Creating a Digital Archive: Segmenting and Transcribing Articles from the Daily Newspaper 'Bujku'

Kushtrim Visoka

September 29, 2024

Abstract

This paper details the creation of a dataset consisting of segmented articles from historical issues of the daily newspaper 'Bujku,' a crucial source of independent information for Kosovo Albanians during the 1990s under Serbian oppression. The dataset was developed by segmenting articles from scanned issues and applying optical character recognition (OCR) using Gemini Flash 1.5 and Tesseract. Additionally, dense vector representations of the articles were generated, and a user-friendly platform was constructed for browsing and performing semantic searches.

Key challenges included low-quality scans, OCR difficulties with articles featuring multiple columns and unclear boundaries. The resulting dataset includes over 130,000 segmented and transcribed articles from 1991 to 1998, serving as a valuable resource for researchers, historians, and digital humanities scholars exploring Kosovo's sociopolitical context during this period.

1 Introduction

After the violent and illegal suppression of Kosovo's autonomy by Serbian authorities in 1989, a repressive regime was imposed in Kosovo that aimed to crush all forms of resistance and gain full control over the public and institutional life of Kosovo Albanians [1]. As part of these measures, almost all written and audiovisual media were shut down, leaving citizens without independent sources of information [2]. In this harsh context of censorship and pressure, a bold effort was made to keep information in the Albanian language alive.

'Bujku' (*'The Farmer'* in English), originally an annex of 'Rilindja', which primarily focused on agricultural issues and rural life, was transformed into a daily newspaper covering key issues of the time [3]. Despite facing extraordinary challenges such as daily threats, arrests of journalists, and continuous pressure from Serbian forces, 'Bujku' continued to be published and distributed. The newspaper quickly became an important voice for Kosovo Albanians, providing reliable information and maintaining the spirit of peaceful resistance.

Majority of the journalists contributing to 'Bujku' were from 'Rilindja', the largest media outlet in Kosovo until it was forcibly banned by the Serbian regime. They were dedicated to reporting on events and developments in Kosovo, despite numerous difficulties. Essentially, 'Bujku' became the primary medium of peaceful civil resistance in Kosovo, informing the public and conveying resilience and determination in the face of repression. For many years, 'Bujku' remained a beacon of hope and solidarity for Kosovo Albanians during this dark historical period.

The Non-Governmental Organization (NGO) Free Libre Open Source Software Kosova (FLOSSK) scanned issues of the daily newspaper 'Bujku' from the period 1991 to 1998 and published them on their website, significantly contributing to the preservation of historical archives. Additionally, they have performed OCR on all the pages of the newspaper, making the text searchable and readable in digital format. However, the newspaper's articles have not been segmented and separated as individual text from other articles found on the same full page. For this reason, this dataset was created to ensure that the newspaper articles are segmented and processed through OCR, separated from other articles within the page, thereby enabling archiving and data access at the article level. This advanced segmentation and organization provides a valuable resource for researchers, historians, and any interested parties seeking to explore and analyze the newspaper's content in a detailed and accurate manner. This initiative helps in preserving and promoting Kosovo's cultural and historical heritage during one of the most sensitive periods in its history.

2 Dataset Creation Process

2.1 Data Collection

The files were obtained from the publicly accessible platform hosted by FLOSSK at `https://books.flossk.org/gazetat/`. This platform provides a comprehensive archive of the daily newspaper 'Bujku', enabling researchers and the public to access historical issues. A total of 1834 issues were manually downloaded and organized into a structured directory based on publication date to facilitate efficient processing and retrieval.

The files were preserved with their original filenames as provided by FLOSSK, adhering to the format `BU-YYYYMMDD.pdf`, where `BU` denotes 'Bujku' and the date indicates the publication date of each issue. This naming convention ensured that each file was easily identifiable and maintained the historical context of the publication. However, many issues were missing from the archive, creating gaps in the dataset, which may limit researchers' ability to fully analyze the historical record from this period.

2.2 Segmentation Methodology

Segmenting the articles from the scanned issues of 'Bujku' was one of the most complex tasks due to the newspaper's intricate layout. Articles often spanned multiple columns, and content such as advertisements or editorials frequently overlapped, making it difficult to accurately define article boundaries. Various automatic segmentation tools were tested but proved inadequate, as they failed to reliably detect boundaries and often resulted in fragmented or misclassified text.

