Thematic Landscapes of the Past: Analysing Slovene Historical Periodicals With Topic Modelling

Filip Dobranić[1], **Uroš Šmajdek**[3], **Oliver Pejić**[1], **Ciril Bohak**[3], **Vojko Gorjanc**[1,2], **Tina Munda**[2], **Darja Fišer**[1]
[1] Institute of Contemporary History
Privoz 11, SI-1000 Ljubljana
{filip.dobranic, oliver.pejic, darja.fiser}@inz.si
[2] Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
{tina.munda, vojko.gorjanc}@ff-uni-lj.si
[3] Faculty of Computer Science, University of Ljubljana Večna pot 113, SI-1000 Ljubljana
{uros.smajdek, ciril.bohak}@fri.uni-lj.si

Abstract

This paper explores the thematic landscapes of three Slovene historical periodicals—*Slovenka*, *Slovenec*, and *Slovenski narod*—from the *sPeriodika* corpus, a comprehensive collection of Slovene press published between 1771 and 1914 (Dobranić et al., 2024). Using BERTopic, we analyse the thematic profiles of these periodicals, enriched with diachronic perspectives. Our study examines the thematic commonalities and specificities of the selected periodicals, highlighting their distinct political orientations, target audiences, and the increasing nationalist polarisation in public discourse. This work contributes to digital humanities by demonstrating the potential of modern topic modelling techniques, such as BERTopic, to advance historical and cultural research.

Keywords: historical newspapers, digital humanities, diachronic analysis, topic modelling, BERTopic, collective identities, sPeriodika

1. Introduction

We present a thematic comparative analysis of three periodicals (Slovenka, Slovenec, and Slovenski narod) in the public domain from the *sPeriodika* corpus, a comprehensive collection of Slovene historical press published between 1771 and 1914 (Dobranić et al., 2023).

Using the popular transformer-based topic modelling framework BERTopic (Grootendorst, 2022), we examine these periodicals to answer the following questions:

1. What are the thematic profiles, commonalities and specificities of the selected periodicals?

2. What is the thematic landscape of the selected newspapers over time, from their inception until 1914 (copyright expiration cutoff)?

By addressing these questions, we explore how these periodicals represented and influenced the cultural and political dynamics of their time.

2. Related Work

Research in historical newspapers has increasingly integrated topic modelling into collaborative digital humanities workflows, bringing together humanities scholars, computer scientists, and information specialists (Oberbichler et al., 2022; Villamor Martin et al., 2023). These interdisciplinary infrastructures are based on processes such as digitization, metadata enrichment, and post-OCR correction, which provide the foundational methods for large-scale computational analyses including topic modelling (Lombardi and Marinai, 2020).

Topic modelling has emerged as one of the key methods for uncovering long-term thematic patterns in such studies, with evaluations of its performance in the context of historical (newspaper) data. For example, Murugaraj et al. (2025b) evaluated topic modelling approaches for newspaper archives, comparing traditional probabilistic Latent Dirichlet Allocation (LDA), matrix factorization-based Non-Negative Matrix Factorization (NMF), and neural-based models such as BERTopic (Grootendorst, 2022). Their findings demonstrate that BERTopic outperforms classical models in all tested aspects, particularly in contextual sensitivity and thematic coherence. In subsequent studies, Murugaraj et al. (2025a,c) highlighted how traditional topic modelling methods often fail to fully capture the dynamic and complex nature of discourse in historical texts.

By contrast, BERTopic proved to be effective in identifying the most relevant topics for specific queries and in restricting retrieval to documents or segments within those topics. Ginn and Hulden (2024) applied both traditional statistical models

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 92–99
May 16, 2026.

(LDA and NMF) and BERT-based models to historical literary texts. Although quantitative metrics tended to favour statistical models, their qualitative evaluation revealed that neural models provided deeper insights into the data, highlighting the potential of modern transformer-based approaches for historical text analysis.

