Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Named Entity Recognition for Pre-Modern Serbian Cyrillic: Using Transformer-Based Models and Generative Data Augmentation

Marija Đokić Petrović

mdjokicpetrovic@raf.rs

Union University

Vladimir Polomac

University of Kragujevac

Mihailo Popović

Austrian Academy of Sciences

Research Article

Keywords: Named Entity Recognition, Pre-modern Serbian Cyrillic, Large Language Models, Generative Data Augmentation, Digital Humanities, Cultural Heritage

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10275617/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Named Entity Recognition for Pre-Modern Serbian Cyrillic: Using Transformer-Based Models and Generative Data Augmentation

Marija Đokić Petrović[1*], Vladimir Polomac[2] and Mihailo St. Popović[3]

[1*]Department of Computer Science, The School of Computing, Union University, Kneza Mihaila VI/6, Belgrade, 11000, Serbia.
[2]Faculty of Philology and Arts, University of Kragujevac, Liceja Kneževine Srbije 1A, Kragujevac, 34000, Serbia.
[3]Institute for Medieval Research, Department of Byzantine Studies, Austrian Academy of Sciences, Georg-Coch-Platz 2, Vienna, 1010, Austria.

*Corresponding author(s). E-mail(s): mdjokicpetrovic@raf.rs;
Contributing authors: v.polomac@filum.kg.ac.rs;
mihailo.popovic@oeaw.ac.at;

Abstract

In this paper, we evaluate the effectiveness of transformer-based models for Named Entity Recognition (NER) in pre-modern Serbian Cyrillic texts, with a particular focus on improving model generalization through generative data augmentation. We construct a domain-specific NER dataset, based on medieval Serbian charters and archival documents, annotated using the BIO tagging scheme, and extend the training corpus with synthetically generated examples. Two transformer-based language models, BERTić and XLM-RoBERTa, are fine-tuned and evaluated in both in-domain and out-of-domain settings. The results show that transformer-based models, particularly fine-tuned XLM-RoBERTa, achieve strong in-domain performance and substantially improve out-of-domain robustness compared with a previously developed CNN-based NER model. Old-Serb-XLM-RoBERTa trained with 30% synthetic data achieves the highest in-domain performance, with an overall F1-score of ≈ 0.92. The configuration trained with 100% synthetic data performs best on some individual out-of-domain documents, whereas the 50% configuration achieves the highest average weighted F1-score in out-of-domain evaluation and is therefore selected as the

recommended model for broader use. These findings show that synthetic data augmentation can improve generalization, especially for low-resource historical NER, but its effectiveness depends on the model architecture, augmentation ratio, and the degree of domain shift. The recommended model, together with the annotated and augmented training dataset, is publicly available via Hugging Face, supporting further research on historical South Slavic languages.

Keywords: Named Entity Recognition, Pre-modern Serbian Cyrillic; Large Language Models, Generative Data Augmentation, Digital Humanities, Cultural Heritage

1 Introduction

Named Entity Recognition (NER) is an important task in Natural Language Processing (NLP), enabling the extraction of entities such as persons, locations, socio-cultural groups, and others from unstructured text (Munnangi, 2024). While NER has achieved strong performance for contemporary languages and domains, its application to low-resource languages and their historical varieties remains challenging (Keraghel, Morbieu, & Nadif, 2024). These challenges are particularly pronounced when annotated corpora are scarce, and linguistic variation limits the effectiveness of standard approaches (Zhong et al., 2026).

As a low-resource language, Serbian lacks large-scale annotated datasets and robust NER tools, especially for historical texts (Marovac, Avdić, & Milošević, 2023). The work (Đokić Petrović, Popović, & Polomac, 2025) has demonstrated that NER can be applied to pre-modern Serbian Cyrillic using traditional approaches, such as Convolutional Neural Networks (CNN), but it relies on limited training data and shows reduced performance when applied to unseen documents.

The emergence of Large Language Models (LLMs) have significantly improved NER performance for modern Serbian (Torge, Politov, Lehmann, Saffar, & Tao, 2023). For example, Ljubešić and Lauc (2021) developed the BERTić model, fine-tuned on multiple datasets, which achieved strong performance on NER tasks for contemporary Serbian, among others. Ikonić Nešić, Petalinkar, Škorić, and Stanković (2024) explored the integration of BERT-based models with the spaCy framework and compared various transformer architectures and training strategies for NER in a diverse Serbian-language dataset. Kalušev and Brkljač (2026) presented an LLM-based NER model for legal documents written in modern Serbian. However, the application of transformer-based models to pre-modern Serbian remains largely unexplored.

In this paper, we investigate the effectiveness of transformer-based models for NER in pre-modern Serbian Cyrillic texts. We build on an existing corpus of medieval charters and archival documents developed in (Đokić Petrović et al., 2025) and convert the annotations into the BIO tagging scheme. We fine-tune two transformer-based models, BERTić and XLM-RoBERTa, and evaluate their performance in both in-domain and out-of-domain settings. The out-of-domain evaluation is conducted on previously unseen historical documents that differ in their chronological, genre-related, and orthographic proximity to the training corpus. This evaluation setup enables us to assess the robustness of the models under different degrees of domain shift.

To address the limited size of the original corpus, we apply a generative data augmentation strategy and produce additional synthetic training examples at augmentation ratios of 30%, 50%, and 100%. This experimental setup enables us to systematically investigate the impact of synthetic augmentation on model performance and generalization. In particular, it allows us to examine whether the effectiveness of synthetic data varies depending on the amount of generated material and the degree of similarity between the training corpus and the evaluation document.

Our results show that transformer-based models, especially Old-Serb-XLM-RoBERTa, achieve strong in-domain performance and improve out-of-domain robustness compared with the CNN-based NER model developed in (Đokić Petrović et al., 2025). Old-Serb-XLM-RoBERTa trained with 30% synthetic data achieves the highest in-domain performance, with an overall F1-score of ≈ 0.92. The configuration trained with 100% synthetic data performs best on some individual out-of-domain documents, whereas the 50% configuration achieves the highest average weighted F1-score in out-of-domain evaluation and is therefore selected as the recommended model for broader use. These findings show that synthetic data augmentation can improve generalization, especially for low-resource historical NER, but its effectiveness depends on the model architecture, augmentation ratio, and the degree of domain shift. The recommended model, together with the annotated and augmented datasets, is publicly released to support further research on historical Serbian and other South Slavic language varieties.

The main contributions of this paper are as follows:

- BIO-annotated (non)augmented NER dataset(s) for pre-modern Serbian Cyrillic texts based on the Serbian medieval charters and archival material publicly available via Hugging Face[1]
- A systematic study of the effectiveness of a generative data augmentation technique for the Serbian NER in a historical domain
- A comparative evaluation of transformer-based models (BERTić and XLM-RoBERTa) and a CNN-based model for the Serbian NER in a historical domain
- A publicly available NER model (Old-Serb-XLM-RoBERTa) for pre-modern Serbian Cyrillic released via Hugging Face[2].

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the original corpus and the BIO-annotation methodology. Section 4 presents the fine-tuning of the models and describes synthetic data generation. Section 5 presents the results of the models' evaluation. Section 6 discusses the evaluation results. Finally, Section 7 concludes the paper and proposes directions for future work.

2 Related Work

The application of NER to Serbian became more prominent in the early 2010s, with the emergence of more systematic and domain-specific research efforts (Marovac et al., 2023). The first approaches relied primarily on rule-based and hybrid methods. For

[1] https://huggingface.co/datasets/marijadjokic/pre-modern-serbian-cyrillic
[2] https://huggingface.co/marijadjokic/Old-Serb-XLM-RoBERTa

instance, Vitas and Pavlović-Lažetić (2008) developed a system combining morphological and lexical analysis with dictionary-based resources to identify personal names and geographical entities. Similarly, Krstev, Obradović, Utvić, and Vitas (2014) extended earlier rule-based frameworks (Krstev, 1997) by incorporating transducers and lexical resources such as thesauri, enabling the recognition of personal and geopolitical entities.

The introduction of machine learning approaches marked a significant step forward. Šandrih, Krstev, and Stankovic (2019) developed Stanford- and spaCy-based NER models using a dataset of news articles annotated with personal names, produced by the system introduced in (Krstev et al., 2014). Ljubešić, Stupar, Jurić, and Agić (2013) developed a Conditional Random Field (CRF)-based NER system for Croatian and Slovenian, closely related to the Serbian language, trained on annotated web and news corpora such as SETimes.HR (Agić & Ljubešić, 2014) and Vjesnik (Glavaš & Šnajder, 2015). Šandrih Todorović, Krstev, Stanković, and Ikonić Nešić (2021) introduced SrpCNNER, a CNN-based model trained on the SrpELTeC-gold corpus (Stanković, Krstev, Šandrih Todorović, & Škorić, 2022), which consists of Serbian novels from 1840 to 1920. This model supports multiple entity types and achieves strong performance, with an F1 score of $\approx 91\%$ on the test set. Building on this work, Nešić, Petalinkar, Stanković, Utvić, and Kitanović (2024) proposed SrpCNNeL, a system for NER and entity linking that extends the pre-trained SrpCNNER2 model, which follows the same architecture as SrpCNNER. It is implemented in spaCy, but is trained on a more diverse dataset that includes the SrpELTeC-gold-extended dataset, newspaper articles, and automatically generated sentences derived from Wikidata and the Leximirka (Lazić & Škorić, 2020) database.

