

# A Multilingual Sentence-Transformer Baseline for Serbian News Topic Classification

Saša Petalinkar\*, Milica Ikonić Nešić<sup>†</sup>, Ranka Stanković\*, Jelena Graovac<sup>‡</sup>

\*University of Belgrade, Faculty of Mining and Geology, Serbia

{sasa.petalinkar, ranka.stankovic}@rgf.bg.ac.rs

<sup>†</sup>University of Belgrade, Faculty of Philology, Serbia

milica.ikonic.nesic@fil.bg.ac.rs

<sup>‡</sup>University of Belgrade, Faculty of Mathematics, Serbia

jgraovac@matf.bg.ac.rs

## Abstract

We present a baseline for topic classification of Serbian newspaper articles. Starting from a corpus of roughly 123,000 newspaper articles annotated with section labels, we apply a deterministic label-normalization procedure that corrects encoding-corrupted spellings, merges truncated labels, and removes a non-coherent catch-all category, resulting in 35 normalized classes. The multilingual MiniLM sentence-transformer is fine-tuned with a multiple-negatives ranking objective and, at test time, all 35 section labels are ranked by cosine similarity, with the top-ranked label treated as a single direct classification hit. On a held-out test split of 18,487 articles, the model reaches a top-1 accuracy of 0.784 (Wilson 95% interval 0.778–0.790), with a weighted F1 of 0.790 and a macro F1 of 0.653. Per-section analysis shows strong performance on high-frequency sections and substantial over-prediction of small, semantically adjacent classes. The fine-tuned model, per-section metrics, a confusion analysis, and documented label-cleaning corrections are released to support further work on Serbian news classification.

## 1 Introduction

Serbian is a less-resourced South Slavic language for which large, cleanly labeled text-classification resources remain scarce. However, newspaper archives are routinely tagged with editorial section (*rubrika*) labels that describe the topical section to which an article belongs. These labels offer a natural supervisory signal for topical classification, which is useful for archival indexing, retrieval, and media monitoring.

In this paper, a baseline section classifier is built and evaluated over a large Serbian news corpus. The central modeling choice is to treat section assignment as a ranking problem: a multilingual sentence-transformer encodes both the article text

and each candidate section label, and sections are ranked by cosine similarity. At evaluation time, the first ranked label is taken as the single direct classification hit, which makes the model directly comparable to a standard top-1 classifier.

The main contributions of this work are:

- a normalized 35-section Serbian news dataset with a fully documented, deterministic label-cleaning procedure;
- a multilingual MiniLM sentence-transformer baseline that reaches a top-1 accuracy of 0.784 (Wilson 95% interval 0.778–0.790) and a weighted F1 of 0.790;
- per-section metrics and a confusion analysis that characterize head- and tail-section behavior and expose the dominant inter-section confusions.

## 2 Related Work

Early approaches contrasted document clustering and latent semantic indexing with probabilistic topic models such as Latent Dirichlet Allocation (Blei et al., 2003; Blei, 2012), which represent each article as a mixture over latent topics rather than assigning it to a single cluster. Newman et al. (2006) built on this lineage by coupling statistical topic models with named-entity recognition to jointly model the topics discussed and the persons, organizations, and places mentioned across 330,000 New York Times articles, extracting topics, topic trends, and entity-linking topics. More recent embedding-based methods reframe topic discovery as clustering over contextual document vectors (Grootendorst, 2022). These methods are unsupervised; in contrast, editorial section labels provide direct supervision against a fixed inventory.

Using supervised classification, Daud et al. (2023) categorize real-world Reuters news with

classical models over TF-IDF features, reporting that a hyperparameter-optimized SVM with a linear kernel outperforms stochastic gradient descent, logistic regression,  $k$ -nearest neighbors, random forests, and naive Bayes. Their survey of the area surfaces several recurring findings: support vector machines are strong text categorizers, as in the Reuters-21578 work of Debole and Sebastiani on supervised term weighting; SVM and naive Bayes have been compared for news classification on social-media streams; and frequency-based descriptors such as TF\*PDF have been used for topic extraction and tracking over news archives. The review also notes a move toward deep models for some languages (e.g. CNN/LSTM hybrids for Chinese news), feature selection over “bags of topics” rather than words, and neural topic models. Two cross-cutting observations are directly relevant here: much work labelled news “classification” in fact targets fake-news detection rather than genuine topical categorization, and most studies rely on pre-organized benchmark datasets whose clean structure does not reflect noisy real-world news.

Hu (2025) survey small-sample text classification and group existing methods into data augmentation (synonym replacement, back-translation), meta-learning (MAML, prototypical networks), and transfer learning by fine-tuning pre-trained encoders (Devlin et al., 2019), noting that augmentation can inject noise, meta-learning is unstable on text, and fine-tuning needs substantial data and is sensitive to distribution shift. They instead adopt prompt engineering, casting classification as masked language generation, and combine dynamic prompt generation with knowledge distillation from a large pre-trained teacher over unlabeled news, reporting gains in 5/10/20-shot settings and identifying class imbalance as a key difficulty.

The present work sits between these lines. As in the supervised classifiers above, articles are mapped to a fixed label set, but rather than TF-IDF features or a softmax head, a multilingual sentence-transformer (Reimers and Gurevych, 2019, 2020; Wang et al., 2020) is used and classification is cast as embedding-based ranking over section labels. Unlike the balanced, accuracy-only setup of Daud et al. (2023) and the genuinely few-shot regime of Hu (2025), the supervision here is abundant but highly skewed over 35 Serbian sections, so the central difficulty is tail-class precision. We therefore report class-balanced and distribution-weighted

metrics and compare the sentence-transformer models with conventional TF-IDF classifiers. Large-scale section classification for Serbian remains comparatively underexplored, although a few related studies address smaller samples. Živanović and Stefanović (2025) investigate topic modeling in newspaper headlines from Serbia, Bosnia and Herzegovina, and Croatia using LDA and NMF models. The study identifies dominant topics and compares thematic similarities across the three countries. Model performance is evaluated using coherence, perplexity, Hellinger distance, and Silhouette score. Batanović et al. (2018) presents the Serbian Semantic Textual Similarity News Corpus (STS.news.sr), a dataset of 1,192 Serbian sentence pairs annotated with fine-grained semantic similarity scores where several STS models are evaluated on the corpus. A WordNet-based approach for Serbian text categorization was proposed (Graovac, 2013), leveraging morphological, syntactic, and semantic information from lexical resources and representative synsets to improve classification performance on Ebart corpus, classifying on 3 classes: politics, economy and sport. Graovac (2014) uses the same corpus for classification with different,  $n$ -gram-based methods. This line of work also includes language-independent  $n$ -gram-based categorizers with weighting factors (Graovac et al., 2015), a structural SVM over  $n$ -gram features for Serbian (Kovačević and Graovac, 2015), and a hierarchical formulation (Graovac et al., 2017) of the kind this work proposes as a coarse-to-fine extension, all on small, balanced datasets with only a few classes. Đukić (2024) explore topic classification of economic newspaper articles for Serbian.