3. Data and Methodology

3.1. Dataset

Our analysis is based on a 1 bn word corpus *sPeriodika* of Slovene periodicals published between 1771 and 1914 (Dobranić et al., 2023), from which we selected three periodicals based on their historical significance:

- *Slovenka* [The Slovene Woman] (1897–1902)
- *Slovenec* [Slovene] (1873–1945)
- *Slovenski narod* [The Slovene Nation] (1868–1943)

Slovenec and *Slovenski narod* were the two most prominent and widely read Slovene-language political dailies during the turn of the century, catering to readerships with opposing political views. While *Slovenec* served as the leading voice of Slovene political Catholicism, *Slovenski narod* was closely aligned with liberal-progressive politics. Both newspapers eventually became the official organs of the Slovene Catholic and liberal parties respectively (Amon and Erjavec, 2011).

The principal difference between the two newspapers laid in their stance towards secularism and the Church's role in society. *Slovenec* campaigned for preserving the Church's independence against state meddling and its supremacy in education and civic life. While it also supported Slovene nationalist demands, its primary discursive enemy was liberalism, which it often equated with German politics (Amon and Erjavec, 2011, 144-147). Conversely, *Slovenski narod*'s rhetoric placed greater emphasis on Slovene nationalism and heavily criticized the unequal status of Austria's non-dominant nationalities. Its main discursive enemies were German nationalism and Slovene political Catholicism, and its core readership consisted of educated professionals as well as wealthier peasants (Amon and Erjavec, 2011, 120-130).

Slovenka began as a supplement to the liberal Slovene newspaper *Edinost* [Unity] from Trieste and later became an independent monthly publication. While its content initially mostly focused on literature, nationalist politics and domestic life, it also published more nuanced commentary on women's and social issues during its last two years. It was published for a much shorter period compared to the other two newspapers but deserves special attention due to its specificity as the first female-oriented and female-edited journal in the history of Slovene journalism (Amon and Erjavec, 2011, 136-138). The size of each dataset in the number of tokens, paragraphs and issues is presented in Table 1.

The corpus itself as well as individual newspaper subcorpora are not evenly distributed through time. *Slovenka* shows a slight decline from its initial to final year, consistent with its start as a biweekly supplement and transformation to a monthly publication in 1900. Opposite to that both *Slovenec* and *Slovenski narod* show a gradual increase of the number of paragraphs through the years, consistent with their development and growth in the second half of the 19th century. In order to control for the uneven distribution of paragraphs in our subcorpora, the analysis presented discusses relative paragraph frequencies when engaging in diachronic thematic analysis in section 5

While all paragraphs were considered in the analysis of *Slovenka*, the analysis of *Slovenec* and *Slovenski narod* excluded a third of the paragraphs from our analysis (see section 3.2). The shape of yearly distribution of excluded paragraphs matches the distribution of all paragraphs in the newspaper which indicates that the paragraphs from merged topics are diachronically representative of the corpus.

	Tokens	Paragraphs	Issues
S. narod	183,294,799	4,404,531	14,039
Slovenec	137,506,802	3,158,842	10,897
Slovenka	1,633,570	56,330	113

Table 1: Corpus size.

3.2. Topic Modelling

We use BERTopic (Grootendorst, 2022) to model the topics in each of the periodicals on individual paragraphs. We use the linguistically annotated texts from sPeriodika in the CONLL-u format and use the paragraph annotations to extract individual paragraphs. These are then assigned metadata from the periodical (issue date, periodical name, text, lemmatised text, paragraph annotations etc.) which is then used for our analysis.

Each of the periodicals is modeled individually using the same set of parameters and random seeds for the UMAP (McInnes et al., 2018) and the topic model. We used the `paraphrase-multilingual-MiniLM-L12-v2` model (Reimers and Gurevych, 2019) for our embeddings and the parameters presented in Table 2.