The introduction of transformer-based models has further advanced NER for Serbian and related South Slavic languages. In particular, the BERT family of transformer architectures (Devlin, Chang, Lee, & Toutanova, 2019) has enabled substantial improvements in that domain. For example, XLM-RoBERTa (Conneau et al., 2020), trained on large-scale multilingual corpora covering 100 languages, supports NER for modern Serbian. Ljubešić and Lauc (2021) proposed BERTić, fine-tuned on multiple datasets (SETimes.SR (Batanović, Ljubešić, Samardžić, & Erjavec, 2018; Batanović, Ljubešić, & Samardžić, 2018), ReLDI-sr (Ljubešić, Erjavec, Batanović, Miličević, & Samardžić, 2019), and news corpora), which achieve strong performance on NER tasks for contemporary Bosnian, Croatian, Montenegrin, and Serbian (BCMS). Ikonić Nešić et al. (2024) explored the integration of BERT-based models with the spaCy framework and compared different transformer architectures and training strategies for NER on a diverse Serbian-language dataset containing seven entity types. Furthermore, Kalušev and Brkljač (2026) introduce a specialized LLM-based NLP model that automatically extracts and classifies entities in Serbian legal texts, leveraging pre-trained BERT.

Despite recent advances in transformer-based architectures for NER in modern Serbian (Ikonić Nešić, Petalinkar, Stanković, & Mitkov, 2025), their application to historical varieties of the language remains largely underexplored. One of the few studies addressing this domain is (Đokić Petrović et al., 2025), which introduced a CNN-based NER model for pre-modern Serbian Cyrillic texts. Trained on medieval charters and legal documents, the model demonstrated the feasibility of NER in historical corpora,

achieving an F1 score of ≈ 0.92 on the test set. However, its performance declined on unseen data, indicating limited generalization across different historical sources. Motivated by these limitations, we investigate the effectiveness of transformer-based architectures for NER in pre-modern Serbian.

One of the key challenges for low-resource languages is the scarcity of resources and the limited availability of annotated data (Hiltmann et al., 2025). In the context of historical language varieties, it is especially relevant, as annotated resources are limited, and linguistic variation, non-standard orthography, and domain-specific entity types make manual annotation both challenging and costly (Gladstone, Fang, & Stewart, 2025). This limitation, also highlighted in (Đokić Petrović et al., 2025), motivated us to investigate the effectiveness of data augmentation in enriching the training dataset and improving model performance in pre-modern Serbian.

Data augmentation has become an important strategy for addressing data scarcity in NER (Huang, Gao, & Ren, 2025). Early data augmentation approaches were largely rule-driven and word-level modifications. For example, Dai and Adel (2020) investigated synonym and mention replacement, whereas Elwing Torres, Silva de Moura, Soares da Silva, Nascimento, and Mesquita (2026) applied comparable methods, highlighting the potential loss of semantic coherence introduced by heuristic rules.

Addressing the limitations of rule-based methods, generative neural approaches have been proposed to synthesize fully labeled sentences. Ding et al. (2020) employ an Long Short-Term Memory (LSTM)-based language model trained on linearized sequences, while (Zhou et al., 2022) applied Masked Language Modeling (MLM) to generate diverse entities within contextual constraints. These methods, however, often depend on the sequential dependency of entities. To mitigate this limitation, Hu et al. (2023) introduced an entity-to-text framework that decouples entity dependencies.

More recently, research has shifted to prompting LLMs to enhance diversity and robustness. Song, Shen, and Zhao (2024) introduce a prompt-based method that employs continuous soft prompts on pre-trained language models. This method introduces operations such as label flipping and context augmentation, combined with self-consistency filtering to remove low-quality samples and a mix-up strategy to reduce overfitting to adversarial examples. Building on the rewriting capabilities of LLMs, authors in Ye et al. (2024) perform augmentation at both the contextual and entity levels. They use multiple rewriting strategies, such as varying sentence length and style, to preserve semantic coherence while increasing diversity. To ensure that the generated augmentations maintain linguistic consistency with the target domain, Evuru et al. (2024) extracted certain lexical, syntactic, and semantic constraints from the source data and created natural language instructions to guide the LLM.

Despite significant advancements, research on generative data augmentation for low-resource languages remains limited. For example, Moreira, Silva, Bencke, and Moreira (2026) evaluate data augmentation techniques for NER, focusing on domain-specific Portuguese datasets. They employed data augmentation techniques based on rules, backtranslation, and LLMs on four datasets of varying sizes to train transformer-based NER models. The authors in (Kamath & Vajjala, 2025) use LLM-based synthetic data generation with a small number of gold examples as an alternative to

automatically generated datasets for low-resource NER across 11 languages from 3 language families: Indo-European, Niger-Congo, and Dravidian.

To our knowledge, data augmentation has not yet been investigated for modern Serbian NER or its historical varieties. In this study, we aim to bridge this gap by combining transformer-based models with generative data augmentation. We further examine whether the effectiveness of synthetic augmentation varies across historical documents representing different degrees of chronological, genre-related, and orthographic distance from the training corpus.

3 Original Corpus

3.1 Data Sources

The datasets developed in this study build upon the corpus introduced in (Đokić Petrović et al., 2025), consisting of medieval Serbian administrative and legal texts, including the *Banjska Chrysobull* (Trifunović, 2011), the third example of the *Dečani Chrysobull* (Ivić & Grković, 1976), and a collection of charters and letters from the Dubrovnik archive (Stojanović, 1929). The *Banjska Chrysobull*, issued in the early 14$^{\text{th}}$ century during the reign of King Milutin (reigned 1282-1321), is a royal charter that regulates monastic endowments and property relations. The *Dečani Chrysobull*, issued in the early 14$^{\text{th}}$ century during the reign of King Stefan Dečanski (reigned 1322-1331), constitutes a comprehensive legal and administrative document detailing land ownership, privileges, and obligations associated with the Dečani monastery. The *Dubrovnik archival material* consists of administrative and diplomatic correspondence and charters from the 14$^{\text{th}}$ century, reflecting interactions between the Serbian medieval realm and the Republic of Ragusa (Dubrovnik). This domain profile is relevant to the subsequent evaluation, in which the models are applied to unseen documents representing different degrees of chronological, genre-related, and orthographic distance from the training corpus.

The given corpus includes personal names (`PER`), locations (`LOC`), and demographic or ethnonymic references (`DEMO`), whose distribution across the 4,112 manually annotated sentences is shown in Table 1.

Table 1: Distribution of annotated named entities in the corpus introduced in (Đokić Petrović et al., 2025)

Entity Type	Total	Unique
PER	7,562	2,490
LOC	1,525	1,275
DEMO	85	54

3.2 BIO-annotation

Based on the manually annotated corpus introduced in (Đokić Petrović et al., 2025), the original span-based annotations are converted into the BIO tagging scheme (Ramshaw & Marcus, 1995), which is widely used in sequence labeling tasks such as NER and is fully compatible with transformer-based architectures.

For this purpose, a programmatic conversion pipeline is implemented that first tokenizes each sentence in the original corpus and subsequently assigns BIO labels to individual tokens: `B-XXX`, indicating the beginning of an entity; `I-XXX`, indicating a token inside an entity; and `O`, marking tokens outside any annotated entity. This representation enables precise identification of entity boundaries, including multi-token entities.

Following the automatic conversion process, the resulting annotations are manually verified in order to correct conversion inconsistencies and previously mislabeled entities. Consequently, the entity counts in the final BIO-annotated dataset differ from those reported for the original span-based corpus, as some entity boundaries and category assignments were corrected during this process.

The distribution of tokens in the final BIO-annotated dataset is present in Table 2, while the complete BIO-annotated dataset is publicly available via Hugging Face[3]. Table 3 presents examples of sentences with the original span-based annotations and their corresponding conversion into the BIO format.

