24th International Symposium INFOTEH-JAHORINA, 19-21 March 2025

Processing Balkan News Articles Using Selected Methods of Natural Language Processing

Marko M. Živanović*, Nenad Stefanović**
Academy of Technical and Art Applied Studies Belgrade,
The School of Electrical and Computer Engineering, Belgrade, Serbia
Email:*markoz@gs.viser.edu.rs
University of Kragujevac, Faculty of Tehnical Sciences
Čačak, Serbia, **nenad.stefanovic@ftn.kg.ac.rs

***Abstract*— The paper explores topics in newspaper articles using machine learning and natural language processing methods. The focus is on the automatic identification of key topics from the headlines of tabloid articles from Serbia, Bosnia and Herzegovina, and Croatia. Using Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), the study analyzed cross-country differences and similarities in dominant topics. Additionally, the quality of the models was evaluated using coherence, perplexity, Hellinger distance, and the Silhouette score. The paper highlights both the technical and sociological contributions to the analysis of regional media interests, as well as the potential for further research on the influence of media on societal attitudes.**

Keywords*— topic analysis, natural language processing (NLP), LDA, NMF, media content, regional analysis, Hellinger distance, coherence, newspaper articles.

I. Introduction

This paper focuses on topic analysis in newspaper articles using modern methods of machine learning and natural language processing (NLP). The main contribution of the work lies in the development of an automated system that enables the identification of key topics based on the content of newspaper articles, providing insights into the primary interests and concerns of societies in the region. This approach has both technical and sociological dimensions, as it allows for the quantitative monitoring of topics significant to various communities.

The research analyzed data from three popular tabloids from Serbia, Bosnia and Herzegovina, and Croatia, collected via RSS feeds from the Internet. The processing included a total of 93 documents from Serbia, 73 from Bosnia and Herzegovina, and 90 from Croatia, representing a comprehensive sample for regional media content analysis. Article titles were preprocessed by removing stopwords, special characters, and unnecessary elements, ensuring high-quality input for the models.

The methods used in this study include Latent Dirichlet Allocation (LDA) for topic modeling, topic coherence analysis, perplexity, and Hellinger distance for model quality evaluation, as well as the Silhouette score for assessing topic clustering. Additionally, a Non-negative Matrix Factorization (NMF) model was implemented to enrich the analysis further. The results provide insights into the diversity and commonalities of topics among the three analyzed countries, shedding light on dominant interests within the media space. Beyond the technical contribution to the application of advanced natural language processing methods, the paper opens avenues for further research on the impact of media topics on societies in the region, with a particular focus on their perception and effects on public opinion. The paper consists of a review of similar studies based on sociological and social phenomena, the mathematical preparation of data for analysis, natural language processing algorithms, evaluation metrics of machine learning such as Perplexity and Silhouette, a discussion of the results and related works, as well as the conclusion.

II. Related Works

Research in the field of NLP in the Balkans provides valuable insights into the development of resources and methods for the languages of this region. Kosmajac and Kešelj (2023) analyzed the application of the *TextRank* algorithm for extractive summarization of Serbian texts, using a dataset from online sources in Bosnia and Herzegovina [1]. Marovac et al. highlight that the Serbian language, as a highly inflected and under-resourced language, poses challenges for NLP, but note significant initiatives in corpus and method development over the last three decades [2]. In the context of the COVID-19 pandemic, Beliga et al. analyzed Croatian media content using NLP techniques, pointing to the overlap of terminology during the pandemic waves and the importance of named entity recognition [3]. Madjar et al. examined public awareness of air pollution in the Western Balkans through social media and media text analysis, emphasizing the link between public discussions and pollutant concentrations [4]. Furlan et al. developed a methodology for measuring semantic similarity in short Serbian texts, using a custom corpus to assess system performance [5]. Additional research includes analyses of media narratives and their impact on social processes in the former Yugoslavia, such as Marko studies on reporting civil rights, and Abazi and Doja work on the representation of the Balkan Wars in the media [6,7], while Brnović and Marija explore post-war media manipulations. Andreasen et al. discuss the role of journalism in transitional societies of the Western Balkans [8], highlighting the synergy between media and politics and the potential for the development of peace journalism [9]. These studies emphasize the significance of NLP tools and analyses in studying language, media, and social processes in the region.