Annotated news corpora are a common way to build foundational resources for lower-resourced languages, most often for named-entity recognition. Ruokolainen et al. (2020) release a Finnish news corpus of 953 articles (193,742 tokens) with six entity classes (organization, location, person, product, event, and date) drawn from a single online outlet, with rule-based and neural baselines on in- and out-of-domain test sets. Yusufu et al. (2025) similarly present an Uzbek news corpus of 500 manually annotated articles (222,536 tokens) with three entity classes (person, location, organization) from a single source, reporting that the pre-trained CINO model outperforms other multilingual encoders. At the opposite scale, Mackenzie et al. (2020) derive CC-News-En, a 44-million-

document English news corpus from Common Crawl collected between 2016 and 2018, primarily for information-retrieval experiments. For Serbian, newspaper text features in the TESLA-NER-NEL corpus (Ikonić Nešić and Utvić, 2024), which provides more than 150,000 annotated sentences with four named-entity classes (person, organisation, location, and role) over 12,216 publicly available news articles drawn from the SrpKor reference corpus (Vitas et al., 2025). In addition, newspaper articles annotated with seven entity classes have been used to train a CNN model for named-entity recognition and linking (NER and NEL) (Ikonić Nešić et al., 2026). The resource introduced here differs from all of these along two axes: it is labeled with editorial sections (topical sections) rather than named entities or IR topics, and it is large (123,000 articles) for a single lower-resourced language, which allows topical section classification and its class-imbalance behavior to be studied directly.

### 3 Corpus Preparation and Label Normalization

The corpus is a slice of the Ebart media archive for 2023. Digitization, OCR correction, and the assignment of editorial sections (topics) were performed manually by archive staff.

The training dataset is prepared in several steps. The raw corpus index contains one row per newspaper article with the fields `unid`, `file_name`, `naslov` (title), and `rubrika` (section). Raw metadata is read with fallback decoding, and each row is linked to a plain-text article file. The dominant raw encoding is cp1250.

Label normalization is deterministic and is applied before any split generation, so that all train, validation, and test statistics are computed over the same normalized label inventory. The pass (i) standardizes known encoding-corrupted (mojibake<sup>1</sup>) spellings of canonical labels, (ii) merges truncated fragment labels into their parent section, (iii) drops a generic non-coherent catch-all label, and (iv) excludes one malformed row with an empty article identifier. Table 1 summarizes the normalized and raw corpus accounting; Tables 2–4 document the individual corrections.

The spelling correction pass standardized 10,316 rows whose canonical labels were corrupted by

<sup>1</sup>Mojibake is the corruption of readable text into garbled characters caused by interpreting encoded text with the wrong character encoding.

| Quantity                    | Value                                               |
|-----------------------------|-----------------------------------------------------|
| Raw metadata rows           | 123,196                                             |
| Raw article files on disk   | 123,635                                             |
| Linked metadata rows        | 123,196                                             |
| Metadata rows missing files | 0                                                   |
| Raw orphan article files    | 439                                                 |
| Raw topics before norm.     | 43                                                  |
| Normalized rows             | 123,179                                             |
| Dropped rows in norm.       | 17                                                  |
| Canonical spelling rewrites | 10,316                                              |
| Fragment-label merges       | 28                                                  |
| Final normalized sections   | 35                                                  |
| Raw encodings               | cp1250=122,926, utf-8=270                           |
| Raw words/file              | min 7, median 298, mean 402.35, p75 517, max 23,005 |

Table 1: Raw and normalized corpus accounting.

encoding. Table 2 lists the affected labels, where “[?]” marks a byte that was corrupted (mojibake) in the raw label.

Twenty-eight rows carried truncated fragment labels, isolated content words left over from clipped or corrupted section strings that were deterministically merged into their parent sections. The largest group comprised 11 rows of the fragment *nepogode* (“disasters”), folded into *Nesreće i elementarne nepogode* (Accidents and natural disasters) and raising that section from 4,288 to 4,299 articles. Seven rows of *finansije* (“finance”) were merged into *Ekonomija i finansije* (Economy and finance; 3,692 to 3,699), three rows of *informatika* (“IT”) into *Telekomunikacije i informatika* (Telecommunications and IT; 600 to 603), and three rows of *potroša[?]ja*, a mojibake-truncated form of *potrošača* (“consumers”) into *Zaštita prava potrošača* (Consumer rights protection; 174 to 177). The remaining merges were single- or double-row corrections: *rudarstvo* (“mining”) into *Energetika i rudarstvo* (Energy and mining; 1,563 to 1,565), *društvene* (“social”) into *Ostale društvene teme* (Other social topics; 4,911 to 4,912), and *prava* (“rights”) into *Ljudska prava* (Human rights; 440 to 441).

After that, seventeen rows were removed during normalization. Sixteen of them belonged to the generic label *teme* (“topics”), a catch-all bucket whose mixed semantics and low internal coherence made it unsuitable as a research-grade class. The remaining row was a malformed metadata entry, a blank article identifier paired with a placeholder `.txt` file, that could not be traced through pre-processing, splitting, or evaluation. Both were ex-

| Raw label                        | Normalized label               | English gloss                   | Rows  |
|----------------------------------|--------------------------------|---------------------------------|-------|
| Nesre[?]e i elementarne nepogode | Nesreće i elementarne nepogode | Accidents and natural disasters | 4,288 |
| Pravosu[?]e                      | Pravosuđe                      | Judiciary                       | 3,264 |
| Saobra[?]aj                      | Saobraćaj                      | Traffic / transport             | 1,839 |
| Ratni zlo[?]ini/Haški tribunal   | Ratni zločini/Haški tribunal   | War crimes / Hague Tribunal     | 751   |
| Zaštita prava potroša[?]a        | Zaštita prava potrošača        | Consumer rights protection      | 174   |

Table 2: Canonical spelling corrections for encoding-corrupted section labels.

| Category        | Item                               | Rows | Rationale                                                                                                   |
|-----------------|------------------------------------|------|-------------------------------------------------------------------------------------------------------------|
| Dropped section | teme                               | 16   | Generic catch-all bucket with mixed semantics and too little internal coherence for a research-grade class. |
| Malformed row   | blank article id, placeholder .txt | 1    | Could not be traced through preprocessing, splitting, or evaluation.                                        |

Table 3: Noisy or malformed items removed during normalization.

cluded before any split generation, so they affect neither the normalized label inventory nor the reported statistics.