The model produced 554 topics on *Slovenka*. Due to their larger size, modelling for *Slovenec* and

Parameter name	Parameter value
UMAP `n_neighbors`	15
UMAP `n_components`	5
UMAP `min_dist`	0.1
UMAP `metric`	cosine
Topic model `top_n_words`	100

Table 2: Topic modelling parameters.

Slovenski narod returned two orders of magnitude more topics.

Given these differences, we needed a more manageable number of less fine-grained thematic clusters that would be easier to compare across the periodicals and across time. Furthermore, after manually inspecting the topics we observed the model splitting otherwise thematically-coherent topics based on proper nouns (for example country names). This was expected due to BERTopic's use of cTF-IDF, but ultimately, we were interested in thematic areas of reporting like State administration regardless of the country the paragraph is referring to. In order to merge these topics we tested automated hierarchical clustering but it produced unsatisfactory results. Instead, we decided to group the individual topics into manually curated "themes".

In order to determine the viability of a manual approach we started by manually grouping all 554 topics from *Slovenka* and merged them into the resulting themes. These were created through the manual grouping process by viewing representative words and close reading representative paragraphs of BERTopic-suggested topics. We observed that when the number of paragraphs in the topic starts approaching the minimum, the content and the topics themselves become more likely to contain text fragments and increadingly hard to thematically categorise. The latter, along with our estimates of the labor required to merge all the topics produced for *Slovenec* and *Slovenski narod* led us to consider the 500 most-represented topics in these newspapers, comprising roughly two thirds of all paragraphs present. The rest of the paragraphs were excluded from our analysis. The coarse-grained themes are presented in Table 3 and typeset in Smallcaps. In this paper we use the coarse-grained themes for analysis. However, the original individual topics can still be zoomed in for more in-depth analysis, which is planned as future work.

After grouping the topics, as a filtering criterion, we first excluded out of scope paragraphs (uncategorised topics from *Slovenec* and *Slovenski narod*), followed by textual fragments, and garbled text due to OCR errors, which comprise the theme Outliers and noise and represent 58.3% of all analysed paragraphs in *Slovenka*, 85.1% of *Slovenec*, and 88.4% of paragraphs in *Slovenski narod*. The final set of themes used in our analysis is presented in Figure 1.

	Themes
Slovenski narod	Advertisements and announcements, Art and culture, Countries and nationalities, Criminality and natural disasters, Education, Family, Finance, Food production, Health and mortality, Infrastructure, Narrative, Nature and weather, Newspaper publishing, Outliers and noise, Occupations, Paratext, Political life, Religious practice, Social life, State administration, Travel and communications
Slovenec	Advertisements and announcements, Art and culture, Countries and nationalities, Criminality and natural disasters, Education, Family, Finance, Food production, Health and mortality, Narrative, Nature and weather, Newspaper publishing, Outliers and noise, Non-Slovene text, Occupations, Paratext, Political life, Religious practice, Slovene identity, Social life, State administration, Travel and communication
Slovenka	Abroad, Art, Body and emotion, Culinary arts, Family, Female identities, Group identities, Material culture and objects, Metatext, Morals, Narrative, Nature and environment, Newspaper publishing, Outliers and noise, Non-Slovene text, Occupations, Paratext, Railway, Religious practice, Russian culture, Slovene language, State institutions, Time

Table 3: Topic groups per periodical in alphabetical order.

4. General Thematic Analysis

The themes identified in each of the three periodicals are visualised in Figure 1. Due to *Slovenka* being two orders of magnitude smaller than the other two newspapers, theme size and theme rankings are not directly comparable across periodicals. The population of themes *Slovenka* can be as low as 28 paragraphs for themes such as Slovenian language. For cross-periodical comparisons, we use relative frequencies and shares.