This paper stands out by analyzing tabloid articles from three Balkan countries (Serbia, Bosnia and Herzegovina, and Croatia), providing a cross-country comparison of dominant topics. Unlike previous studies, it uses both LDA and NMF for topic modeling and evaluates model quality using multiple metrics (coherence, perplexity, Hellinger distance, and Silhouette score). Additionally, it connects the results to societal interests, offering unique insights into the influence of media on public opinion in the region.

III. METHODOLOGY

A. Data preparation for analysis

Data preparation for Really Simple Syndication (RSS) feed analysis involves several key steps to transform data from various sources into a format suitable for analytical tasks. Essentially, RSS (Really Simple Syndication) services enable automatic retrieval of the latest posts from news, blogs, forums, and other sources. RSS feeds contain structured information such as headlines, descriptions, links, and publication dates. This information can be extracted and analyzed to track trends, analyze content, or perform other analyses related to textual data.

First, it is necessary to download the RSS feeds from the appropriate URLs. Each RSS address provides news about specific fields, such as politics, health, culture, sports, etc. These data can then be processed to extract relevant attributes such as:

- **Title**: A brief description of the article,
- **Description**: A more detailed text describing the article's content,
- **Link**: A URL pointing to the full article,
- **Publication Date**: The date the article was published;

Once the data is downloaded, the next step is to organize it into an appropriate format, such as an Excel file, to enable further analysis. Data organization involves adding each entity (article) into a table with the four basic attributes: Title, Description, Link, and Publication Date [10].

B. Text preprocessing

Text preprocessing is a crucial stage in NLP, as it ensures that the data is in a format suitable for precise and efficient analysis. Proper preprocessing can significantly improve the performance of machine learning models and other analytical techniques. In this context, the preprocessing process includes several important steps that prepare the text for further analysis, such as topic modeling, text classification, and other tasks.

Tokenization is the first step in the preprocessing process, and its goal is to break the text into basic units called tokens. In the context of natural language analysis, tokens usually represent words, but they can also be punctuation marks or other relevant units.

Mathematically, tokenization can be represented as a function $f(t)$, where t is the input text, and the output is a set of tokens T:

$$f(t) \rightarrow T = \{t_1, t_2, \ldots, t_n\} \quad (1)$$

After tokenization, the next step is the removal of punctuation marks and stop words. Stop words are words that frequently appear in texts but do not provide significant information for analysis (e.g., "and", "or", "the", "to", "of"). By using a predefined list of stop words (STOP_WORDS) and removing punctuation, the quality of the data is improved. Additionally, words shorter than three characters are removed, as they typically do not provide relevant information in the context of text analysis [11].

This step can be mathematically represented as a function$g(T)$, which removes punctuation and stop words from the set of tokens T:

$$g(T) \rightarrow T' = \{t_1', t_2', \ldots, t_m'\} \quad (2)$$

where T' represents the filtered tokens that do not contain stop words or punctuation marks.

Lemmatization is the process of reducing words to their base form (lemma), which reduces the number of variations that the same word can have. For example, words like "running", "ran", and "runs" can be reduced to the common lemma "run". While this step is not explicitly shown in the previous code, it is crucial for further analysis because it ensures that different forms of the same word are treated as identical, reducing the number of unique words (vocabulary).

Mathematically, lemmatization can be represented as a function $h(t_i')$, where each token t_i' from the set T' is reduced to its lemma l_i:

$$h(t_i') \rightarrow l_i \quad (3)$$

Thus T', the set T' becomes a set of lemmatized words L:

$$g(T) \rightarrow L = \{l_1, l_2, \ldots, l_m\} \quad (4)$$

Special characters, such as numbers, symbolic signs, and other unnecessary elements, can introduce noise into the data, which may impact the precision of the models. Using regular expressions (regEx) allows for efficient identification and removal of these characters, thereby cleaning the text and ensuring that only relevant data remain for analysis [12].

This step can be mathematically represented as a function $P(t)$, which applies regular expressions to the set of tokens T' to remove special characters and numbers:

$$p(T') \rightarrow T'' = \{t_1'', t_2'', \ldots, t_k''\} \quad (5)$$

where T'' represents the cleaned text without special characters and numbers.