As a result, manual segmentation was necessary. Each page was carefully reviewed, with articles delineated based on visual cues such as headlines and formatting differences. This approach allowed for a more accurate segmentation, despite the variability in layouts across different issues, such as differences in column width, font size, and content placement. After completing this labor-intensive process, a total of 130,048 articles were successfully segmented, ensuring the accuracy and integrity of the dataset.

2.3 Data Processing

Converting the segmented articles into a digital format using OCR presented further challenges, particularly due to the poor quality of the original newspaper pages, either because of aging or the quality of the paper itself. Many issues exhibited poor resolution, ink bleed-through, and other artifacts that made text recognition difficult. Initially, Tesseract was used for OCR, but it struggled with the poor scan quality, leading to a high number of errors, especially in articles with complex layouts or faint text.

To resolve these challenges, Gemini Flash 1.5 was introduced, providing superior accuracy in handling multi-column layouts and in recognizing Albanian special characters (e.g., ç, ë) as well as older typographical symbols (« and »). Despite this improvement, a small number of articles remained problematic and were processed using Tesseract as a fallback, though less than 5% of the dataset relied on it. The combination of Gemini Flash 1.5 and thorough post-OCR text cleaning resulted in a high-quality, reliable dataset.

2.4 Embedding Extraction Using intfloat/multilingual-e5-large

To enable advanced semantic analysis and facilitate research on the segmented articles, text embeddings were extracted using the `intfloat/multilingual-e5-large` model. This model, described in the technical report by Wang et al. (2024) [4], is designed to produce high-quality text embeddings across multiple languages, making it well-suited for processing the Albanian text in the 'Bujku' dataset.

The model generates dense vector representations of the articles, capturing their semantic content in a high-dimensional space. These embeddings allow for various downstream tasks such as clustering articles by topic, detecting thematic trends, and performing large-scale similarity searches. By leveraging the

multilingual capability of the model, the specific characteristics of the Albanian language, including unique characters and context, are effectively encoded.

Using the `intfloat/multilingual-e5-large` model enhances the utility of the dataset for researchers and historians by enabling more sophisticated computational analysis, including linguistic trend identification, sociopolitical discourse analysis, and historical content clustering. The embeddings provide a robust foundation for exploring relationships between articles and understanding the evolving narrative of 'Bujku' over time.

3 Platform for Browsing and Semantic Search

In addition to the creation of the dataset, a web platform, `https://bujku.news`, has been developed to make the segmented articles from 'Bujku' easily accessible to the public. Built using React, this platform enables users to explore the digitized content of the newspaper, offering powerful semantic search capabilities based on the extracted embeddings. By leveraging the `intfloat/multilingual-e5-large` model, users can conduct searches that go beyond simple keywords, allowing for deeper exploration of the articles' meaning and context.

4 Conclusion and Future Work

The segmented 'Bujku' dataset is an important resource for preserving historical journalism from Kosovo during the 1990s. It provides useful insights into the sociopolitical climate of the time and makes it easier for researchers and historians to access and study this period in detail.

Despite the high quality of the resulting dataset, a few limitations remain. The absence of many newspaper issues due to incomplete archival efforts leaves gaps in the historical record. Additionally, while manual segmentation and advanced OCR tools resulted in high-quality text extraction, further automation in article segmentation could improve efficiency in future projects.

Future work could focus on adding metadata for each article, improving segmentation accuracy with newer machine learning models, and expanding the dataset by incorporating missing issues, should they become available. This dataset could also serve as the foundation for various research applications, such as linguistic analysis, historical trend studies, and sociopolitical discourse analysis.

References

[1] Clark, Howard. *Civil Resistance in Kosovo*. Pluto Press, 2000. JSTOR, `https://doi.org/10.2307/j.ctt18fsc6d`. Accessed 29 Sept. 2024.

[2] Gazeta Bujku, 18 Janury 1991, page 1.

[3] Sinani, Resul. "The Challenges on Native Language Information During Occupation: Metamorphoses of the Albanian Language Media in Kosovo." *European Journal of Research and Reflection in Arts and Humanities*, vol. 3, no. 3, 2015.

[4] Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. "Multilingual E5 Text Embeddings: A Technical Report." *arXiv preprint arXiv:2402.05672*, 2024.