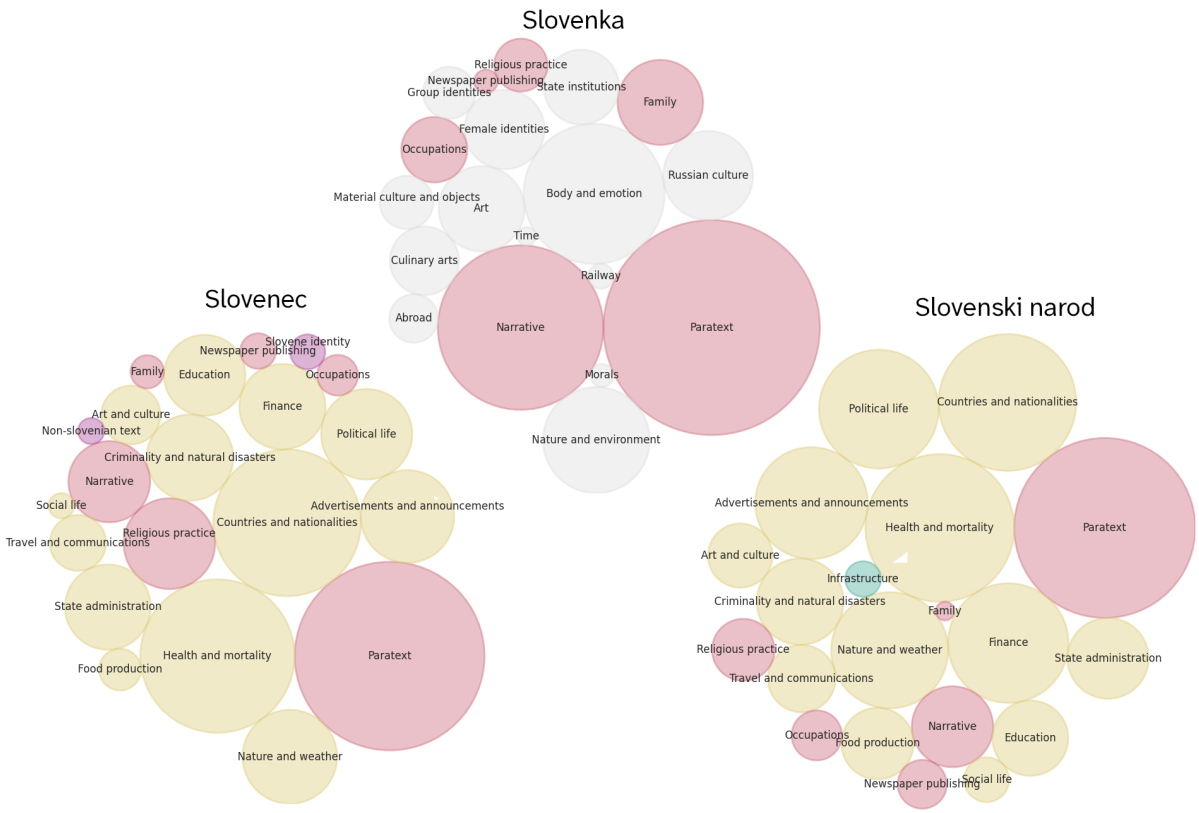

Figure 1: Circle-packed themes (excluding OUTLIERS AND NOISE) per periodical. Circle size indicates share of paragraphs. Themes in red are shared across all 3 periodicals. Themes in yellow are shared between *Slovenec* and *Slovenski narod*. Other colours signify themes unique to that periodical.

4.1. Shared themes

The list of all six themes with absolute and relative paragraph counts that are present in all three periodicals is presented in Table 4.

PARATEXT: Metatextual content, such as titles, attributions, places and dates from correspondents' reports dominate in all three periodicals, which is not surprising, given the journalistic genre of our dataset.

NARRATIVE: Textual fragments describing situations with no specific contextual cues are more frequent in *Slovenka* than in *Slovenec* and *Slovenski narod*, which reflects differences in reporting style, as well as the fact that *Slovenka* contained a much larger share of literary text.

