 (41.9%), suggesting that some financial content shares vocabulary with political news while other financial content forms a distinct cluster.

Topic Confusion: Science and technology exhibit notable confusion, both splitting primarily between Topics 1 and 3. This overlap suggests shared vocabulary around research, innovation, and technical terminology that the model struggles to differentiate with only 6 topics.

Cultural Content: Culture distributes across Topics 3 and 4, likely reflecting the diversity of cultural coverage (events, arts, entertainment) that overlaps with sports events and general interest content.

Figure 4 provides a visual representation of these correspondences.

D. Topic Word Interpretation

Table V presents the top words for each topic, enabling qualitative interpretation of topic semantics.

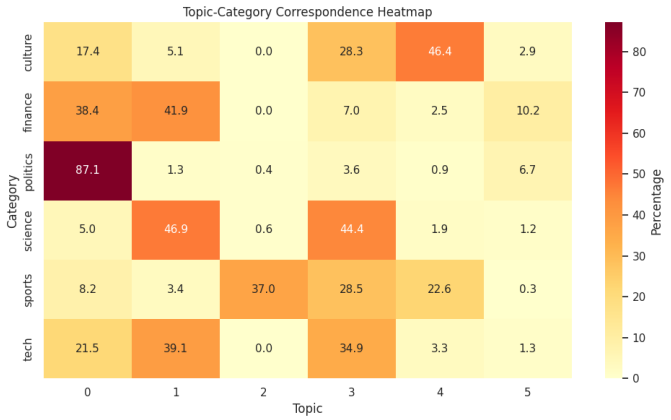

Fig. 4. Heatmap showing the percentage of documents from each category assigned to each topic. Darker colors indicate stronger correspondence.

TABLE V
TOP WORDS PER TOPIC

Topic	Top Words (Romanian/English)
0	stat (state), țară (country), președinte (president), proiect (project), ministru (minister), politic (political)
1	companie (company), lună (month), leu (leu), euro (euro), piață (market), creștere (growth)
2	meci (match), echipă (team), minut (minute), scor (score), turneu (tournament), finală (final)
3	loc (place), lume (world), viață (life), utilizator (user), studiu (study), lucru (thing)
4	oră (hour), loc (place), eveniment (event), sportiv (sports), metru (meter), film (film)
5	ban (money), vot (vote), leu (leu), alegere (election), primar (mayor), referendum

Based on these word distributions, we assign interpretive labels:

- **Topic 0**: Politics/Government
- **Topic 1**: Economy/Business
- **Topic 2**: Sports/Competitions
- **Topic 3**: Science/Life
- **Topic 4**: Events/Culture
- **Topic 5**: Elections/Local Politics

Notably, Topic 5 captures a sub-domain of political content focused on elections and local government, explaining why some political documents do not map to Topic 0. Figure 5 visualizes the word weights.

E. Document Assignment Confidence

We analyze the confidence of topic assignments by examining the maximum topic probability for each document. Table VI summarizes these statistics.

The mean confidence of 0.65 indicates that most documents have a moderately strong affiliation with their dominant topic. The distribution, shown in Figure 6, reveals that while many documents have confidence above 0.6, a substantial fraction