Several small but semantically coherent sections were deliberately kept as distinct classes rather than merged for balance (Table 4).

| Section         | English gloss     | Rows  | Rationale                                                                                     |
|-----------------|-------------------|-------|-----------------------------------------------------------------------------------------------|
| Civilno društvo | Civil society     | 23    | Small but semantically coherent and distinct from the generic social-topics bucket.           |
| Korupcija       | Corruption        | 312   | Corruption coverage is substantively distinct and analytically useful.                        |
| EU              | European Union    | 422   | EU coverage is narrower than general foreign policy and remains traceable on its own.         |
| Nauka           | Science           | 427   | Science content is thematically coherent even if relatively small.                            |
| Ljudska prava   | Human rights      | 441   | Human-rights coverage is analytically meaningful as its own class.                            |
| Religija        | Religion          | 973   | Religion is a clear semantic class, not merged solely for balance.                            |
| Istorijske teme | Historical topics | 1,518 | Historical topics form a coherent genre that would be diluted by broader cultural categories. |

Table 4: Small coherent sections preserved as distinct classes after normalization.

## 4 Model and Training

The baseline uses the multilingual MiniLM sentence-transformer<sup>2</sup> (Wang et al., 2020; Reimers and Gurevych, 2019, 2020). It is fine-tuned with a multiple-negatives ranking loss over positive article-text/section pairs, relying on in-batch negatives (Henderson et al., 2017). Validation during training uses an explicit validation pair file with

<sup>2</sup>paraphrase-multilingual-MiniLM-L12-v2

a binary-classification evaluator. At test time, the model ranks all 35 section labels for each held-out article by cosine similarity between  $\ell_2$ -normalized article and section embeddings; the direct-hit classification prediction is the top-ranked section.

We compared the sentence-transformer models with two conventional classifiers trained on the same training split and evaluated on the same held-out test set. Articles were represented by unigram TF-IDF vectors with at most 100,000 features, a minimum document frequency of two, a maximum document frequency of 0.95, and sublinear term-frequency scaling. We trained a Linear SVM and multinomial logistic regression on these representations. Both conventional models perform direct single-label classification over the 35 sections, without ranking section labels.

Because the top-1 prediction of the ranking model corresponds to a single-label decision, micro precision, micro recall, micro F1, and accuracy are numerically identical: each article contributes exactly one gold section and one top-1 predicted section. Macro- and weighted-averaged precision, recall, and F1 are additionally reported.

We also report balanced accuracy, Matthews correlation coefficient (MCC), and Cohen’s  $\kappa$ . In this single-label multiclass setting, balanced accuracy equals macro recall: it assigns equal weight to the recall of each section, irrespective of its frequency. This measure complements overall accuracy, which is dominated by the most frequent sections in this highly imbalanced dataset. Equal treatment of sections is relevant to archival indexing and media monitoring because the observed frequency of a section reflects the news cycle and editorial organization rather than its analytical importance. Low-frequency sections such as Corruption, Human rights, Science, Civil society, and Consumer rights may therefore remain important despite their limited test support. We interpret balanced accuracy as a measure of class-balanced thematic coverage and consider it together with accuracy, macro

| Split | Articles | Ratio  | Pos. pairs | Neg. pairs | Total pairs | Median words | Mean words |
|-------|----------|--------|------------|------------|-------------|--------------|------------|
| train | 86,224   | 0.6999 | 86,224     | 172,448    | 258,672     | 298.0        | 402.2      |
| val   | 18,468   | 0.1499 | 18,468     | 36,936     | 55,404      | 300.0        | 405.0      |
| test  | 18,487   | 0.1501 | 18,487     | 36,974     | 55,461      | 295.0        | 400.3      |

Table 5: Article split sizes, pair counts, and article-length statistics.

F1, weighted F1, MCC, and per-class precision and recall.

For the direct-hit rate, Wilson score intervals (Wilson, 1927) are reported, which are appropriate for binomial proportions.

The normalized index is split at the article level with iterative multilabel stratification (Sechidis et al., 2011; Szymański and Kajdanowicz, 2017), using a 70/15/15 train/validation/test ratio and seed 42. The resulting artifacts are effectively single-label: the average number of sections per article is 1.0 in all three splits. Training data are prepared as article-text/section pairs, with two sampled negative labels per positive article-label pair. Table 5 reports split sizes, pair counts, and article-length statistics; Table 6 reports the section distribution across splits.

The modeling index does not contain newspaper or publication-date fields. Newspaper and date analyses are therefore joined from an auxiliary export and reported with explicit coverage. The join matched 11,782 of 123,179 normalized articles (Table 7); all matched rows with year metadata are from 2023, with 8,264 articles in train, 1,776 in validation, and 1,742 in the test set. The auxiliary join covers twelve distinct national newspapers, each present in all three splits. Because the auxiliary source was sampled at roughly one thousand articles per outlet, the per-newspaper totals are near-uniform (about 980 to 1,000 articles each), and the counts within each split simply track the 70/15/15 ratio. The five most frequent outlets in the training split are Blic, Politika, NIN, Novi Magazin and Vreme. The set mixes daily newspapers (Blic, Politika, Večernje novosti, Kurir, Alo!, Informer, Danas) with weekly news magazines (NIN, Vreme, Novi Magazin, Nedeljnik, Pečat); this editorial mix matters because section conventions differ across outlets, which plausibly contributes to some of the inter-section overlap observed in the confusion analysis. A sample of the training-data export is shown in Listing 1, which includes the article text together

with its newspaper, date, and topic labels.