RELIGIOUS PRACTICE: It is three times as likely for a paragraph to be religiously themed in *Slovenec* compared to *Slovenski narod*, which corresponds with the political profiling of the opposing newspapers. While the theme was identified in *Slovenka* as well, references to religion in this newspaper differs almost completely from the other two as it contains almost exclusively listings of Slavic baptismal names in the Catholic calendar, as well as the names of priests, details of their appointments and liturgical duties.

FAMILY: This theme contains references to family relationships, but also home and homemaking. It represents a relatively small portion of *Slovenec* and *Slovenski narod* and even when it does, it is often in texts offering lodging than discussions of familial life. In *Slovenka*, however, we see both a much larger relative presence and a wider array of relationship expressions as well as references to homemaking, motherhood, and care.

OCCUPATIONS: Expectedly, this theme in *Slovenka* is characterised by predominantly feminine occupation variants, contrary to *Slovenec* and *Slovenski narod* where masculine variants are the norm. Another key difference when comparing *Slovenka* with the other two newspapers is that paragraphs about EDUCATION are grouped in this category whereas *Slovenec* and *Slovenski narod* have a standalone theme for it. Not only were discussions on the politics and practice of education so much more prominent in these two newspapers, education-related paragraphs in *Slovenka* are more focused on the day to day profession of teaching as opposed to the politics of it.

4.2. Thematic profiles of *Slovenec* and *Slovenski narod*

While disctinct from *Slovenka*, the thematic profiles of *Slovenec* and *Slovenski narod* are very similar as there are only three marginal themes that are unique to just one of them.

The range of detected themes for *Slovenec* and *Slovenski narod* is much more diverse compared to *Slovenka* and reflects their function as political dailies covering vast aspects of everyday life.

Health and mortality: This is the most represented theme in both newspapers (2.1% vs. 1.4%) and it contains obituaries as well as advertisements for cures and tonics. This is consistent with the structure of newspapers at the time.

Countries and nationalities: Nearly as frequent as the previous one (2% vs. 1.2%), this theme is composed mostly of the reporting on events in foreign lands. As noted in 5.2, while *Slovenski narod* contains fewer paragraphs in this theme on average, we observe much greater spikes in times of significant foreign events.

While both newspapers feature relatively similar amounts of Advertisements and announcements (0.5% ; 1.2%), the liberal *Slovenski narod*, catering to a relatively wealthier (and ultimately progressive) audience, contains more paragraphs in both the Finance (0.5% ; 1.4%) and Political life (0.5% ; 1.3%) themes than the more rural, conservative *Slovenec*.

The topic Nature and weather (0.5% ; 1.3%) in both newspapers primarily features weather forecasts and reports on weather conditions, as well as accounts of interesting or unusual natural phenomena. It also includes reports on animals, ranging from wild animals to pets, with *Slovenski narod* placing a particular emphasis on animal husbandry, which forms part of broader reporting on resource management, such as forestry.

Theme	Slovenka	Slovenec	Slovenski narod
Paratext	2,732 4.84%	64,207 2.21%	60,794 3.06%
Occupations	619 1.10%	3,002 0.10%	4,702 0.24%
Narrative	3,901 6.93%	11,920 0.41%	12,304 0.62%
Newspaper publishing	79 0.14%	2,293 0.08%	4,482 0.22%
Religious Practice	405 0.72%	14,877 0.51%	7,242 0.36%
Family	1,042 0.185%	2,008 0.07%	661 0.03%
Total (all themes)	56,330	2,904,362	1,987,638

Table 4: Absolute and relative paragraph counts for themes present in all three periodicals.

4.3. Thematic profiles of *Slovenka*

Body an emotion and Nature and Environment: These are *Slovenka*'s most prominent unique theme (5% ; 2.9%). They contain references to body parts, emotions, illnesses and similar, which are common in literary texts. The literary character of the content of *Slovenka* is further reinforced by prominent themes such as Art (1.9%) and Russian culture (2%), since the journal also published many translations and discussions of contemporary literature. While Russian culture might as well be subsumed by the Art theme, it was featured so prominently (due to many publications and discussions of Russian literary works) that we decided to present it separately. While we can find mentions of Russian authors in some of the fine-grained topics in Slovenec and Slovenski narod these are nowhere near as prolific as they are in Slovenka.