C. LDA model and NMF technique

The methodology of the work should cover several key areas related to topic modeling in textual data, including LDA, model evaluation using coherence and perplexity, as well as NMF.

LDA is a generative probabilistic model that explains a set of documents as mixtures of topics.

The basic principle of the LDA model is that each document in the corpus represents a mixture of multiple topics, with each topic being represented as a distribution of words. In the LDA model, we assume that documents are generated from several topics, with each topic being a distribution over words. The model uses a Dirichlet distribution to model the distribution of topics in documents and the distribution of words in topics [13].

NMF is a technique used to decompose a matrix into two factors, with all elements of the matrices being non-negative. For textual data, NMF is used as a topic modeling technique, where two matrices are derived: one representing the relevance of topics in documents, and the other representing the word distribution in topics. NMF is similar to LDA but is a deterministic model, unlike the probabilistic LDA model [14].

Mathematically, the NMF model is based on the factorization of the matrix V into two non-negative factors W and H, such that

$$V \approx WH \tag{6}$$

In the context of text analysis:

- V is the word frequency matrix in documents,
- W represents the topic distribution in documents,
- H represents the word distribution in topics;

D. Model evaluation

Additional measures for model evaluation include coherence, which assesses the semantic relatedness of words within topics; perplexity, which measures the model's ability to predict new data; and Hellinger distance, which quantifies the similarity between probabilistic topic distributions.

Perplexity is a metric that measures how well the model predicts new data. It is calculated as the exponential of the negative average log-likelihood of the model for generating words in documents. The formula for perplexity is:

$$Perplexity(D) = \exp\left(-\frac{\sum_{d=1}^{|D|} \log P(w_d)}{\sum_{d=1}^{|D|} N_d}\right) \tag{7}$$

Where is :

- D is collection of documents (corpus),
- $P(w_d)$ is likelihood of words in document d under the model,
- N_d is total number of words in document d ;

A **lower perplexity score** indicates that the model is better at predicting the words in unseen documents, meaning it is more effective at capturing the topics in the data [15].

Coherence measures the semantic similarity between words in a topic, i.e., how meaningfully related the words in a topic are. Several methods exist to calculate coherence, with the most common being the method that measures pointwise mutual information (PMI) between words [16].

Pointwise Mutual Information (PMI) is typically calculated as the average PMI over all pairs of words within a topic:

$$Coherence(T) = \frac{1}{|W_T| \cdot (|W_T| - 1)} \sum_{w_i, w_j \in W_T, i \neq j} \log \frac{P(w_i, w_j)}{P(w_i) \cdot P(w_j)} \tag{8}$$

Where:

- T represents the topic,
- W_T is the set of words in topic T ,
- $P(w_i, w_j)$ is the probability of co-occurrence of words w_i and w_j within a sliding window in the corpus,
- $P(w_i)$ and $P(w_j)$ are the individual probabilities of w_i and w_j occurring in the corpus;

This formula measures how much more likely two words are to co-occur in the same context than would be expected by chance. A higher coherence score indicates a stronger semantic similarity between the words in the topic [17].

When applied to NLP, the **Silhouette Score** measures how similar a text element (like a document or word) is to other elements within the same cluster (cohesion) compared to how similar it is to elements in other clusters (separation). In NLP tasks, this can be useful for evaluating document clustering, topic modeling, or word embedding clustering, where the goal is to group similar textual data together [18].

Mathematically, for a given element (e.g., a document or word), the Silhouette score is computed as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{9}$$

The following definitions apply:

- $a(i)$ is the average distance between element i and all other elements within the same cluster, reflecting the cohesion of the cluster,
- $b(i)$ is the average distance between element i and all elements in the nearest cluster (excluding the cluster to which i belongs), indicating the separation between clusters;

IV. Results and Discussion

This research yielded several key results for text analysis using topic modeling methods, with the results presented for three different regions: Serbia, Croatia, and Bosnia and Herzegovina. These results provide valuable insights into the efficiency and quality of the topic analysis models and can be used for further discussion of the various applied methods.

The results are provided in Table 1.