Listing 1: A sample record from the training-data export.

```
<record n="9" articleId="
C1258927004CD0B8C125896C0044DDC6" split="train"
sourceWordCount="111" sourceCharCount="792"
sampleWordCount="80" sampleCharCount="557"
sampleSentenceCount="4" truncated="true"
candidateCountForNewspaper="709">
<newspaper>NIN</newspaper>
<date year="2023" month="2023-02">2023-02-16</date>
<sourceDatabase>Develop\Arhiva2023.nsf</
sourceDatabase>
<title>FILMOVI DINKA TUCAKOVI?A</title>
<topicMatch>
<goldTopic>Kultura</goldTopic>
<sourceTopic>Kultura</sourceTopic>
<canonicalTopic>Kultura</canonicalTopic>
</topicMatch>
<textPath>articles/C1258927004CD0B8C125896C0044DDC6.
txt</textPath>
<text encoding="cp1250" role="shortCorpusSample">
FILMOVI DINKA TUCAKOVI?A KINOTEKA Nakon što je
ovu nedelju i ciklus „Velikani svetskog filma"
posvetila Đini Lolobrididi, prikazavši njene
filmove Nevernice (1953), Hotel Raj (1966),
Dobro ve?e, gospo?o Kembel (1968), Zavodnice
(1965), Žena od slame (1964), Hleb, ljubav i ma
šta (1953), Carska venera (1962), Hleb, ljubav
i ljubomora (1954) – Jugoslovenska kinoteka
obeleži?e desetogodišnjicu smrti Dinka Tucakovi
?a. Tako su u nedelju i ponedeljak, 19. i 20.
februara, na programu ostvarenja The Rubber
Soul Project (2004), Šest dana juna (1985),
Zabranjeni</text>
</record>
<record n="10" articleId="
C1258927004CD0B8C125899B003CF0D0" split="train"
sourceWordCount="36" sourceCharCount="259"
sampleWordCount="36" sampleCharCount="259"
sampleSentenceCount="3" truncated="false"
candidateCountForNewspaper="706">
<newspaper>Novi Magazin</newspaper>
<date year="2023" month="2023-04">2023-04-20</date>
<sourceDatabase>Develop\Arhiva2023.nsf</
sourceDatabase>
<title>MOTOCIKLISTI – HUMANITARCI</title>
<topicMatch>
<goldTopic>Religija</goldTopic>
<sourceTopic>Religija</sourceTopic>
<canonicalTopic>Religija</canonicalTopic>
</topicMatch>
<textPath>articles/C1258927004CD0B8C125899B003CF0D0.
txt</textPath>
<text encoding="cp1250" role="shortCorpusSample">
MOTOCIKLISTI – HUMANITARCI Mališani iz Niša
takmi?ila su se u kucanju uskršnjim jajima sa
motociklistima – učesnicima humanitarne akcije
„Odaberi svog bajkera i odmeri snage". Cilj je
bio prikupljanje novca za lečenje bolesne dece.
(BETAPHOTO/SASA DJORDJEVIC)</text>
</record>
```

## 5 Results and Discussion

On the held-out test split of 18,487 articles the model produces 14,502 correct top-1 predictions and 3,985 errors, giving a top-1 accuracy (equivalently micro F1) of 0.784. Table 9 reports the full

| Section                   | English gloss                   | Full   | Train  | Val   | Test  |
|---------------------------|---------------------------------|--------|--------|-------|-------|
| Civilno društvo           | Civil society                   | 23     | 16     | 3     | 4     |
| EU                        | European Union                  | 422    | 295    | 63    | 64    |
| Ekologija                 | Environment                     | 1,108  | 776    | 166   | 166   |
| Ekonomija i finansije     | Economy and finance             | 3,699  | 2,589  | 555   | 555   |
| Energetika i rudarstvo    | Energy and mining               | 1,565  | 1,095  | 235   | 235   |
| Estrada/Zabava            | Showbiz/entertainment           | 6,980  | 4,886  | 1,047 | 1,047 |
| Infrastruktura            | Infrastructure                  | 1,252  | 876    | 188   | 188   |
| Istorijske teme           | Historical topics               | 1,518  | 1,063  | 227   | 228   |
| Korupcija                 | Corruption                      | 312    | 218    | 47    | 47    |
| Kosovo                    | Kosovo                          | 5,877  | 4,114  | 881   | 882   |
| Kriminal                  | Crime                           | 4,663  | 3,264  | 699   | 700   |
| Kultura                   | Culture                         | 14,764 | 10,335 | 2,214 | 2,215 |
| Ljudska prava             | Human rights                    | 441    | 309    | 66    | 66    |
| Mediji                    | Media                           | 1,681  | 1,177  | 252   | 252   |
| Nauka                     | Science                         | 427    | 299    | 64    | 64    |
| Nesreće i elem. nepogode  | Accidents and natural disasters | 4,299  | 3,009  | 645   | 645   |
| Obrazovanje               | Education                       | 2,743  | 1,920  | 411   | 412   |
| Ostale društvene teme     | Other social topics             | 4,912  | 3,438  | 737   | 737   |
| Policija                  | Police                          | 3,331  | 2,332  | 499   | 500   |
| Politika                  | Politics                        | 8,489  | 5,942  | 1,273 | 1,274 |
| Poljoprivreda             | Agriculture                     | 1,629  | 1,140  | 244   | 245   |
| Pravosuđe                 | Judiciary                       | 3,264  | 2,285  | 489   | 490   |
| Privreda                  | Industry and commerce           | 2,048  | 1,434  | 307   | 307   |
| Ratni zločini/Haški trib. | War crimes / Hague Tribunal     | 751    | 526    | 112   | 113   |
| Region                    | Regional (Balkans)              | 3,831  | 2,682  | 574   | 575   |
| Religija                  | Religion                        | 973    | 681    | 146   | 146   |
| Saobraćaj                 | Traffic / transport             | 1,839  | 1,287  | 276   | 276   |
| Socijalna politika        | Social policy                   | 725    | 507    | 109   | 109   |
| Spoljna politika          | Foreign policy                  | 2,830  | 1,981  | 424   | 425   |
| Sport                     | Sports                          | 16,175 | 11,322 | 2,426 | 2,427 |
| Svet                      | World                           | 15,761 | 11,033 | 2,364 | 2,364 |
| Telekom. i informatika    | Telecommunications and IT       | 603    | 422    | 90    | 91    |
| Vojska                    | Military                        | 751    | 526    | 112   | 113   |
| Zaštita prava potrošača   | Consumer rights protection      | 177    | 124    | 26    | 27    |
| Zdravstvo                 | Healthcare                      | 3,316  | 2,321  | 497   | 498   |

Table 6: Section distribution in the full normalized corpus and the train/validation/test splits, with an English gloss for each section.

| Split | Matched | Unmatched | Coverage |
|-------|---------|-----------|----------|
| train | 8,264   | 77,960    | 0.0958   |
| val   | 1,776   | 16,692    | 0.0962   |
| test  | 1,742   | 16,745    | 0.0942   |

Table 7: Coverage of auxiliary newspaper/date metadata joined by article identifier.

set of aggregate metrics. The weighted F1 (0.790) exceeds the macro F1 (0.653), reflecting strong performance on frequent sections and weaker performance on small ones.