The second interesting strand of themes relate to content on women's identity, domestic life as well as gender-defined public engagement, which is illustrated by themes such Female identities (1.6%), Culinary arts (1.2%), and Material culture and objects (0.7%), which, while present in the other two newspapers represent a relatively insignificant portion of their content.

5. Diachronic Thematic Analysis

Due to their positioning, we compare *Slovenec* and *Slovenski narod* but analyse *Slovenka* separately due to its shorter publishing period and specificity as a topical women's newspaper.

5.1. Slovenka

When observed diachronically, the distribution of themes confirms established historiographical knowledge regarding a shift in the journal's editorial policy: moving from literary and female-interest content to theoretical feminist and social issues topics.

This shift in editorial policy also manifests itself in the results of our analysis, see Figure 2. The clear predominance of the themes such as Art (2.3% vs. 1.3%), Body and emotion (5.6% vs. 4.3%), Nature (3.3% vs. 2.4%), and Russian culture during Marica Nadlišek Bartol's editorship (1897-1899) point to the overtly literary character of the journal at the time. The relative dominance of Morals (0.2% vs. 0.1%) likewise points to a predominance of literary or didactic content. The first period was also characterized by an emphasis on domesticity-related themes such as Culinary arts (1.6% vs. 0.7%) and Material culture and objects (1% vs. 0.4%).

Conversely, a rise in more elaborate social commentary during Ivanka Anžič Klemenčič's editor-

ship (1900-1902) is illustrated by an increase of themes such as FEMALE IDENTITIES (0.9% vs. 2.5%); STATE INSTITUTIONS (0.7% vs. 2.3%) and OCCUPATIONS (0.8% vs. 1.4%). Some themes take both angles, domesticity and social commentary. FAMILY (2% vs. 1.6%), for example, was more dominant in the early years of *Slovenka*, but experienced a slight uptick in 1901, possibly due to the presence of theoretical texts dealing with women's roles in family life.

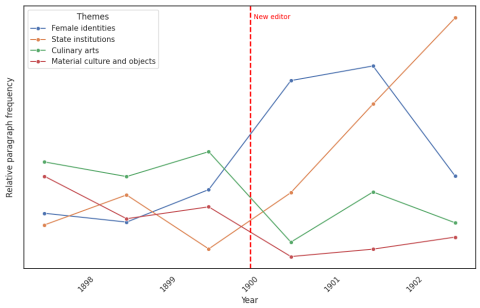

Figure 2: Themes signifying the change of editorial policy of *Slovenka*.

5.2. Slovenec in Slovenski narod

The comparative diachronic thematic analysis of *Slovenec* and *Slovenski narod* is performed by exploring the timelines visualising diachronic behaviour of themes to detect out of the ordinary trends. By zooming in the key terms that characterize the themes, we identify two groups of observations: (1) observations that confirm established historiographical knowledge and align with its expectations, and (2) observations that reveal patterns that could only be discerned by means of distant reading and open potential avenues for deeper historiographic inquiry.

5.2.1. Confirmation of established historiographical knowledge

The diachronic development of the topic COUNTRIES AND NATIONALITIES shows that reporting on developments abroad typically spiked in periods of heightened geopolitical conflict, see Figure 3. The most representative examples include spikes in 1885 (the Mahdist War, more noticeable in *Slovenski narod*), 1900 (the Second Boer War), 1904 (the Russo-Japanese War), 1912-3 (the Balkan Wars), and 1914 (the First World War). As a general tendency, *Slovenski narod* reported with less absolute frequency but with more observable spikes compared to *Slovenec*.