Table 2: Distribution of tokens in the BIO-annotated dataset built upon the corpus introduced in (Đokić Petrović et al., 2025)

BIO Label	Total	Unique Tokens	Unique Tokens (B+I)
B-PER	6,611	1,781	2,493
I-PER	881	712	
B-LOC	1,500	1,197	1,463
I-LOC	521	266	
B-DEMO	86	50	57
I-DEMO	7	7	

4 Model Fine-Tuning

4.1 Transformer Models

To investigate the effectiveness of transformer-based architectures for NER in pre-modern Serbian Cyrillic texts, we fine-tuned two pretrained language models: BERTić (`classla/bcms-bertic`)(Ljubešić & Lauc, 2021) and XLM-RoBERTa (`xlm-roberta-base`)(Conneau et al., 2020), both introduced in Section 2. BERTić, due to its specialization in closely related South Slavic languages and its demonstrated strong performance on NER tasks for contemporary Serbian (Ljubešić &

[3] https://huggingface.co/datasets/marijadjokic/pre-modern-serbian-cyrillic/blob/main/bio_annotated_dataset.json

Table 3: Examples of span-based annotations from the corpus introduced in (Đokić Petrović et al., 2025), converted into BIO format.

Sentence	**Span-based annotation**
Село чиготово доброслава грьбнце и съ своимн мегіамн ігова половина	{[5,13,"LOCATION"], [14,24,"PERSON"], [25, 32, "PERSON"]}
кралиев'ствоующаго в'сею срьв'ською Землею	{[24, 40, "LOCATION"]}
Tokens	**BIO labels**
["Село", "чиготово", "доброслава", "грьбнце", "и", "съ", "своимн", "мегіамн", "ігова", "половина"]	["O", "B-LOC", "B-PER", "I-PER", "O", "O", "O", "O", "O", "O"]
["кралиев'ствоующаго", "в'сею", "срьв'ською", "Землею"]	["O", "O", "B-LOC", "I-LOC"]

Lauc, 2021), is particularly suitable for evaluating the extent to which regionally specialized language representations can effectively transfer to historical Serbian texts. XLM-RoBERTa, as a multilingual transformer model with strong transfer capabilities across low-resource languages (including Serbian) and domains (Dorkin & Sirts, 2024), is particularly suitable for evaluating whether multilingual contextual representations can effectively capture the orthographic, lexical, and morphosyntactic variation characteristic of pre-modern Serbian Cyrillic texts.

4.2 Fine-Tuning Setup

Both transformer models are fine-tuned with the following setup:
(i) using the BIO-annotated dataset described in Section 3.2, and
(ii) using augmented datasets, where generated synthetic data was added to the training part of the dataset from (i), as described in Section 4.3.

To ensure a fair comparison between Old-Serb-BERTić and Old-Serb-XLM-RoBERTa, the same data split, fine-tuning procedure, and hyperparameters were used for all configurations. The BIO-annotated dataset was divided into training, development, and test sets using an 80/10/10 ratio. The same split was reused across all experiments. Synthetic examples, generated using the procedure described in Section 4.3, were added exclusively to the training set, while the development and test sets remained unchanged, thereby preventing data leakage. The number of sentences and distribution of entities across these sets are given in Table 4.

The models were fine-tuned for token classification using PyTorch (version `2.8.0+cu128`)[4] and the Hugging Face Transformers (version `5.12.0`)[5] libraries on an `NVIDIA RTX PRO 4500 Blackwell` in a `RunPod` cloud environment. The corpus is tokenized using the corresponding pretrained tokenizer for each model. Since transformer tokenizers may split individual words into multiple subword units, BIO labels

[4] https://pytorch.org/get-started/locally/
[5] https://huggingface.co/docs/transformers/en/index

Table 4: Distribution of sentences and entities across sets before data augmentation

	Training		Development		Test	
Sentences	3,289		411		412	
Entity	**Total**	**Unique**	**Total**	**Unique**	**Total**	**Unique**
B-PER	5,333	1,550	654	366	624	365
I-PER	708	582	78	76	94	90
B-LOC	1,191	980	147	144	142	134
I-LOC	410	224	55	46	55	41
B-DEMO	69	41	9	8	7	6
I-DEMO	5	5	0	0	2	2

are aligned with tokenized sequences by assigning the entity label only to the first subword token, while subsequent subword tokens are ignored during loss computation.

Training was performed using cross-entropy loss. Evaluation was conducted on the development set after each epoch, and the checkpoint achieving the highest entity-level F1-score on the development set was selected for final evaluation on the test set. The remaining fine-tuning parameters are summarized in Table 5.

Table 5: Fine-tuning parameters

Parameter	**Value**
Learning rate	2e-5
Number of epochs	6
Batch size per device	4
Gradient accumulation steps	4
Maximum sequence length	256
Weight decay	0.01
Warmup ratio	0.1
Evaluation strategy	epoch
Effective batch size	16
Random seed	42
Hardware	NVIDIA RTX PRO 4500 Blackwell
CUDA available	True
Model-selection metric	Entity-level F1 on the development set
Evaluation framework	seqeval

4.3 Data Augmentation

Although the *Banjska Chrysobull* and the *Dečani Chrysobull* are among the richest medieval Serbian sources of personal and geographical names (Grković, 1983; Loma, 2013), making them particularly valuable for NER tasks, the overall number of unique annotated entities in the training set is relatively small. This scarcity of annotated

material, combined with substantial orthographic and lexical variation, motivates the use of data augmentation to enrich the training set and improve model generalization. In that sense, we employ a generative data augmentation strategy.

The synthetic data were generated using the `GPT-5.4-mini` model (OpenAI, 2026) through a few-shot prompting approach, in which the model was conditioned exclusively on annotated sentences sampled from the training set. We first generated a synthetic dataset containing 3,289 sentences, corresponding to the size of the original training set. From this synthetic dataset, subsets containing 30% and 50% of the generated sentences were derived, while the complete synthetic dataset was used for the 100% configuration. Each synthetic subset was then added separately to the original training set[6], producing three augmented training configurations. The development and test sets remained unchanged across all experiments.

The generation pipeline consists of several steps (Figure 1). In the first step, three annotated sentences were randomly sampled from the training set for each prompt and used as few-shot demonstrations. Each example is represented as a JSON object containing a sequence of tokens and the corresponding BIO-formatted NER labels. In the second step, the prompting strategy constrains generation to the predefined entity categories `PER`, `LOC`, and `DEMO`. The model is additionally instructed to in general: (i) resemble the language represented in the examples; (ii) use only graphemes, abbreviations, diacritics, and orthographic patterns attested in the provided examples; (iii) preserve the same historical period and textual tradition as the examples; (iv) contain between 5 and 30 tokens and at least one named entity; (v) contain realistic entities; (vi) avoid copying the provided examples; and (vii) strictly follow the BIO tagging scheme. In the third step, each generated output was automatically validated to ensure structural correctness and BIO-format consistency. The generation process was repeated until the target dataset of 3,289 synthetic sentences was obtained.

Table 6 presents selected sentences produced using the proposed generative data augmentation pipeline. In addition to automatic structural validation, the generated data were manually inspected to assess their linguistic and orthographic plausibility. This inspection revealed that the `GPT-5.4-mini` model occasionally introduced graphemes associated with Old Church Slavonic or later Russian-Slavonic textual traditions that were not attested in the training set. These inconsistencies were not systematically corrected and were retained, allowing us to evaluate the robustness of the models under naturally occurring synthetic noise.

The complete synthetic dataset is publicly available on Hugging Face[7]. The synthetic subsets were separately merged with the original training set and used to fine-tune both Old-Serb-BERTić and Old-Serb-XLM-RoBERTa. The same synthetic examples were used for both model architectures, ensuring that differences in performance were not caused by differences in the augmented training data. Table 7 summarizes the distribution of sentences and named-entity labels in the original training set, the synthetic set, and the augmented training sets. As the proportion of

[6] https://huggingface.co/datasets/marijadjokic/pre-modern-serbian-cyrillic/tree/main/augmented-datasets

[7] https://huggingface.co/datasets/marijadjokic/pre-modern-serbian-cyrillic/blob/main/synthetic_dataset.json

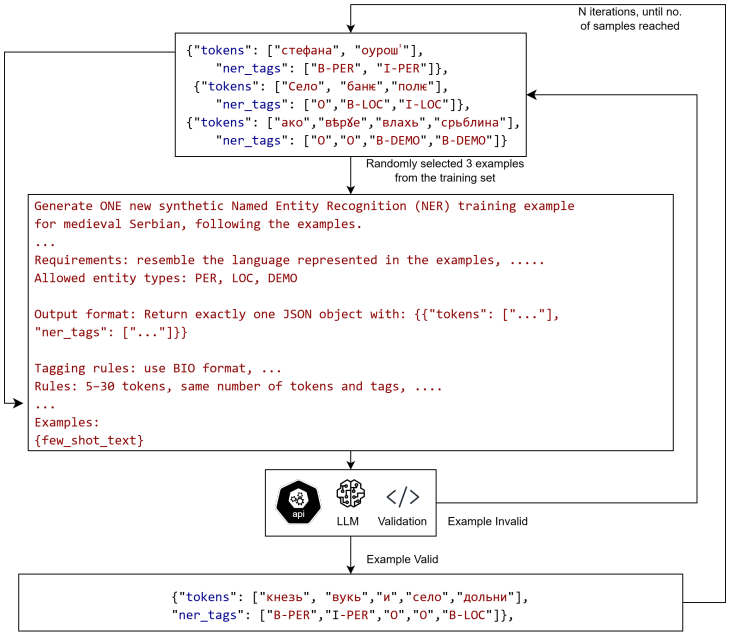


Figure 1: High-level overview of the synthetic data generation process. Model `GPT-5.4-mini` is used to generate new BIO-annotated NER sentences based on the training set.

synthetic data increases from 30% to 100%, both the total and unique numbers of entity labels increase progressively. The largest relative increases are observed for `LOC` and `DEMO`, suggesting that augmentation not only expands the training data but also increases entity diversity and partially reduces the strong dominance of `PER` entities in the original training set.