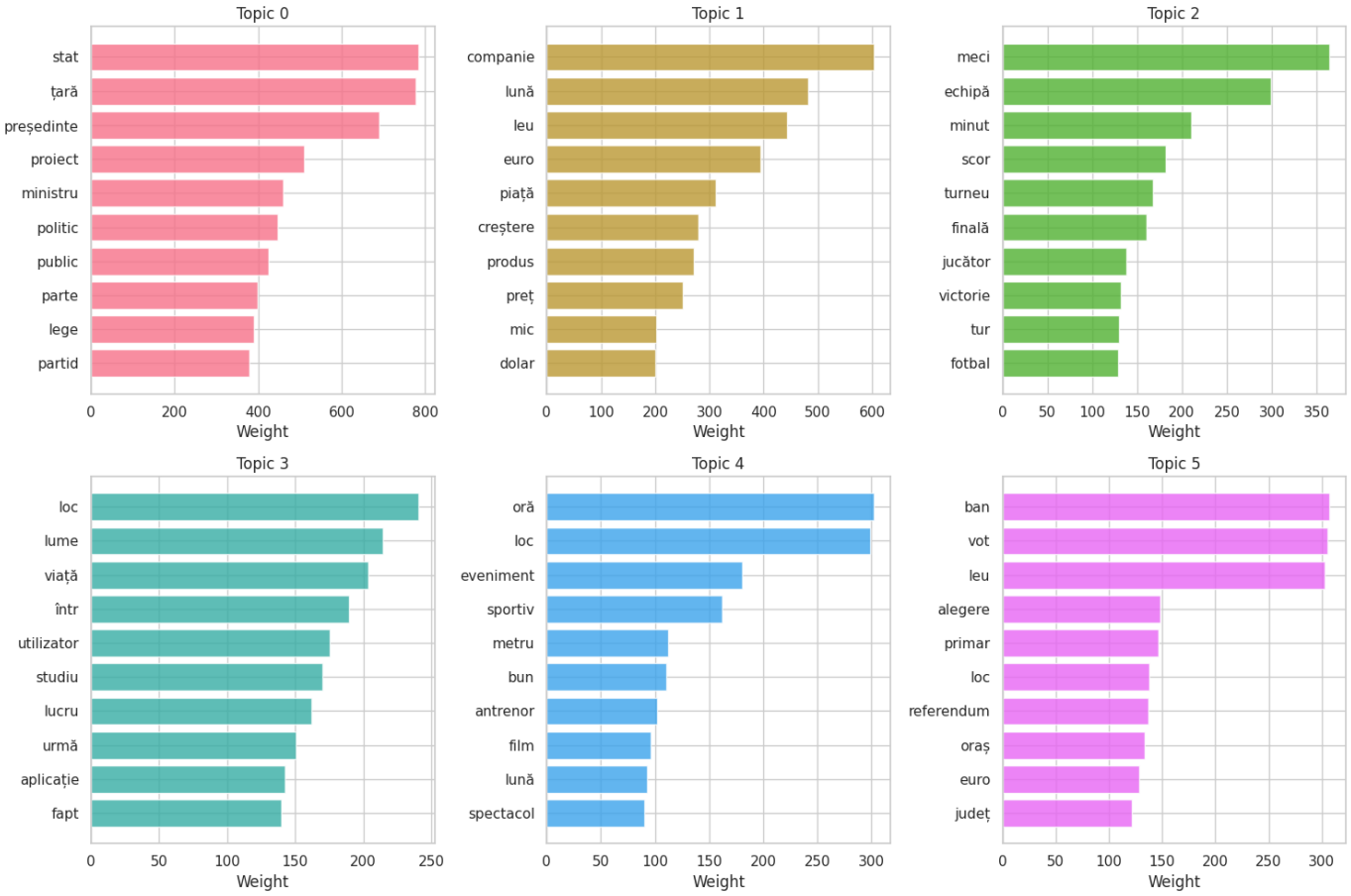

Fig. 5. Horizontal bar charts showing the top 10 words per topic with their weights. Higher weights indicate stronger association with the topic.

TABLE VI
DOCUMENT ASSIGNMENT CONFIDENCE STATISTICS

Statistic	Value
Mean	0.65
Median	0.63
Minimum	0.17
Maximum	0.99

exhibits lower confidence, suggesting multi-topic content or documents with ambiguous thematic content.

Documents with minimum confidence near 0.17 (approximately $1/6$, the uniform probability) represent cases where the model finds no clear topic signal, possibly due to short document length or unusual content.

V. CONCLUSIONS

This paper presented a topic modeling system for Romanian news articles, combining TF-IDF feature selection with LDA topic extraction. Our approach addresses Romanian-specific preprocessing challenges including diacritic normalization and dialect harmonization, while employing part-of-speech filtering to extract semantically relevant vocabulary.