TABLE I. Results of Topic Modeling Analysis by Region

Region	Coherence	Perplexity	Silhouette Score
Serbia	0.544048	-9.124896	0.503288
Croatia	0.539516	-8.823417	0.626429
Bosnia & Herzegovina	0.656993	-7.812511	0.819308

As shown in Figure 1, a higher coherence score indicates that the words grouped within the same topic are more closely related. The coherence results are quite good overall, with Bosnia and Herzegovina achieving the highest score of 0.656993, indicating that the topics in this region are the most well-connected. Serbia and Croatia have slightly lower coherence scores but still fall within a good range, allowing us to conclude that all regions produced meaningful thematic models.

Your perplexity results are negative (which is typical for log-perplexity), but all regions demonstrated similar values in terms of perplexity. Bosnia and Herzegovina had the lowest value (**-7.812511**), indicating that the model for this region was slightly better at predicting new documents.

The Silhouette score is a metric used to evaluate the quality of clusters (topic models in this case). Values closer to 1 indicate well-separated clusters. Bosnia and Herzegovina has the highest Silhouette score (**0.819308**), indicating that topics in this region are the most well-separated. Serbia (**0.503288**) has somewhat lower results, which may suggest that the topics in this region are less distinct or the clusters are less cohesive. Croatia falls in between with a score of **0.626429**, indicating moderately well-separated topics.

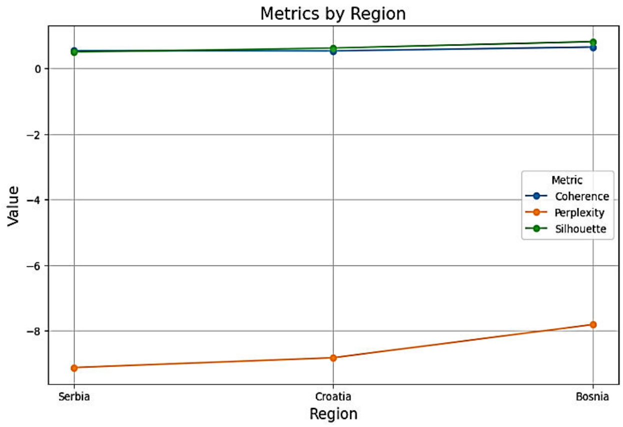

Figure 1. Metrics by Region (Coherence, Perplexity, Silhouette)

As shown in Figure 2, the histogram distribution of documents by topics for Serbia reveals that Topic 3 has the highest number of documents, making it the most prevalent topic in Serbian media. Topic 1 and Topic 5 are also present in a larger number, though not as dominant as Topic 3. Similarly, as seen in Figure 3, the distribution of topics for Croatia shows that Topic 3 leads in the number of documents, with Topic 2 and Topic 1 also having significant representation, indicating a wider distribution of topics. In Figure 4, the distribution of topics for Bosnia and Herzegovina highlights that Topic 1 and Topic 4 dominate, being the most prevalent in the analyzed documents, suggesting specific interest areas recognized in the media of this country.