5 Evaluation Results

The evaluation was conducted under two experimental conditions: (i) in-domain evaluation on the unchanged test set and (ii) out-of-domain evaluation on three previously unseen historical documents. In-domain performance is reported using entity-level precision (P), recall (R), and F1-score, whereas the out-of-domain results are presented as token-level weighted averages across BIO labels. Detailed class-level results and confusion matrices for the out-of-domain evaluation are provided in Appendix A.

This experimental setup allows us to compare: (i) transformer-based models against the previously developed CNN-based NER model in (Đokić Petrović et al., 2025), and (ii) the impact of synthetic data on NER performance in a low-resource Serbian historical setting.

5.1 In-Domain Evaluation

Table 8 presents the in-domain evaluation results.

Table 6: Examples of synthetically generated sentences applying the generative data augmentation technique

Sentence/Tokens	BIO labels
[”мастиконъ”, ”кнѩзь”, ”влахо”, ”приде”, ”в”, ”село”, ”дѣчане”]	["O", "B-PER", "I-PER", "O", "O", "O", "B-LOC"]
[”а”, ”сѩ”, ”имь”, ”воислављь”, ”и”, ”жупанъ”, ”дражинъ”, ”от”, ”стоупьника”]	["O", "O", "O", "B-PER", "O", "B-DEMO", "B-PER", "O", "B-LOC"]
[”в”, ”село”, ”жрѣбчане”, ”приде”, ”деспоть”, ”стефань”]	["O", "O", "B-LOC", "O", "O", "B-PER"]
[”драгославь”, ”и”, ”братоіе”, ”радоіе”, ”вх”, ”село”, ”мирово”]	["B-PER", "O", "B-PER", "I-PER", "O", "B-LOC", "I-LOC"]

Table 7: Distribution of sentences and named entities across the training set and augmented sets (training set + synthetic sets). Values are presented as total/unique.

Dataset	No. of sent.	B-PER	I-PER	B-LOC	I-LOC	B-DEMO	I-DEMO
Training	3,289	5,333/1,550	708/582	1,191/980	410/224	69/41	5/5
Synthetic	3,289	4,577/1819	527/442	4,035/2,247	537/404	319/258	32/32
Augmented 30%	4,276	6,677/2,098	855/703	2,390/1,765	565/345	172/127	15/15
Augmented 50%	4,933	7,587/2,397	974/792	3,185/2,168	659/412	241/179	24/24
Augmented 100%	6,578	9,910/3,049	1,235/981	5,226/3,085	947/585	388/283	37/37

Table 8: In-domain evaluation results. Values are reported as P/R/F1.

Model	Synth.	DEMO	LOC	PER	Overall
Old-Serb-BERTić	0%	0.00/0.00/0.00	0.63/0.75/0.68	0.91/0.92/0.92	0.85/0.88/0.8649
	30%	0.00/0.00/0.00	0.61/0.75/0.67	0.90/0.93/0.92	0.84/0.89/0.8649
	50%	0.00/0.00/0.00	0.65/0.72/0.68	0.90/0.93/0.92	0.85/0.88/0.8684
	100%	0.33/0.14/0.20	0.66/0.75/0.70	0.89/0.92/0.90	0.84/0.88/0.8604
Old-Serb-XLM-RoBERTa	0%	0.33/0.14/0.20	0.77/0.90/0.83	0.92/0.95/0.93	0.89/0.93/0.9073
	30%	0.57/0.57/0.57	0.78/0.93/0.85	0.94/0.95/0.94	0.90/0.94/0.9217
	50%	0.36/0.57/0.44	0.77/0.87/0.82	0.93/0.95/0.94	0.89/0.93/0.9084
	100%	0.71/0.71/0.71	0.81/0.88/0.84	0.92/0.96/0.94	0.90/0.94/0.9172
CNN-based NER	–	0.78/0.54/0.64	0.83/0.69/0.75	0.94/0.96/0.95	0.93/0.91/0.92

5.2 Out-of-Domain Evaluation

To evaluate the robustness and generalization capabilities of the models beyond the training distribution, we conducted out-of-domain evaluation on a collection of unseen historical documents spanning different centuries and textual genres. The evaluation corpus includes:

Table 9: Out-of-domain weighted evaluation results on unseen historical documents

Test Document	Model	Synth.	Precision	Recall	F1-score
Will of Miloš Belmužević	Old-Serb-BERTić	0%	0.80	0.85	0.800
		30%	0.80	0.84	0.816
		50%	0.80	0.84	0.798
		100%	0.83	0.85	0.816
	Old-Serb-XLM-RoBERTa	0%	0.86	0.87	0.851
		30%	0.85	0.86	0.853
		50%	0.87	0.88	0.871
		100%	0.87	0.89	0.878
	CNN-based NER model	0%	0.12	0.16	0.14
St. Georg Church (Vienna)	Old-Serb-BERTić	0%	0.84	0.84	0.837
		30%	0.84	0.82	0.824
		50%	0.83	0.85	0.820
		100%	0.85	0.80	0.814
	Old-Serb-XLM-RoBERTa	0%	0.84	0.85	0.839
		30%	0.91	0.89	0.892
		50%	0.90	0.90	0.897
		100%	0.87	0.88	0.871
	CNN-based NER model	0%	0.31	0.16	0.20
Charters of Emperor Stefan Uroš	Old-Serb-BERTić	0%	0.94	0.94	0.940
		30%	0.94	0.92	0.932
		50%	0.95	0.96	0.947
		100%	0.95	0.95	0.947
	Old-Serb-XLM-RoBERTa	0%	0.96	0.96	0.954
		30%	0.97	0.96	0.964
		50%	0.97	0.97	0.964
		100%	0.97	0.96	0.965
	CNN-based NER model	0%	0.06	0.11	0.06

- *Will of Miloš Belmužević* (dated to 1500) (Obradović, 2024; Polomac, 2025)
- *Manuscript preserved in the Greek Orthodox Church of St. Georg in Vienna* (dated to 1778) (Plöchl, 1983; Ransmayr, 2018; Seirinidou, 2011; Turczynski, 1959)
- *Charters of Emperor Stefan Uroš to the people of Dubrovnik* (Porčić, 2017) (dated between 1356 and 1362).

The three documents represent different degrees of domain shift. The *Charters of Emperor Stefan Uroš* is chronologically and textually closest to the medieval documentary material used for training, while the *Will of Miloš Belmužević* represents a somewhat later but still related legal-documentary context. The *St. Georg Church (Vienna)* document represents the greatest chronological, genre-related, and orthographic distance from the non-augmented training corpus. Table 9 summarizes the weighted out-of-domain results.

6 Discussion

The evaluation results show that transformer-based models, particularly Old-Serb-XLM-RoBERTa, are more robust than the earlier developed CNN-based NER model, especially in out-of-domain evaluation. The results also suggest that synthetic data augmentation can improve model generalization, although the two transformer architectures respond differently to augmentation. Synthetic data yields only limited and

inconsistent improvements for Old-Serb-BERTić: its in-domain scores remain similar across the four configurations, while its out-of-domain results fluctuate without a clear systematic trend. Old-Serb-XLM-RoBERTa benefits more consistently from augmentation, but the effect is not linear. The configuration trained with 30% synthetic data achieves the highest in-domain F1-score of 0.9217 (≈ 0.92), while the 50% configuration obtains the highest average weighted F1-score of ≈ 0.91 across the three out-of-domain documents, with the best individual result on the *St. Georg Church (Vienna)* document. The 100% configuration performs best on the *Will of Miloš Belmužević* and the *Charters of Emperor Stefan Uroš*. For this reason, the 50% configuration is selected as the recommended model for publication and broader use.