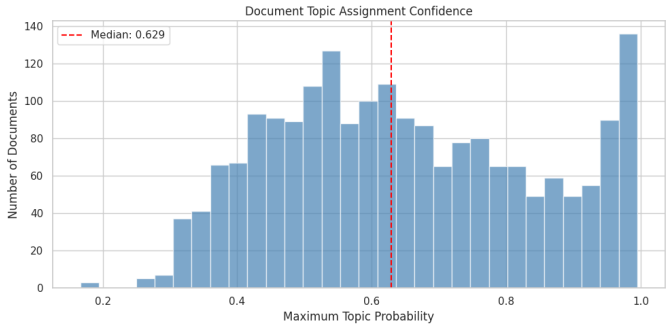

Fig. 6. Distribution of maximum topic probability (confidence) across documents. The median confidence of 0.63 indicates moderate topic specificity.

A. Summary of Findings

The experimental results demonstrate several achievements:

- The Romanian preprocessing pipeline successfully normalizes text across diacritic and dialect variations, producing a unified vocabulary of 4,678 features
- The TF-IDF + LDA pipeline extracts 6 interpretable topics that correspond to recognizable semantic themes (politics, economy, sports, science/life, events, elections)
- Strong topic-category alignment exists for politics

(87.1%) and moderate alignment for sports (37%) and finance (42%)
- Document assignment confidence averages 0.65, indicating reasonable topic specificity

B. Limitations

Several limitations affect our results:

- **Class Imbalance**: The dominance of politics and finance in the source corpus influences topic discovery, with Topic 0 capturing a disproportionate share of documents
- **Category Confusion**: Science and technology exhibit substantial overlap, suggesting that 6 topics may be insufficient to capture the full semantic diversity of these domains
- **Sample Size**: Our analysis uses a 2,000-document sample rather than the full corpus, potentially affecting topic stability
- **Evaluation Scope**: We rely on category correspondence as a proxy for topic quality; intrinsic coherence measures would provide complementary evaluation

C. Future Work

Several directions could improve upon this work:

1) **Balanced Sampling**: Training on category-balanced samples would reduce the influence of dominant categories on topic discovery
2) **Outlier Filtering**: Removing extremely long documents (>1,000 words) may improve topic coherence by excluding compilations and aggregated content
3) **Hyperparameter Optimization**: Systematic exploration of topic counts (K), prior values (α, β), and vocabulary size could identify improved configurations
4) **Embedding-Based Models**: Modern approaches such as BERTopic leverage contextual embeddings and may better capture semantic nuances in Romanian text
5) **Cross-Lingual Evaluation**: Comparing topic quality with multilingual models (e.g., XLM-RoBERTa) would assess the benefit of Romanian-specific preprocessing

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006, pp. 113–120.

[3] D. M. Blei and J. D. Lafferty, "A correlated topic model of Science," *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.

[4] A. M. Butnaru and R. T. Ionescu, "MOROCO: The Moldavian and Romanian dialectal corpus," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 688–698.

[5] F. Can and E. A. Ozkarahan, "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases," *ACM Transactions on Database Systems*, vol. 15, no. 4, pp. 483–517, 1990.

[6] S. D. Dumitrescu, A. M. Avram, and S. Pyysalo, "The birth of Romanian BERT," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4324–4328.

[7] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[8] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent Dirichlet allocation," in *Advances in Neural Information Processing Systems*, 2010, pp. 856–864.

[9] F. Martin and M. Johnson, "More efficient topic modelling through a noun only approach," in *Proceedings of the Australasian Language Technology Association Workshop*, 2015, pp. 111–115.

[10] J. D. McAuliffe and D. M. Blei, "Supervised topic models," in *Advances in Neural Information Processing Systems*, 2007, pp. 121–128.

[11] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016, pp. 1727–1736.

[12] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers, "Analyzing entities and topics in news articles using statistical topic models," in *IEEE International Conference on Intelligence and Security Informatics*, 2006, pp. 93–104.

[13] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[14] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," in *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[15] A. T. Wilson and P. A. Chew, "Term weighting schemes for latent Dirichlet allocation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 2010, pp. 465–473.

[16] Explosion, "spaCy: Industrial-strength natural language processing," 2024. [Online]. Available: https://spacy.io/