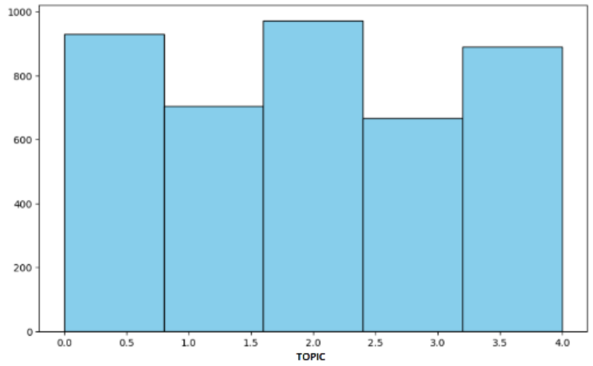

Figure 2. Histogram distribution of documents by topics for Serbia

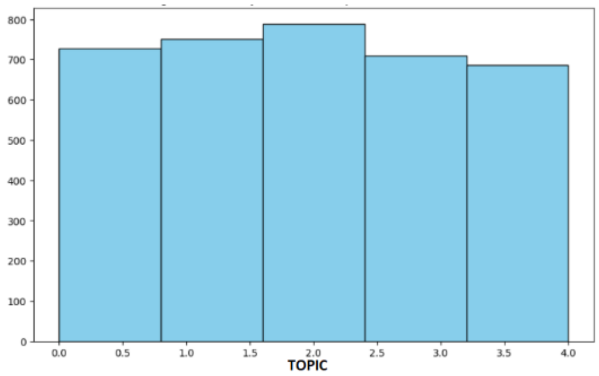

Figure 3. Histogram distribution of documents by topics for Croatia

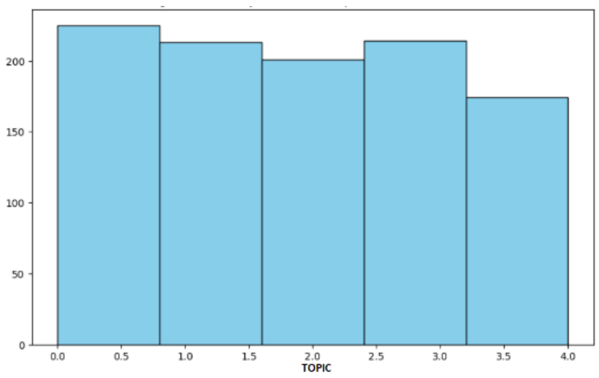

Figure 4. Histogram distribution of documents by topics for Bosnia and Herzegovina

As shown in Figure 5, the analysis of topic similarities across Serbia reveals that Topics 3 and 1 have the highest similarity (1.00), indicating a strong overlap in content. The similarity between Topic 3 and Topic 2 is also high, though slightly lower. In Figure 6, the topic similarity for Croatia shows that Topics 1 and 2 have a high similarity (0.99), and Topics 4 and 1 also exhibit strong similarity. However, the strongest connection is observed between Topics 4 and 5, as indicated by the maximum value in the heatmap. In Figure 7, the topic similarity for Bosnia and Herzegovina highlights a very high similarity between Topics 1 and 4, as well as between Topics 1 and 2 (up to 0.99), suggesting a significant overlap in the words and themes they address.

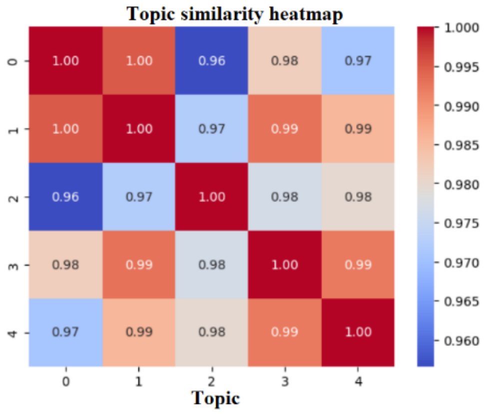

Figure 5. Topic similarity heatmap for Serbian tabloids

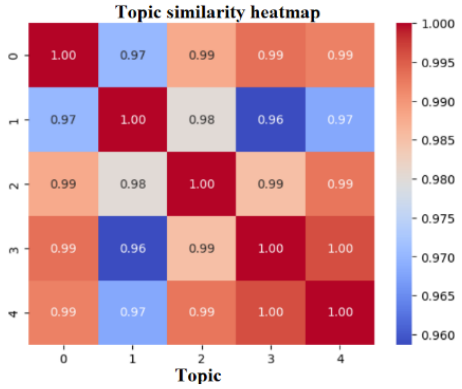

Figure 6. Topic similarity heatmap for Croatian tabloids

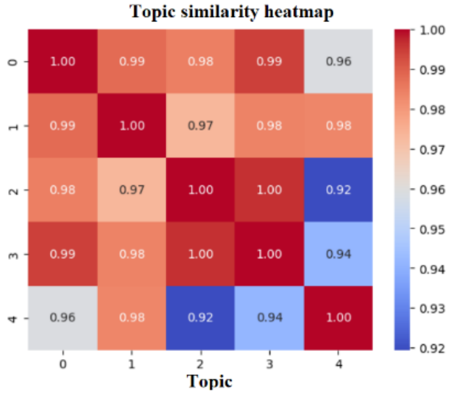

Figure 7. Topic similarity heatmap for Bosnian tabloids

In the analysis of the most prominent words in topics across **Serbia, Croatia, and Bosnia and Herzegovina**, distinct patterns are observed. In **Serbia**, the most notable words include **Vučićević, Vučić, Novi Sad, Novak Đoković, Saobraćajna nesreća, foto, and Nova Godina**, indicating a dominance of political and sports topics in Serbian tabloids. In **Croatia**, prominent words such as **Zagreb, Hajduk, Hrvatska, and Evro** reflect the significance of sports and political themes. In Bosnia and Herzegovina, the key words like **Žena, Svijet, Muškarac, Sarajevo, Hercegovina, and Bosna** point to topics related to social and political events in the country.