These findings indicate that no single augmentation ratio is optimal across all evaluation settings and confirm that the degree of domain shift strongly influences model performance. The *Will of Miloš Belmužević* and the *St. Georg Church (Vienna)* documents are more challenging for both transformer models, while the *Charters of Emperor Stefan Uroš* produces the highest weighted F1-scores, probably because it is closer to the medieval diplomatic tradition represented in the training corpus.

The results also indicate that Old-Serb-XLM-RoBERTa performs better than Old-Serb-BERTić, particularly in out-of-domain evaluation settings. While BERTić is primarily pretrained on contemporary BCMS language varieties, XLM-RoBERTa benefits from multilingual pretraining on a substantially larger and more diverse corpus. This broader linguistic exposure may contribute to greater robustness in handling orthographic inconsistencies and morphologically variable expressions characteristic of historical Serbian Cyrillic texts. However, this explanation should be interpreted cautiously, since the present experiments do not isolate the effects of pretraining data, tokenizer coverage, and model architecture.

The results in Appendix A indicate that `PER` entities are recognized most reliably, which can be explained by their higher frequency in the training data and by relatively stable contextual patterns surrounding personal names. `LOC` entities show more variable performance, reflecting not only the high degree of orthographic variation in historical toponyms, but also difficulties with multi-token institutional place expressions and forms that overlap with demonymic references. `DEMO` entities remain the most challenging category, primarily because they are rare in the dataset and often overlap semantically with personal, common, or geographic references. In addition, a substantial number of gold entity tokens are predicted as `O`, indicating that the model often fails to detect entities rather than merely assigning them to the wrong category. These findings suggest that low-frequency entity types require either more annotated examples or more targeted augmentation strategies.

The qualitative examples in Table 10, taken from Old-Serb-XLM-RoBERTa trained with 50% synthetic data, illustrate the main error patterns. The model correctly recognizes many complex personal names and locations, but demonymic expressions remain difficult to distinguish. Also, in some false-positive examples, common nouns are interpreted as personal names or locations. These examples show that synthetic data expands the range of recognizable forms but does not fully resolve semantic ambiguity or historical spelling variation.

Table 10: Qualitative out-of-domain examples for Old-Serb-XLM-RoBERTa trained with 50% synthetic data

Document	Example	Gold	Prediction	Type
Will of Miloš Belmužević	ѹгарски арсагь	LOC	LOC	Correct
	погане	DEMO	LOC	Label confusion
	тимотею	PER	O	Missed entity
	ѹтофиѝ	O	PER	False positive
St. Georg Church (Vienna)	іоанна сечани	PER	PER	Correct
	георгія	LOC	PER	Label confusion
	вїенићь	LOC	O	Missed entity
	граначнра	O	PER	False positive
Charter of Emperor Uroš	стефанѹ ѹрѡѹ	PER	PER	Correct
	грьке	DEMO	LOC	Label confusion
	призрѣнь	LOC	O	Missed entity
	севастъ	O	PER	False positive

Overall, the results confirm that transformer-based models are more suitable than the earlier developed CNN-based NER model. Synthetic data augmentation can improve generalization, but its effectiveness depends on the augmentation ratio, the pretrained model, and the characteristics of the evaluation document. Old-Serb-XLM-RoBERTa trained with 30% synthetic data remains the strongest in-domain model, while the 50% configuration offers the most balanced out-of-domain performance and is therefore the primary published model.

7 Conclusions and Future Work

This study investigated transformer-based NER for pre-modern Serbian Cyrillic texts and examined whether synthetic data augmentation can improve performance in a low-resource historical setting. Two pretrained language models, BERTić and XLM-RoBERTa, were fine-tuned on the synthetic augmented training sets and evaluated both in-domain and on three unseen documents representing different periods, genres, and degrees of linguistic distance from the training data.

The results show that Old-Serb-XLM-RoBERTa delivers better, more robust performance than Old-Serb-BERTić and the previously developed CNN-based NER model, particularly in out-of-domain evaluation. The configuration trained with 30% synthetic data achieves the highest in-domain F1-score, while the 50% configuration obtains the strongest average performance in the out-of-domain evaluation. It is therefore selected as the primary model for publication and broader use. However, the results also demonstrate that no single augmentation ratio is optimal across all evaluation settings. The effectiveness of synthetic data depends on the underlying model and on the historical, linguistic, and genre-related characteristics of the evaluation document.

Synthetic data augmentation is particularly useful for increasing the diversity of `LOC` and `DEMO` entities, but it does not fully address the challenges posed by class imbalance, orthographic variation, and semantic overlap among entity categories.

Future work will include repeated training with different random seeds to assess the stability and statistical significance of the observed differences. In addition, the generation of targeted synthetic data for rare entity categories, improved control over the linguistic quality and distribution of the tags of generated examples, and the expansion of the training set with additional annotated medieval Serbian texts will be explored.

Declarations

- **Funding**: Not applicable.
- **Competing interests**: The authors declare that they have no competing interests.
- **Ethics approval**: Not applicable. This study is based on historical documents and does not involve human participants or living individuals.
- **Consent to participate**: Not applicable.
- **Consent for publication**: Not applicable.
- **Data availability**: The BIO-annotated dataset(s) generated in this study are publicly available on Hugging Face and GitHub. The data that was used for creating BIO-annotated dataset(s) in this study is available on GitHub.
- **Code availability**: The code used for synthetic data generation, model training, and evaluation is publicly available on GitHub.
- **Authors' contributions**: All authors contributed to the conception and design of the study. The datasets were developed by M. Đ. P. and V. P. Model training was performed by M. Đ. P. The evaluation of the models and analysis of the results were conducted by M. Đ. P. and V.P. The first draft of the manuscript was written by M. Đ. P. The overview of the historical materials was prepared by M. P. All authors contributed to the review and editing of the manuscript and approved the final version.

References

Agić, Ž., & Ljubešić, N. (2014, May). The SETimes.HR Linguistically Annotated Corpus of Croatian. N. Calzolari et al. (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1724–1727). Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L14-1542/

Batanović, V., Ljubešić, N., Samardžić, T., Erjavec, T. (2018). *Training corpus SETimes.SR 1.0.* Retrieved from http://hdl.handle.net/11356/1200 (Slovenian language resource repository CLARIN.SI)

Batanović, V., Ljubešić, N., Samardžić, T. (2018). SETimes. SR–a reference training corpus of Serbian. *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)* (p. 11-17). Retrieved from https://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020, July). Unsupervised Cross-lingual Representation Learning at Scale. D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.747/

Dai, X., & Adel, H. (2020, December). An Analysis of Simple Data Augmentation for Named Entity Recognition. D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3861–3867). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.343/

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423/

Ding, B., Liu, L., Bing, L., Kruengkrai, C., Nguyen, T.H., Joty, S., ... Miao, C. (2020, November). DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks. B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6045–6057). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.488/

Dorkin, A., & Sirts, K. (2024, March). TartuNLP @ SIGTYP 2024 Shared Task: Adapting XLM-RoBERTa for Ancient and Historical Languages. M. Hahn et al. (Eds.), *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP* (pp. 120–130). St. Julian's, Malta: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.sigtyp-1.15/

Elwing Torres, A., Silva de Moura, E., Soares da Silva, A., Nascimento, M.A., Mesquita, F. (2026, March). An Experimental Study on Data Augmentation Techniques for Named Entity Recognition on Low-resource Domains. *Natural Language Processing*, *32*(2), 184–203, https://doi.org/10.1017/nlp.2026.10019

Evuru, C.K., Ghosh, S., Kumar, S., S, R., Tyagi, U., Manocha, D. (2024, June). CoDa: Constrained Generation based Data Augmentation for Low-Resource NLP. K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3754–3769). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://

aclanthology.org/2024.findings-naacl.238/

Gladstone, C., Fang, Z., Stewart, S.D. (2025). Ground Truth Generation for Multilingual Historical NLP using LLMs. T. Arnold, M. Fantoli, & R. Ros (Eds.), *Computational humanities research 2025* (Vol. 3, pp. 1323–1335). Retrieved from https://doi.org/10.63744/UWoDSxRk90Vn

Glavaš, G., & Šnajder, J. (2015, September). Resolving entity coreference in Croatian with a constrained mention-pair model. J. Piskorski, L. Pivovarova, J. Šnajder, H. Tanev, & R. Yangarber (Eds.), *The 5th Workshop on Balto-Slavic Natural Language Processing* (pp. 17–23). Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA. Retrieved from https://aclanthology.org/W15-5303/

Grković, M. (1983). *Imena u Dečanskim hrisovuljama*. Filozofski fakultet u Novom Sadu, Institut za južnoslovenske jezike. Retrieved from www.svevlad.org.rs/imenoslov_files/grkovic_imena_decanskih_hrisovulja.pdf