Table 8 compares the sentence-transformer models with the conventional classifiers. The TF-IDF Linear SVM achieves the highest accuracy (0.815), macro F1 (0.713), weighted F1 (0.812), and MCC (0.801). MiniLM achieves the highest balanced accuracy (0.734), compared with 0.694 for the Linear SVM. The Linear SVM therefore performs better under metrics influenced by the observed class dis-

tribution, whereas MiniLM has the highest mean recall across the 35 sections.

Table 10 reports per-section support, predictions, true/false positives and negatives, precision, recall, F1, and a Wilson 95% interval on recall. The macro recall (0.734) exceeds the macro precision (0.618), indicating that the model tends to over-predict minority sections: for many small classes the number of predictions far exceeds the support (e.g. EU, with support 64 but 156 predictions), which inflates recall while depressing precision.

Table 11 lists the largest top-1 confusions. The dominant errors form interpretable clusters: a law-enforcement cluster in which *Kriminal* (Crime), *Policija* (Police), and *Pravosuđe* (Judiciary) are mutually confused; a cultural overlap in which *Kultura* (Culture) is drawn toward *Estrada/Zabava* (Showbiz/entertainment) and *Istorijske teme* (Historical topics); and a political adjacency in which *Politika* (Politics) is drawn toward *Kosovo*, *Spoljna politika*

| Model                      | Accuracy     | Bal. acc.    | Macro P      | Macro R      | Macro F1     | Weighted F1  | MCC          |
|----------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| TF-IDF Linear SVM          | <b>0.815</b> | 0.694        | 0.741        | 0.694        | <b>0.713</b> | <b>0.812</b> | <b>0.801</b> |
| TF-IDF logistic regression | 0.802        | 0.616        | <b>0.751</b> | 0.616        | 0.657        | 0.797        | 0.786        |
| MiniLM                     | 0.784        | <b>0.734</b> | 0.618        | <b>0.734</b> | 0.653        | 0.790        | 0.770        |
| GTE-512                    | 0.722        | 0.639        | 0.534        | 0.639        | 0.555        | 0.736        | 0.705        |

Table 8: Aggregate single-label classification results on the common held-out test set. Balanced accuracy equals macro recall in this setting. Bold indicates the highest value in each column.

| Metric                        | Value       |
|-------------------------------|-------------|
| Evaluated articles            | 18,487      |
| Correct top-1 predictions     | 14,502      |
| Top-1 errors                  | 3,985       |
| Micro precision / recall / F1 | 0.784       |
| Wilson 95% interval           | 0.778–0.790 |
| Macro precision               | 0.618       |
| Macro recall                  | 0.734       |
| Macro F1                      | 0.653       |
| Weighted precision            | 0.805       |
| Weighted recall               | 0.784       |
| Weighted F1                   | 0.790       |
| Ranking evaluator Hit@1       | 0.784       |

Table 9: Held-out test results when the ranking model is evaluated as a direct top-1 section classifier.

(Foreign policy), and *Mediji* (Media). Several confusions reflect genuine semantic proximity, such as the mutual confusion between *Saobraćaj* (Traffic/transport) and *Infrastruktura* (Infrastructure), and the tendency for *Spoljna politika* (Foreign policy) to be predicted as *EU* (European Union).

The Linear SVM’s overall accuracy advantage is concentrated in several frequent sections. Differences in *Culture*, *World*, *Politics*, and *Crime* account for approximately 97% of its net accuracy advantage over MiniLM. By contrast, MiniLM has higher recall in 19 of the 35 sections. Its largest recall advantages occur for *Corruption* (+0.298), *EU* (+0.281), *Infrastructure* (+0.202), *Human rights* (+0.121), *Consumer rights protection* (+0.111), and *Science* (+0.109). This pattern is consistent with MiniLM’s higher balanced accuracy.

Higher minority-class recall is generally accompanied by lower precision. Relative to the Linear SVM, MiniLM increases recall for *Corruption* by 0.298 but decreases precision by 0.318. For *EU*, recall increases by 0.281 and precision decreases by 0.309. The Linear SVM has higher per-class F1 in 31 of the 35 sections; MiniLM has higher F1 for *Infrastructure*, *Region*, *Judiciary*, and *Economy and finance*. Thus, the Linear SVM has stronger overall and class-level F1 performance, while MiniLM has higher mean class recall.

The baseline is strong on high-frequency sections: *Sport* (Sports; F1 0.965), *Svet* (World; 0.905), and *Kultura* (Culture; 0.865) are all reliably separated. The weakest classes are small and semantically adjacent to larger ones; for example, *Civilno društvo* (Civil society; support 4, precision 0.094), *Zaštita prava potrošača* (Consumer rights protection; precision 0.228), and *EU* (European Union; precision 0.282) are heavily over-predicted. The gap between macro recall and macro precision quantifies this tendency: the ranking objective favors recall on rare labels at the cost of precision, since small sections receive far more top-1 predictions than their true support.

Two regularities make this concrete. First, top-1 quality is strongly governed by support: every section with more than 2,000 test instances scores  $F1 \geq 0.865$  (Sport, Svet, Kultura), whereas nine of the ten smallest sections (support below 120) fall below 0.60. Second, the over-prediction is systematic rather than incidental: for almost every minority section the model assigns the label far more often than it actually occurs. *Civilno društvo* (Civil society) is predicted eight times as often as its true support, *Zaštita prava potrošača* (Consumer rights protection) nearly three times, and EU about 2.4 times, whereas the head sections are predicted slightly less often than their support (Sport and Kultura at 0.97 and 0.88 of their support, respectively). The ranking objective therefore redistributes decision mass toward rare labels, which is precisely what pushes macro recall (0.734) above macro precision (0.618) and opens the 0.137-point gap between weighted F1 (0.790) and macro F1 (0.653).