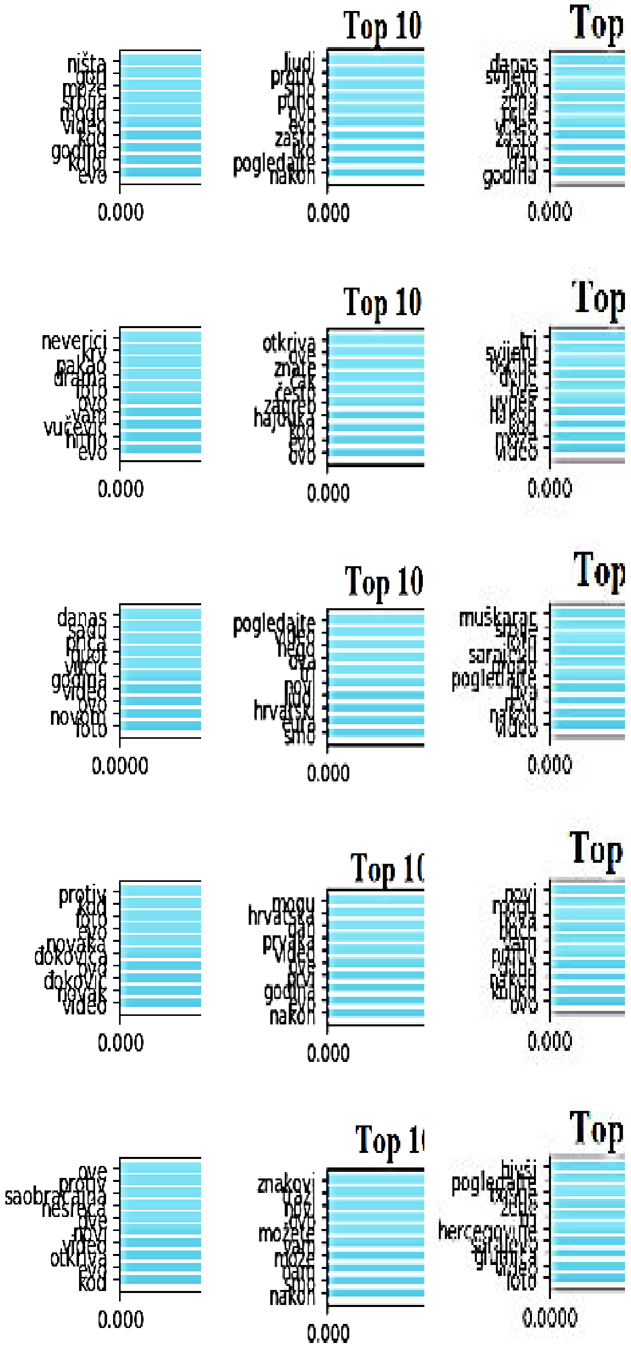

Figure 8. The top 10 frequent words for the topics extracted from Serbian, Croatian, and Bosnian tabloids

V. CONCLUSION

This paper presents the development of an automated system using machine learning techniques—LDA and NMF—to analyze topics within newspaper articles from Serbia, Bosnia and Herzegovina, and Croatia. The study offers a novel approach to identifying and comparing dominant media topics across these countries, providing insights into the interests shaping public discourse. Evaluation of model quality using metrics like coherence, perplexity, Hellinger distance, and the Silhouette score ensures a robust framework for assessing topic modeling performance. The sociological implications are significant, highlighting the commonalities and differences in media content, and offering a deeper understanding of the regional media landscape. However, the research is limited by a small sample size and challenges arising from linguistic variations in the Balkans, which may impact topic modeling accuracy. Additionally, while quantitative metrics are valuable, they don't fully capture topic relevance or societal impact, and further qualitative analysis could enhance the findings. Future research could expand the analysis to other countries and media types, incorporate NLP techniques like sentiment analysis, and explore how topics evolve over time in response to socio-political events, thereby offering a more comprehensive understanding of media's influence on public opinion and societal attitudes in the Balkans.