Hiltmann, T., Dröge, M., Dresselhaus, N., Grallert, T., Althage, M., Bayer, P., ... others (2025). NER4all or Context is All You Need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach. *arXiv preprint arXiv:2502.04351*, , Retrieved from https://arxiv.org/abs/2502.04351 arXiv:2502.04351 [cs.CL]

Hu, X., Jiang, Y., Liu, A., Huang, Z., Xie, P., Huang, F., ... Yu, P.S. (2023). *Entity-to-Text based Data Augmentation for various Named Entity Recognition Tasks* (A. Rogers, J. Boyd-Graber, & N. Okazaki, Eds.). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.578/

Huang, Y., Gao, Y., Ren, C. (2025). A Survey of Data Augmentation in Named Entity Recognition. *Neurocomputing*, *651*, 130856, https://doi.org/https://doi.org/10.1016/j.neucom.2025.130856 Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231225015280

Ikonić Nešić, M., Petalinkar, S., Stanković, R., Mitkov, R. (2025, September). From Zero to Hero: Building Serbian NER from Rules to LLMs. A. Picazo-Izquierdo, E.L. Estevanell-Valladares, R. Mitkov, R. Munoz Guillena, & R. Garcia Cerda (Eds.), *Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models* (pp. 87–96). Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria. Retrieved from https://aclanthology.org/2025.r2lm-1.10/

Ikonić Nešić, M., Petalinkar, S., Škorić, M., Stanković, R. (2024). BERT Downstream Task Analysis: Named Entity Recognition in Serbian. M. Trajanović, N. Filipović, & M. Zdravković (Eds.), *Disruptive information technologies for a smart society* (Vol. 860). Cham: Springer. Retrieved from https://doi.org/10.1007/

978-3-031-71419-1_29

Ivić, P., & Grković, M. (1976). *Dečanske hrisovulje*. Novi Sad, Institut za lingvistiku.

Kalušev, V., & Brkljač, B. (2026). Named Entity Recognition for Serbian Legal Documents: Design, Methodology and Dataset Development. M. Trajanović, N. Filipovic, & M. Zdravković (Eds.), *Transformative Technologies Shaping a Smarter Society* (pp. 403–420). Cham: Springer Nature Switzerland.

Kamath, G., & Vajjala, S. (2025, December). Does Synthetic Data Help Named Entity Recognition for Low-Resource Languages? K. Inui et al. (Eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (pp. 159–167). Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.ijcnlp-short.15/

Keraghel, I., Morbieu, S., Nadif, M. (2024). *Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study.* Retrieved from https://arxiv.org/abs/2401.10825

Krstev, C. (1997). *Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije* (Doctoral dissertation, University of Belgrade (Serbia)). Retrieved from https://phaidrabg.bg.ac.rs/view/o:9574

Krstev, C., Obradović, I., Utvić, M., Vitas, D. (2014). A system for Named Entity Recognition based on Local Grammars. *Journal of Logic and Computation*, *24*(2), 473-489, https://doi.org/10.1093/logcom/exs079

Lazić, B., & Škorić, M. (2020). From DELA based dictionary to Leximirka lexical database. *Infotheca - Journal for Digital Humanities*, *19*(2), 81–98, https://doi.org/10.18485/infotheca.2019.19.2.4 Retrieved from https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2019.19.2.4_en

Ljubešić, N., Erjavec, T., Batanović, V., Miličević, M., Samardžić, T. (2019). *Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.1.* Retrieved from http://hdl.handle.net/11356/1240 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., & Lauc, D. (2021, April). BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. B. Babych et al. (Eds.), *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing* (pp. 37–42). Kiyv, Ukraine: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.bsnlp-1.5/

Ljubešić, N., Stupar, M., Jurić, T., Agić, Ž. (2013). Combining available datasets for building named entity recognition models of Croatian and Slovene. *Slovenščina*, *2*(1), 35–57, Retrieved from http://www.dlib.si/?URN=URN:NBN:SI:DOC-VSWXF4CE

Loma, A. (2013). *Toponomija Banjske hrisovulje: ka osmišljenju starosrpskog toponomastičkog rečnika i boljem poznavanju opšteslovenskih imenoslovnih obrazaca.* Beograd: Srpska akademija nauka i umetnosti. Retrieved from https://hdl.handle.net/21.15107/rcub_dais_2187

Marovac, U.A., Avdić, A.R., Milošević, N.L. (2023). *A Survey of Resources and Methods for Natural Language Processing of Serbian Language.* Retrieved from https://arxiv.org/abs/2304.05468

Moreira, H., Silva, P.F.d., Bencke, L., Moreira, V. (2026, April). Data Augmentation for Named Entity Recognition in Domain-Specific Scenarios in Portuguese. M. Souza, I. de Dios-Flores, D. Santos, L. Freitas, J.W.d.C. Souza, & E. Ribeiro (Eds.), *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1* (pp. 250–259). Salvador, Brazil: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2026.propor-1.25/

Munnangi, M. (2024). *A Brief History of Named Entity Recognition.* Retrieved from https://arxiv.org/abs/2411.05057

Nešić, M.I., Petalinkar, S., Stanković, R., Utvić, M., Kitanović, O. (2024). SrpCNNeL: Serbian Model for Named Entity Linking. *2024 19th conference on computer science and intelligence systems (fedcsis)* (p. 465-473).

Obradović, N. (2024). Testament Miloša Belmuževića. *Mešovita građa (miscellanea)*(45), 35–56, Retrieved from https://www.iib.ac.rs/assets/files/MGIC2473035O.pdf

OpenAI (2026). *GPT-5.4 Mini.* https://developers.openai.com/api/docs/models/gpt-5.4-mini. (OpenAI API Documentation, accessed: 2026-06-14)

Plöchl, W.M. (1983). *Die Wiener orthodoxen Griechen: eine Studie zur Rechts- und Kulturgeschichte der Kirchengemeinden zum Hl. Georg und zur Hl. Dreifaltigkeit und zur Errichtung der Metropolis von Austria.* Wien: Verlag des Verbandes der wissenschaftlichen Gesellschaften Österreichs.

Polomac, V. (2025). Srpski jezik u Ugarskoj krajem XV veka (na primeru testamenta vojvode Miloša Belmuževića). M. Kovačević (Ed.), *Aktuelna pitanja istorije srpskog jezika* (p. 277-309). Andrićgrad: Andrićev institut.

Porčić, N. (2017). *Документи српских средњовековних владара у дубровачким збиркама: доба Немањића, Documents of Serbian Medieval Rulers in Dubrovnik Collections: the Nemanjić Period* (D. Bataković T., Ed.). Београд: Балканолошки институт Српске академије наука и уметности/ Belgrade: Institute for Balkan Studies SASA. Retrieved from https://hdl.handle.net/21.15107/rcub_dais_3926

Ramshaw, L., & Marcus, M. (1995). Text Chunking using Transformation-Based Learning. *Third Workshop on Very Large Corpora.* Retrieved from https://aclanthology.org/W95-0107/

Ransmayr, A. (2018). *Untertanen des Sultans oder des Kaisers: Struktur und Organisationsformen der beiden Wiener griechischen Gemeinden von den Anfängen im 18. Jahrhundert bis 1918* (1st ed.). Deutschland: V&R unipress, Vienna University Press. (Hochschulschrift)

Šandrih, B., Krstev, C., Stankovic, R. (2019, September). Development and Evaluation of Three Named Entity Recognition Systems for Serbian - The Case of Personal Names. R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 1060–1068). Varna, Bulgaria: INCOMA Ltd. Retrieved from https://aclanthology.org/R19-1122/

Šandrih Todorović, B., Krstev, C., Stanković, R., Ikonić Nešić, M. (2021, September). Serbian NER&beyond: The archaic and the modern intertwinned. R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 1252–1260). Held Online: INCOMA Ltd. Retrieved from https://aclanthology.org/2021.ranlp-1.141/

Seirinidou, V. (2011). *Greeks in Vienna (18th-first half of the 19th century).* Athens: Herodotus.

Song, S., Shen, F., Zhao, J. (2024, Mar.). RoPDA: Robust Prompt-Based Data Augmentation for Low-Resource Named Entity Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, p. 19017–19025). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/29868

Stanković, R., Krstev, C., Šandrih Todorović, B., Škorić, M. (2022). Annotation of the Serbian ELTeC Collection. *Infotheca - Journal for Digital Humanities*, *21*(2), 43–59, https://doi.org/10.18485/infotheca.2021.21.2.3 Retrieved from https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.2.3_en

Stojanović, Lj. (1929). *Stare srpske povelje i pisma* (No. 19). Beograd and Sremski Karlovci: Srpska kraljevska akademija. (Knjiga 1, Prvi deo)

Torge, S., Politov, A., Lehmann, C., Saffar, B., Tao, Z. (2023, May). Named Entity Recognition for Low-Resource Languages - Profiting from Language Families. J. Piskorski et al. (Eds.), *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)* (pp. 1–10). Dubrovnik, Croatia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.bsnlp-1.1/

Trifunović, Đ. (2011). *Povelja kralja Milutina manastiru Banjska-Svetostefanska hrisovulja*. Službeni glasnik. Retrieved from https://books.google.lu/books?id=ZjQeMwEACAAJ

Turczynski, E. (1959). *Die deutsch-greichischen Kulturbeziehungen: bis zur Berufung König Ottos*. München, R. Oldenbourg.