The confusion structure is largely interpretable and points to editorial overlap rather than arbitrary error. Law-enforcement sections, cultural sub-genres, and political topics are the principal sources of confusion, all of which are areas where newspaper section boundaries are themselves fuzzy.

Two practical consequences follow from this structure. First, the largest confusions are bidi-

| Section                            | English gloss                   | Support | Pred. | TP    | FP  | FN  | Prec. | Rec.  | F1    | Rec. 95% CI |
|------------------------------------|---------------------------------|---------|-------|-------|-----|-----|-------|-------|-------|-------------|
| Sport                              | Sports                          | 2,427   | 2,358 | 2,309 | 49  | 118 | 0.979 | 0.951 | 0.965 | 0.942–0.959 |
| Svet                               | World                           | 2,364   | 2,181 | 2,056 | 125 | 308 | 0.943 | 0.870 | 0.905 | 0.856–0.883 |
| Kultura                            | Culture                         | 2,215   | 1,954 | 1,804 | 150 | 411 | 0.923 | 0.814 | 0.865 | 0.798–0.830 |
| Politika                           | Politics                        | 1,274   | 1,049 | 893   | 156 | 381 | 0.851 | 0.701 | 0.769 | 0.675–0.725 |
| Estrada/Zabava                     | Showbiz/entertainment           | 1,047   | 1,115 | 944   | 171 | 103 | 0.847 | 0.902 | 0.873 | 0.882–0.918 |
| Kosovo                             | Kosovo                          | 882     | 859   | 736   | 123 | 146 | 0.857 | 0.834 | 0.845 | 0.808–0.858 |
| Ostale društvene teme              | Other social topics             | 737     | 574   | 331   | 243 | 406 | 0.577 | 0.449 | 0.505 | 0.414–0.485 |
| Kriminal                           | Crime                           | 700     | 607   | 410   | 197 | 290 | 0.675 | 0.586 | 0.627 | 0.549–0.622 |
| Nesreće i elem. nepogode           | Accidents and natural disasters | 645     | 641   | 514   | 127 | 131 | 0.802 | 0.797 | 0.799 | 0.764–0.826 |
| Region                             | Regional (Balkans)              | 575     | 634   | 487   | 147 | 88  | 0.768 | 0.847 | 0.806 | 0.815–0.874 |
| Ekonomija i finansije              | Economy and finance             | 555     | 446   | 360   | 86  | 195 | 0.807 | 0.649 | 0.719 | 0.608–0.687 |
| Policija                           | Police                          | 500     | 490   | 301   | 189 | 199 | 0.614 | 0.602 | 0.608 | 0.558–0.644 |
| Zdravstvo                          | Healthcare                      | 498     | 504   | 413   | 91  | 85  | 0.819 | 0.829 | 0.824 | 0.794–0.860 |
| Pravosuđe                          | Judiciary                       | 490     | 491   | 367   | 124 | 123 | 0.747 | 0.749 | 0.748 | 0.709–0.785 |
| Spoljna politika                   | Foreign policy                  | 425     | 413   | 263   | 150 | 162 | 0.637 | 0.619 | 0.628 | 0.572–0.664 |
| Obrazovanje                        | Education                       | 412     | 475   | 338   | 137 | 74  | 0.712 | 0.820 | 0.762 | 0.780–0.854 |
| Privreda                           | Industry and commerce           | 307     | 335   | 196   | 139 | 111 | 0.585 | 0.638 | 0.611 | 0.583–0.690 |
| Saobraćaj                          | Traffic / transport             | 276     | 304   | 194   | 110 | 82  | 0.638 | 0.703 | 0.669 | 0.646–0.754 |
| Mediji                             | Media                           | 252     | 305   | 183   | 122 | 69  | 0.600 | 0.726 | 0.657 | 0.668–0.778 |
| Poljoprivreda                      | Agriculture                     | 245     | 246   | 189   | 57  | 56  | 0.768 | 0.771 | 0.770 | 0.715–0.820 |
| Energetika i rudarstvo             | Energy and mining               | 235     | 269   | 194   | 75  | 41  | 0.721 | 0.826 | 0.770 | 0.772–0.869 |
| Istorijske teme                    | Historical topics               | 228     | 302   | 163   | 139 | 65  | 0.540 | 0.715 | 0.615 | 0.653–0.770 |
| Infrastruktura                     | Infrastructure                  | 188     | 286   | 135   | 151 | 53  | 0.472 | 0.718 | 0.570 | 0.650–0.778 |
| Ekologija                          | Environment                     | 166     | 210   | 123   | 87  | 43  | 0.586 | 0.741 | 0.654 | 0.669–0.802 |
| Religija                           | Religion                        | 146     | 203   | 113   | 90  | 33  | 0.557 | 0.774 | 0.648 | 0.700–0.834 |
| Ratni zločini/Haški trib. Tribunal | War crimes / Hague Tribunal     | 113     | 151   | 89    | 62  | 24  | 0.589 | 0.788 | 0.674 | 0.703–0.853 |
| Vojska                             | Military                        | 113     | 169   | 84    | 85  | 29  | 0.497 | 0.743 | 0.596 | 0.656–0.815 |
| Socijalna politika                 | Social policy                   | 109     | 169   | 56    | 113 | 53  | 0.331 | 0.514 | 0.403 | 0.421–0.606 |
| Telekom. i informatika             | Telecommunications and IT       | 91      | 135   | 63    | 72  | 28  | 0.467 | 0.692 | 0.558 | 0.591–0.778 |
| Ljudska prava                      | Human rights                    | 66      | 120   | 40    | 80  | 26  | 0.333 | 0.606 | 0.430 | 0.485–0.715 |
| EU                                 | European Union                  | 64      | 156   | 44    | 112 | 20  | 0.282 | 0.688 | 0.400 | 0.566–0.788 |
| Nauka                              | Science                         | 64      | 127   | 51    | 76  | 13  | 0.402 | 0.797 | 0.534 | 0.683–0.877 |
| Korupcija                          | Corruption                      | 47      | 98    | 38    | 60  | 9   | 0.388 | 0.809 | 0.524 | 0.675–0.896 |
| Zaštita prava potrošača            | Consumer rights protection      | 27      | 79    | 18    | 61  | 9   | 0.228 | 0.667 | 0.340 | 0.478–0.814 |
| Civilno društvo                    | Civil society                   | 4       | 32    | 3     | 29  | 1   | 0.094 | 0.750 | 0.167 | 0.301–0.954 |