REFERENCES

[1] Kešelj, V. (2023, March). Automated Authorship Attribution using CNG Distance on Blog Posts in the Serbian Language. In 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-8). IEEE.

[2] Marovac, U. A., Avdić, A. R., & Milošević, N. L. (2023). A Survey of Resources and Methods for Natural Language Processing of Serbian Language. arXiv preprint arXiv:2304.05468.

[3] Beliga, S., Martinčić-Ipšić, S., Matešić, M., Petrijevčanin Vuksanović, I., & Meštrović, A. (2021). Infoveillance of the Croatian online media during the COVID-19 pandemic: one-year longitudinal study using natural language processing. JMIR public health and surveillance, 7(12), e31540.

[4] Madjar, A., Gjorshoska, I., Prodanova, J., Dedinec, A., & Kocarev, L. (2023). Western Balkan societies' awareness of air pollution. Estimations using natural language processing techniques. Ecological Informatics, 75, 102097.

[5] Furlan, B., Batanović, V., & Nikolić, B. (2013). Semantic similarity of short texts in languages with a deficient natural language processing support. Decision Support Systems, 55(3), 710-719.

[6] Marko, D. (2012). Citizenship in media discourse in Bosnia and Herzegovina, Croatia, Montenegro, and Serbia.

[7] Abazi, E., & Doja, A. (2017). The past in the present: time and narrative of Balkan wars in media industry and international politics. Third World Quarterly, 38(4), 1012-1042.

[8] Brnović Milorad, M. (2023). Nation, religion and manipulation: Post-war media scene in the countries of the former Yugoslavia.

[9] Andresen, K., Hoxha, A., & Godole, J. (2017). New roles for media in the Western Balkans: A study of transitional journalism. Journalism Studies, 18(5), 614-628

[10] Petrov, A., & Macdonald, C. (2024). RSS: effective and efficient training for sequential recommendation using recency sampling. ACM Transactions on Recommender Systems, 3(1), 1-32.

[11] Gastaldi, J. L., Terilla, J., Malagutti, L., DuSell, B., Vieira, T., & Cotterell, R. (2024). The foundations of tokenization: Statistical and computational concerns. arXiv preprint arXiv:2407.11606.

[12] Gulu-zada, F., Fataliyev, F., Huseynova, G., & Rustamov, S. (2024, September). A hybrid approach to Azerbaijani lemmatization: integrating rule-based methods with ML PoS tagging. In 2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-5). IEEE.

[13] Tian, C., Zhang, J., Liu, D., Wang, Q., & Lin, S. (2024). Technological topic analysis of standard-essential patents based on the improved Latent Dirichlet Allocation (LDA) model. Technology Analysis & Strategic Management, 36(9), 2084-2099.

[14] Huang, Z., Cai, D., & Sun, Y. (2024). Towards more accurate microbial source tracking via non-negative matrix factorization (NMF). Bioinformatics, 40(Supplement_1), i68-i78.

[15] Ankner, Z., Blakeney, C., Sreenivasan, K., Marion, M., Leavitt, M. L., & Paul, M. (2024). Perplexed by Perplexity: Perplexity-Based Data Pruning With Small Reference Models. arXiv preprint arXiv:2405.20541.

[16] Alnatah, H., Yao, Q., Beaumariage, J., Mukherjee, S., Tam, M. C., Wasilewski, Z., ... & Snoke, D. W. (2024). Coherence measurements of polaritons in thermal equilibrium reveal a power law for two-dimensional condensates. Science Advances, 10(18), eadk6960.

[17] Zheng, S., Kwon, Y., Qi, X., & Zou, J. (2024). Truthful Dataset Valuation by Pointwise Mutual Information. arXiv preprint arXiv:2405.18253.

[18] Vardakas, G., Papakostas, I., & Likas, A. (2024). Deep Clustering Using the Soft Silhouette Score: Towards Compact and Well-Separated Clusters. arXiv preprint arXiv:2402.00608.