Vitas, D., & Pavlović-Lažetić, G. (2008). Resources and Methods for Named Entity Recognition in Serbian. *INFOtheca-Journal of Informatics & Librarianship*, *9*(1-2), 35a, Retrieved from https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/509/487

Ye, J., Xu, N., Wang, Y., Zhou, J., Zhang, Q., Gui, T., Huang, X. (2024). LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition. *ArXiv*, *abs/2402.14568*, , Retrieved from https://api.semanticscholar.org/CorpusID:267782766

Zhong, T., Yang, Z., Liu, Z., Zhang, R., You, W., Liu, Y., ... Liu, T. (2026). *Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research*. Retrieved from https://arxiv.org/abs/2412.04497

Zhou, R., Li, X., He, R., Bing, L., Cambria, E., Si, L., Miao, C. (2022, May). MELM: Data Augmentation with Masked Entity Language Modeling for Low-Resource NER. S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2251–2262). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.160/

Đokić Petrović, M., Popović, M.S., Polomac, V. (2025). Named Entity Recognition for Pre-Modern Serbian: A Preliminary Study. J. Moskovljević Popović & R. Stanković (Eds.), *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig: Thematic Collection of Papers* (Vol. 1, pp. 295–309). Belgrade: University of Belgrade — Faculty of Philology. other. Retrieved from http://doi.fil.bg.ac.rs/pdf/eb_ser/judig/2025-1/judig-2025-1-ch17.pdf

A Detailed Out-of-Domain Evaluation Results

This appendix presents detailed classification reports and confusion matrices for all models evaluated on the out-of-domain historical documents introduced in Section 5.2.

A.1 Classification Reports

A.2 Confusion Matrices

Table 11: Classification reports for the out-of-domain evaluation on the *Will of Miloš Belmuževi*ć (dated from 1500)

Model	Synth.	Metric	B-DEMO	B-LOC	B-PER	I-DEMO	I-LOC	I-PER	O
Old-Serb-BERTić	0%	P	0.00	0.33	0.36	0.00	1.00	0.80	0.86
		R	0.00	0.04	0.12	0.00	0.08	0.40	0.99
		F1	0.00	0.07	0.18	0.00	0.14	0.53	0.92
	30%	P	0.00	0.47	0.35	0.00	0.30	0.60	0.88
		R	0.00	0.32	0.18	0.00	0.23	0.30	0.95
		F1	0.00	0.38	0.24	0.00	0.26	0.40	0.91
	50%	P	0.00	0.27	0.38	0.00	0.00	1.00	0.97
		R	0.00	0.15	0.23	0.00	0.00	0.17	0.99
		F1	0.00	0.19	0.29	0.00	0.00	0.29	0.98
	100%	P	0.00	0.18	0.49	0.00	0.00	1.00	0.97
		R	0.00	0.22	0.25	0.00	0.00	0.19	0.99
		F1	0.00	0.19	0.33	0.00	0.00	0.32	0.98
Old-Serb-XLM-RoBERTa	0%	P	0.00	0.56	0.61	0.00	1.00	0.63	0.90
		R	0.00	0.36	0.42	0.00	0.08	0.50	0.97
		F1	0.00	0.44	0.50	0.00	0.14	0.56	0.93
	30%	P	0.00	0.40	0.67	0.00	0.43	0.46	0.91
		R	0.00	0.36	0.49	0.00	0.23	0.60	0.95
		F1	0.00	0.38	0.56	0.00	0.30	0.52	0.93
	50%	P	0.00	0.45	0.74	0.00	0.60	0.86	0.91
		R	0.00	0.32	0.52	0.00	0.46	0.60	0.97
		F1	0.00	0.38	0.61	0.00	0.52	0.71	0.94
	100%	P	0.00	0.56	0.65	0.00	0.75	0.63	0.92
		R	0.00	0.32	0.67	0.00	0.46	0.50	0.96
		F1	0.00	0.41	0.66	0.00	0.57	0.56	0.94
			DEMO	**LOC**	**PER**	**O**			
CNN-based NER		P	0.00	0.14	0.30	0.00			
		R	0.00	0.20	0.38	0.00			
		F1	0.00	0.16	0.33	0.00			

Table 12: Classification reports for the out-of-domain evaluation on a *Manuscript preserved in the Greek Orthodox Church of St. Georg in Vienna* (dated from 1778)

Model	Synth.	Metric	B-DEMO	B-LOC	B-PER	I-DEMO	I-LOC	I-PER	O
Old-Serb-BERTić	0%	P	0.00	0.07	0.57	0.00	0.17	0.86	0.92
		R	0.00	0.08	0.68	0.00	0.08	0.55	0.93
		F1	0.00	0.08	0.62	0.00	0.11	0.67	0.92
	30%	P	0.00	0.29	0.42	0.00	0.20	0.83	0.92
		R	0.00	0.50	0.53	0.00	0.17	0.46	0.89
		F1	0.00	0.36	0.47	0.00	0.18	0.59	0.91
	50%	P	0.00	0.13	0.50	0.00	0.43	1.00	0.89
		R	0.00	0.08	0.37	0.00	0.25	0.09	0.96
		F1	0.00	0.10	0.42	0.00	0.32	0.17	0.93
	100%	P	0.00	0.25	0.69	0.00	0.19	1.00	0.91
		R	0.00	0.67	0.47	0.00	0.25	0.27	0.88
		F1	0.00	0.36	0.56	0.00	0.21	0.43	0.89
Old-Serb-XLM-RoBERTa	0%	P	0.00	0.20	0.60	0.00	0.50	0.50	0.91
		R	0.00	0.25	0.63	0.00	0.17	0.18	0.95
		F1	0.00	0.22	0.62	0.00	0.25	0.27	0.93
	30%	P	0.00	0.33	0.67	0.00	1.00	0.83	0.95
		R	0.00	0.50	0.84	0.00	0.42	0.91	0.93
		F1	0.00	0.40	0.74	0.00	0.59	0.87	0.94
	50%	P	0.00	0.47	0.76	0.00	0.33	0.92	0.95
		R	0.00	0.58	0.84	0.00	0.17	1.00	0.95
		F1	0.00	0.52	0.80	0.00	0.22	0.96	0.95
	100%	P	0.00	0.50	0.53	0.00	0.00	0.60	0.95
		R	0.00	0.42	0.84	0.00	0.00	0.82	0.94
		F1	0.00	0.46	0.65	0.00	0.00	0.69	0.95
			DEMO	**LOC**	**PER**	**O**			
CNN-based NER		P	0.00	0.22	0.47	0.00			
		R	0.00	0.22	0.21	0.00			
		F1	0.00	0.22	0.29	0.00			

Table 13: Classification reports for the out-of-domain evaluation on the *Charters of Emperor Stefan Uroš to the people of Dubrovnik* (dated between 1356 and 1362)

Model	**Synth.**	**Metric**	**B-DEMO**	**B-LOC**	**B-PER**	**I-DEMO**	**I-LOC**	**I-PER**	**O**
Old-Serb-BERTić	0%	P	0.00	0.02	0.25	0.00	0.00	0.87	0.97
		R	0.00	0.02	0.29	0.00	0.00	0.31	0.98
		F1	0.00	0.02	0.27	0.00	0.00	0.46	0.97
	30%	P	0.00	0.12	0.15	0.00	0.00	0.55	0.98
		R	0.00	0.28	0.34	0.00	0.00	0.29	0.95
		F1	0.00	0.17	0.21	0.00	0.00	0.38	0.96
	50%	P	0.00	0.27	0.38	0.00	0.00	1.00	0.97
		R	0.00	0.15	0.23	0.00	0.00	0.17	0.99
		F1	0.00	0.19	0.29	0.00	0.00	0.29	0.98
	100%	P	0.00	0.18	0.49	0.00	0.00	1.00	0.97
		R	0.00	0.22	0.25	0.00	0.00	0.19	0.99
		F1	0.00	0.19	0.33	0.00	0.00	0.32	0.98
Old-Serb-XLM-RoBERTa	0%	P	1.00	0.39	0.32	0.00	0.00	0.70	0.98
		R	0.03	0.30	0.39	0.00	0.00	0.45	0.99
		F1	0.05	0.34	0.35	0.00	0.00	0.55	0.98
	30%	P	0.64	0.42	0.38	0.00	0.00	0.77	0.98
		R	0.40	0.47	0.42	0.00	0.00	0.55	0.99
		F1	0.49	0.44	0.39	0.00	0.00	0.64	0.98
	50%	P	0.60	0.54	0.46	0.00	0.00	0.79	0.98
		R	0.30	0.48	0.43	0.00	0.00	0.52	0.99
		F1	0.40	0.51	0.44	0.00	0.00	0.63	0.98
	100%	P	0.65	0.53	0.35	0.00	0.00	0.69	0.99
		R	0.50	0.52	0.48	0.00	0.00	0.52	0.98
		F1	0.56	0.52	0.40	0.00	0.00	0.60	0.98
			DEMO	**LOC**	**PER**	**O**			
CNN-based NER		P	0.33	0.84	0.18	0.00			
		R	0.02	0.34	0.47	0.00			
		F1	0.04	0.14	0.26	0.00			