Reporting on the topic RELIGIOUS PRACTICE reveals some insights surrounding the wider ideological framing of the newspapers. As expected, it was more present overall in *Slovenec*, being a Catholic-conservative daily. While the former reported on religious issues with more frequency and discernible annual spikes, the two newspapers nevertheless shared a comparable spike in the years 1903-4, when they both intensely reported on Pope Pius X.

The topic HEALTH AND MORTALITY demonstrates how reporting on epidemic diseases spiked during certain years, e.g. in relation to contemporary cholera and tuberculosis epidemics. While both newspapers reported on this topic with similar frequency, we may observe a disproportionate spike in *Slovenski narod*'s reporting in 1885, which seems to coincide with a cholera epidemic in Trieste. Reporting on CRIMINALITY AND NATURAL DISASTERS likewise reveals spikes that coincide with locally or internationally notorious natural disasters. Some examples of spike years include 1895 (the Ljubljana earthquake); 1906 (likely the San Francisco earthquake, more prominent in *Slovenec*), or 1912 (likely the Mürefte earthquake in Turkey).

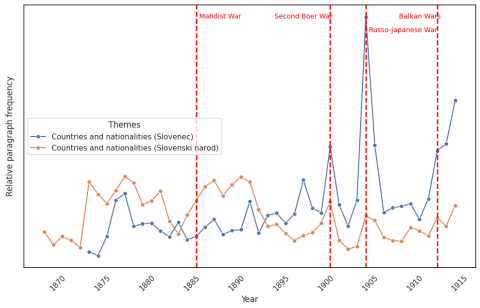

Figure 3: COUNTRIES AND NATIONALITIES theme relative paragraph frequency through time for both *Slovenec* and *Slovenski narod*.

5.2.2. New avenues for research

The topic POLITICAL LIFE reveals some curious differences in reporting among the two newspapers. Throughout the period under investigation and especially before 1907, *Slovenski narod* appears to have reported on political issues more frequently than *Slovenec*. At that point, however, political reporting also increased significantly in *Slovenec* and remained high until the end of our inquiry. One potential explanation for this phenomenon is that the Austrian electoral reform of 1907 marked the beginning of a new period of mass politicization in Austrian society, which could have also manifested itself in heightened political reporting.

Considerable differences are revealed when observing reporting on the topic FINANCE. In *Slovenec*, reporting on the topic began to rise steeply in 1892,

climaxing in 1895, and gradually decreased after stabilization until 1908. Conversely, in *Slovenski narod*, a similar rise began in 1879 and clearly climaxed in 1885, followed by a gradual decline until the end of the graph. The topic ADVERTISEMENTS AND ANNOUNCEMENTS also shows diverging trends in the two newspapers. In *Slovenec*, it rose sharply in 1882, climaxed in 1885 and gradually decreased with localized spikes in the coming decades. In *Slovenski narod*, the presence of this topic gradually increased constantly throughout the entire period, with local spikes in 1874 (perhaps an echo of the Vienna stock exchange crash of 1873?) and 1885.

Finally, reporting on the topic of FOOD PRODUCTION reveals some potentially counter-intuitive insights given the assumed target audiences of the two papers, see Figure 4. While *Slovenski narod* might appear to cater to a more urbanized bourgeois-professional audience, it reported on agricultural products and related topics with double the frequency of *Slovenec*. While *Slovenec* gradually increased its reporting on FOOD PRODUCTION with a climax in 1910, *Slovenski narod*'s reporting on the topic suddenly climaxed in 1885 - perhaps in reaction to the grape phylloxera epidemic - with various localized peaks in the ensuing decades.

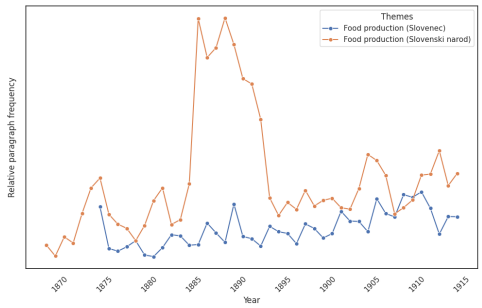

Figure 4: FOOD PRODUCTION theme relative paragraph frequency through time for both *Slovenec* and *Slovenski narod*.