Table 10: Per-section direct-hit classification metrics on the held-out test split.

| Gold section                                     | Predicted section                                | Count | Gold support | Rate within gold |
|--------------------------------------------------|--------------------------------------------------|-------|--------------|------------------|
| Kriminal (Crime)                                 | Policija (Police)                                | 100   | 700          | 0.1429           |
| Kultura (Culture)                                | Estrada/Zabava (Showbiz/entertainment)           | 96    | 2,215        | 0.0433           |
| Policija (Police)                                | Kriminal (Crime)                                 | 82    | 500          | 0.1640           |
| Kriminal (Crime)                                 | Pravosuđe (Judiciary)                            | 57    | 700          | 0.0814           |
| Kultura (Culture)                                | Istorijske teme (Historical topics)              | 56    | 2,215        | 0.0253           |
| Politika (Politics)                              | Kosovo                                           | 47    | 1,274        | 0.0369           |
| Ekonomija i finansije (Economy & finance)        | Privreda (Industry & commerce)                   | 44    | 555          | 0.0793           |
| Kultura (Culture)                                | Ostale društvene teme (Other social topics)      | 44    | 2,215        | 0.0199           |
| Ostale društvene teme (Other social topics)      | Socijalna politika (Social policy)               | 43    | 737          | 0.0583           |
| Politika (Politics)                              | Spoljna politika (Foreign policy)                | 43    | 1,274        | 0.0338           |
| Estrada/Zabava (Showbiz/entertainment)           | Kultura (Culture)                                | 34    | 1,047        | 0.0325           |
| Politika (Politics)                              | Mediji (Media)                                   | 34    | 1,274        | 0.0267           |
| Kosovo                                           | Spoljna politika (Foreign policy)                | 33    | 882          | 0.0374           |
| Svet (World)                                     | Regional (Balkans)                               | 31    | 2,364        | 0.0131           |
| Ostale društvene teme (Other social topics)      | Infrastruktura (Infrastructure)                  | 30    | 737          | 0.0407           |
| Spoljna politika (Foreign policy)                | EU (European Union)                              | 30    | 425          | 0.0706           |
| Saobraćaj (Transport)                            | Infrastruktura (Infrastructure)                  | 29    | 276          | 0.1051           |
| Kriminal (Crime)                                 | Nesreće i elem. nepogode (Accidents & disasters) | 28    | 700          | 0.0400           |
| Politika (Politics)                              | Ostale društvene teme (Other social topics)      | 28    | 1,274        | 0.0220           |
| Pravosuđe (Judiciary)                            | Kriminal (Crime)                                 | 28    | 490          | 0.0571           |
| Ekonomija i finansije (Economy & finance)        | Energetika i rudarstvo (Energy & mining)         | 25    | 555          | 0.0450           |
| Nesreće i elem. nepogode (Accidents & disasters) | Kriminal (Crime)                                 | 25    | 645          | 0.0388           |
| Ostale društvene teme (Other social topics)      | Privreda (Industry & commerce)                   | 25    | 737          | 0.0339           |
| Spoljna politika (Foreign policy)                | Kosovo                                           | 25    | 425          | 0.0588           |
| Svet (World)                                     | EU (European Union)                              | 24    | 2,364        | 0.0102           |
| Policija (Police)                                | Pravosuđe (Judiciary)                            | 23    | 500          | 0.0460           |
| Svet (World)                                     | Spoljna politika (Foreign policy)                | 23    | 2,364        | 0.0097           |
| Infrastruktura (Infrastructure)                  | Saobraćaj (Transport)                            | 22    | 188          | 0.1170           |
| Kultura (Culture)                                | Mediji (Media)                                   | 22    | 2,215        | 0.0099           |
| Spoljna politika (Foreign policy)                | Politika (Politics)                              | 22    | 425          | 0.0518           |

Table 11: Largest top-1 confusions in the held-out test split.

rectional and occur between editorially adjacent sections: Kriminal (Crime) and Policija (Police) are confused in both directions, as are Saobraćaj (Traffic/transport) and Infrastruktura (Infrastructure), and Kultura (Culture) and Estrada/Zabava (Showbiz/entertainment). The embedding space therefore already places semantically related sections close together, and the model fails mainly where human section boundaries are themselves porous; the residual error is dominated by genuine label ambiguity rather than representational weakness. Second, because the head sections are already precision-strong, per-class decision thresholds or calibration targeted at the over-predicted tail (EU, Civilno društvo (Civil society), Zaštita prava potrošača (Consumer rights protection), and Socijalna politika (Social policy)) should raise macro precision and macro F1 at little cost to the head. The consistently symmetric confusions are, in turn, natural candidates for a coarse-to-fine label hierarchy or selective merging that would cut avoidable inter-section leakage while preserving the analytically useful small classes.

The final training-time validation row corresponds to epoch 5.0, with pairwise validation accuracy 0.95 and pairwise validation F1 0.92; the reported average precision and Matthews correlation were 0.0. These are pairwise validation diagnostics on the binary-classification evaluator, not the final held-out top-1 classification metrics, and should not be interpreted as the model’s classification performance.

## 6 Conclusion

A documented label-normalization pipeline and a multilingual sentence-transformer baseline for section classification of Serbian newspaper articles are introduced, addressing a task that has so far been largely unaddressed at this scale for the language. By casting section assignment as embedding-based ranking and evaluating the top-ranked label as a direct classification decision, the model reaches a top-1 accuracy of 0.784 (Wilson 95% interval 0.778–0.790) and a weighted F1 of 0.790 over 35 sections and 18,487 held-out articles, with reliable separation of the high-frequency sections (F1 up to 0.965).

The comparison with conventional classifiers shows that unigram TF-IDF with a Linear SVM is a strong baseline for this task. It outperforms MiniLM in accuracy, macro F1, weighted F1, and

MCC. MiniLM does not outperform the Linear SVM overall, but it achieves higher balanced accuracy and higher recall in 19 of the 35 sections, including several low-frequency sections. The two models therefore differ in their error distribution: the Linear SVM yields better overall and per-class F1 performance, whereas MiniLM yields higher mean recall across sections.