6. Conclusion

This study demonstrates the potential of BERTopic for uncovering thematic and ideological patterns in historical periodicals, offering new insights into the cultural and political landscapes of selected Slovene historical periodicals. By analysing their thematic profiles and diachronic trends, we highlighted the distinct roles these periodicals played in shaping Slovene public discourse.

This research, on the one hand, confirms established historiographical knowledge on topics such as growing nationalist polarisation as well as the impact of key historical events on public discourse. At the same time, it also reveals several novel insights into differences in reporting among the periodicals. These findings underscore the value of computational methods and distant reading in historical and cultural research, particularly in enabling large-scale, data-driven analyses of complex historical texts.

While detailed interpretations of the collected data lie beyond the scope of this study, it provides a preliminary demonstration of how computational distant reading methods can identify phenomena that might otherwise remain elusive through traditional research approaches. Our coarse-grained thematic analysis revealed key reporting patterns and ideological underpinnings; however, a more fine-grained exploration of individual topics is essential to fully understand the nuances of their narrative strategies. Future research could focus on in-depth analyses of individual topics and their linguistic framing, investigating how these periodicals constructed narratives around gender, politics, religion, and other key themes to achieve a deeper understanding of the cultural and political dynamics of the past.

Acknowledgements

This work was supported by the Slovenian Research and Innovation Agency research programme "Digital Humanities: resources, tools and methods" (2022–2027) [grant number P6-0436], the support of the DARIAH-SI research infrastructure, the Slovene Common Language Resources and Technology Infrastructure, CLARIN.SI, and by the project "Large Language Models for Digital Humanities" (2024–2027) [grant number GC-0002].

7. References

Smilja Amon and Karmen Erjavec. 2011. *Slovensko časopisno izročilo: Od začetka do 1918*. Fakulteta za družbene vede, Založba FDV.

Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2024. A lightweight approach to a giga-corpus of historical periodicals: The story of a Slovenian historical newspaper collection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 695–703, Torino, Italia. ELRA and ICCL.

Michael Ginn and Mans Hulden. 2024. Historia magistra vitae: dynamic topic modeling of roman literature using neural embeddings. *arXiv preprint arXiv:2406.18907*.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Francesco Lombardi and Simone Marinai. 2020. Deep learning for historical document analysis and recognition—a survey. *Journal of Imaging*, 6(10):110.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Keerthana Murugaraj, Salima Lamsiyah, Marten During, and Martin Theobald. 2025a. Automating historical insight extraction from large-scale newspaper archives via neural topic modeling. *arXiv preprint arXiv:2512.11635*.

Keerthana Murugaraj, Salima Lamsiyah, Marten Düring, and Martin Theobald. 2025b. Mining the past: a comparative study of classical and neural topic models on historical newspaper archives. In *Proceedings of the 5th international conference on natural language processing for digital humanities*, pages 452–463.

Keerthana Murugaraj, Salima Lamsiyah, Marten During, and Martin Theobald. 2025c. Topic-RAG for historical newspapers: Enhancing information retrieval in humanities research through topic-based retrieval-augmented generation. *Computational Humanities Research*, pages 1–21.

Sarah Oberbichler, Emanuela Boroş, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, and Mikko Tolonen. 2022. Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, 73(2):225–239.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks.

Marta Villamor Martin, David A Kirsch, and Fabian Prieto-Nañez. 2023. The promise of machine-learning-driven text analysis techniques for historical research: topic modeling and word embedding. *Management & Organizational History*, 18(1):81–96.

8. Language Resource References

Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2023. Corpus of slovenian periodicals (1771-1914) sPeriodika 1.0. Slovenian language resource repository CLARIN.SI.