Beyond these headline numbers, the analysis isolates a single, actionable failure mode: performance is governed almost entirely by class support, and the ranking objective systematically overpredicts the long tail, which is what separates the weighted F1 (0.790) from the macro F1 (0.653). Crucially, the residual errors are not random; rather, they are concentrated in bidirectional confusions between editorially adjacent sections. This pattern suggests that the embedding space already captures the underlying topical structure, and that the remaining scope for improvement is constrained primarily by genuine label ambiguity rather than by limitations in the learned representations. This diagnosis directly motivates the most promising next steps: per-class thresholding or calibration on the over-predicted tail, class re-balancing, and a coarse-to-fine or explicitly multilabel formulation for articles that legitimately span several sections. The fine-tuned model, together with the per-section metrics and the documented cleaning corrections, will be released with the final version of this paper, the underlying newspaper corpus itself cannot be made publicly available under the terms of the source archive.

## Acknowledgment

This research was supported by the Science Fund of the Republic of Serbia, #7276, *Text Embeddings - Serbian Language Applications - TESLA*. The authors gratefully acknowledge Media Archive Ebart for providing access to valuable resources that were essential to the completion of this research.

## References

- Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2018. Fine-grained semantic textual similarity for Serbian. In *Proc. of 11th Conference on Language Resources and Evaluation (LREC 2018)*, pages 1370–1378.
- David M. Blei. 2012. *Probabilistic Topic Models*. *Communications of the ACM*, 55(4):77–84.

- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Shahzada Daud, Muti Ullah, Amjad Rehman, Tanzila Saba, Robertas Damaševičius, and Abdul Sattar. 2023. [Topic Classification of Online News Articles Using Optimized Machine Learning Models](#). *Computers*, 12(1):16.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *2019 Conf. of the North American Chapter of ACL: HLT (NAACL-HLT)*, pages 4171–4186.
- Jelena Graovac. 2013. [Wordnet-Based Serbian Text Categorization](#). *Infotheca - Journal for Digital Humanities*, 14(2):2a–17a.
- Jelena Graovac. 2014. Text categorization using n-gram based language independent technique. *Intelligent Data Analysis*, 18(4):677–695.
- Jelena Graovac, Jovana Kovačević, and Gordana Pavlović-Lažetić. 2015. Language Independent n-Gram-Based Text Categorization with Weighting Factors: A Case Study. *Journal of Information and Data Management (JIDM)*, 6(1):4–17.
- Jelena Graovac, Jovana Kovačević, and Gordana Pavlović-Lažetić. 2017. Hierarchical vs. flat n-gram-based text categorization: can we do better? *Computer Science and Information Systems (ComSIS)*, 14(1):103–121.
- Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv preprint arXiv:2203.05794*.
- Matthew Henderson, Rami Al-Rfou, and Brian Strope et al. 2017. Efficient Natural Language Response Suggestion for Smart Reply. *arXiv:1705.00652*.
- Liao Hu. 2025. [Topic Classification of Small Sample News Based on Prompt Engineering](#). In *Proceedings of the 7th International Conference on Computing and Data Science*, Applied and Computational Engineering, pages 101–107.
- Milica Ikonić Nešić, Saša Petalinkar, Olivera Kitanović, Ranka Stanković, and Miloš Utvić. 2026. CNN-based named entity linking: Serbian use case. *Poznan Studies in Contemporary Linguistics*, 62(2):249–280.
- Milica Ikonić Nešić and Miloš Utvić. 2024. [Overview of the Tesla-Ner-Nel-Gold Dataset: Showcase on Serbian-English Parallel Corpus](#). In *South Slavic Languages in the Digital Environment JuDig: Book of Abstracts*, page 57.
- Jovana Kovačević and Jelena Graovac. 2015. Application of a Structural Support Vector Machine method to N-gram based text classification in Serbian. *INFOtheca - Journal of Information and Library Science*, 16(1).
- Joel Mackenzie, Rodger Benham, and Matthias Petri et al. 2020. [CC-News-En: A Large English News Corpus](#). In *Proc.s of 29th ACM International Conference on Information & Knowledge Management, CIKM '20*, page 3077–3084. ACM.
- David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. [Analyzing Entities and Topics in News Articles Using Statistical Topic Models](#). In *Intelligence and Security Informatics (ISI 2006)*, volume 3975 of *LNCS*, pages 93–104. Springer.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](#). In *Proc. of 2019 Conference on Empirical Methods in NLP and the 9th International Joint Conference on NLP (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. ACL.
- Nils Reimers and Iryna Gurevych. 2020. [Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation](#). In *Proc. of 2020 Conference on Empirical Methods in NLP (EMNLP)*, pages 4512–4525. ACL.
- Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2020. A Finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 54(1):247–272.
- Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. [On the Stratification of Multi-Label Data](#). In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2011)*, volume 6913 of *LNCS*, pages 145–158. Springer.
- Piotr Szymański and Tomasz Kajdanowicz. 2017. A Network Perspective on Stratification of Multi-Label Data. In *Proc. of 1st International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA)*, volume 74, pages 22–35.
- Mirko Đukić. 2024. Topic classification of economic newspaper articles in a highly inflectional language—the case of Serbia. Technical report, Working Papers Bulletin of the National Bank of Serbia, VI.
- Duško Vitas, Ranka Stanković, and Cvetana Krstev. 2025. The Many Faces of SrpKor [O familiji korpusa savremenog srpskog jezika SrpKor]. In *South Slavic Languages in the Digital Environment JuDig Book of Abstracts*, pages 3–27.
- Wenhui Wang, Furu Wei, and Li Dong et al. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Also in Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.
- Edwin B. Wilson. 1927. [Probable Inference, the Law of Succession, and Statistical Inference](#). *Journal of the American Statistical Association*, 22(158):209–212.

Aizihaijiaerjiang Yusufu, Kamran Aziz, Aizierguli Yusufu, Abidan Ainiwaer, Fei Li, and Donghong Ji. 2025. Uzbek news corpus for named entity recognition: A. Yusufu et al. *Language resources and evaluation*, 59(3):3139–3152.

Marko M Živanović and Nenad Stefanović. 2025. Processing Balkan News Articles Using Selected Methods of Natural Language Processing. In *24th Int. Symposium INFOTEH-JAHORINA*, pages 11–16.