Table 14: Confusion matrices for the out-of-domain evaluation on the *Will of Miloš Belmuževi*ć (dated from 1500)

Model	Synth.	True Label	P-B-DEMO	P-B-LOC	P-B-PER	P-I-DEMO	P-I-LOC	P-I-PER	P-O
Old-Serb-BERTić	0%	T-B-DEMO	0	0	0	0	0	0	2
		T-B-LOC	0	1	2	0	0	0	25
		T-B-PER	0	1	4	0	0	0	28
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	0	0	1	0	12
		T-I-PER	0	0	0	0	0	4	6
		T-O	0	1	5	0	0	1	449
	30%	T-B-DEMO	0	0	0	0	0	0	2
		T-B-LOC	0	9	1	0	0	1	17
		T-B-PER	0	0	6	0	1	0	26
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	0	0	3	0	10
		T-I-PER	0	0	3	0	0	3	4
		T-O	0	10	7	0	6	1	432
	50%	T-B-DEMO	0	0	0	0	0	0	2
		T-B-LOC	0	2	0	0	0	0	26
		T-B-PER	0	0	5	0	0	0	28
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	0	0	2	0	11
		T-I-PER	0	0	2	0	0	1	7
		T-O	0	4	1	0	5	0	446
	100%	T-B-DEMO	0	0	0	0	0	0	2
		T-B-LOC	0	3	0	0	0	0	25
		T-B-PER	0	1	6	0	0	0	26
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	0	0	2	0	11
		T-I-PER	0	1	1	0	0	3	5
		T-O	0	3	0	0	4	0	449
Old-Serb-XLM-RoBERTa	0%	T-B-DEMO	0	1	0	0	0	0	1
		T-B-LOC	0	10	2	0	0	1	15
		T-B-PER	0	0	14	0	0	0	19
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	2	0	0	1	0	10
		T-I-PER	0	0	0	0	0	5	5
		T-O	0	5	7	0	0	2	442
	30%	T-B-DEMO	0	1	0	0	0	0	1
		T-B-LOC	0	10	2	0	0	1	15
		T-B-PER	0	0	16	0	0	4	13
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	1	0	0	3	0	9
		T-I-PER	0	0	1	0	0	6	3
		T-O	0	13	5	0	4	2	432
	50%	T-B-DEMO	0	1	0	0	0	0	1
		T-B-LOC	0	9	1	0	0	0	18
		T-B-PER	0	1	17	0	0	0	15
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	1	0	0	6	0	6
		T-I-PER	0	0	2	0	0	6	2
		T-O	0	8	3	0	4	1	440
	100%	T-B-DEMO	0	1	0	0	0	0	1
		T-B-LOC	0	9	2	0	0	0	17
		T-B-PER	0	1	22	0	0	1	9
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	0	0	6	0	7
		T-I-PER	0	0	2	0	0	5	3
		T-O	0	5	8	0	2	2	439

		P-DEMO	P-LOC	P-PER	P-O
CNN-based NER	T-P-DEMO	0	0	0	2
	T-P-LOC	0	4	0	16
	T-P-PER	0	0	15	25
	T-P-O	0	25	35	0

Table 15: Confusion matrices for the out-of-domain evaluation on a *Manuscript preserved in the Greek Orthodox Church of St. Georg in Vienna* (dated from 1778)

Model	Synth.	True Label	P-B-DEMO	P-B-LOC	P-B-PER	P-I-DEMO	P-I-LOC	P-I-PER	P-O
Old-Serb-BERTić	0%	T-B-DEMO	0	0	0	0	0	0	0
		T-B-LOC	0	1	0	0	0	0	11
		T-B-PER	0	2	13	0	0	0	4
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	1	2	0	1	0	8
		T-I-PER	0	1	3	0	0	6	1
		T-O	0	9	5	0	5	1	261
	30%	T-B-DEMO	0	0	0	0	0	0	0
		T-B-LOC	0	6	0	0	0	0	6
		T-B-PER	0	2	10	0	0	0	7
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	1	0	2	0	9
		T-I-PER	0	1	4	0	0	5	1
		T-O	0	12	9	0	8	1	251
	50%	T-B-DEMO	0	0	0	0	0	0	0
		T-B-LOC	0	1	0	0	0	0	11
		T-B-PER	0	1	7	0	0	0	11
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	1	0	3	0	8
		T-I-PER	0	1	5	0	0	1	4
		T-O	0	5	1	0	4	0	271
	100%	T-B-DEMO	0	0	0	0	0	0	0
		T-B-LOC	0	8	0	0	0	0	4
		T-B-PER	0	1	9	0	0	0	9
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	2	0	0	3	0	7
		T-I-PER	0	0	3	0	0	3	5
		T-O	0	21	1	0	13	0	246
Old-Serb-XLM-RoBERTa	0%	T-B-DEMO	0	0	0	0	0	0	0
		T-B-LOC	0	3	0	0	0	0	9
		T-B-PER	0	0	12	0	0	1	6
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	3	2	0	2	0	5
		T-I-PER	0	0	3	0	0	2	6
		T-O	0	9	3	0	2	1	266
	30%	T-B-DEMO	0	0	0	0	0	0	0
		T-B-LOC	0	6	0	0	0	0	6
		T-B-PER	0	0	16	0	0	1	2
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	1	0	5	0	6
		T-I-PER	0	0	0	0	0	10	1
		T-O	0	12	7	0	0	1	261
	50%	T-B-DEMO	0	0	0	0	0	0	0
		T-B-LOC	0	7	0	0	0	0	5
		T-B-PER	0	0	16	0	0	1	2
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	2	0	2	0	8
		T-I-PER	0	0	0	0	0	11	0
		T-O	0	8	3	0	4	0	266
	100%	T-B-DEMO	0	0	0	0	0	0	0
		T-B-LOC	0	5	0	0	0	0	7
		T-B-PER	0	0	16	0	0	1	2
		T-I-DEMO	0	0	0	0	0	0	0
		T-I-LOC	0	0	4	0	0	4	4
		T-I-PER	0	0	2	0	0	9	0
		T-O	0	5	8	0	2	1	265

		P-DEMO	P-LOC	P-PER	P-O
CNN-based NER	T-P-DEMO	0	0	0	0
	T-P-LOC	0	2	0	7
	T-P-PER	0	0	7	26
	T-P-O	0	7	8	0

Table 16: Confusion matrices for the out-of-domain evaluation on a *Charters of Emperor Stefan Uroš to the people of Dubrovnik* (dated between 1356 and 1362)

Model	Synth.	True Label	P-B-DEMO	P-B-LOC	P-B-PER	P-I-DEMO	P-I-LOC	P-I-PER	P-O
Old-Serb-BERTić	0%	T-B-DEMO	0	2	3	0	0	0	35
		T-B-LOC	0	1	1	0	1	0	57
		T-B-PER	0	0	19	0	0	0	46
		T-I-DEMO	0	0	0	0	0	0	2
		T-I-LOC	0	0	0	0	0	0	1
		T-I-PER	0	0	7	0	0	13	22
		T-O	0	54	46	0	11	2	4876
	30%	T-B-DEMO	0	9	10	0	1	0	20
		T-B-LOC	0	17	5	0	1	0	37
		T-B-PER	0	3	22	0	0	1	39
		T-I-DEMO	0	0	0	0	0	0	2
		T-I-LOC	0	0	0	0	0	0	1
		T-I-PER	0	1	6	0	1	12	22
		T-O	0	111	101	0	17	9	4751
	50%	T-B-DEMO	0	3	0	0	0	0	37
		T-B-LOC	0	9	0	0	1	0	50
		T-B-PER	0	0	15	0	0	0	50
		T-I-DEMO	0	0	0	0	0	0	2
		T-I-LOC	0	0	0	0	0	0	1
		T-I-PER	0	0	9	0	0	7	26
		T-O	0	22	16	0	10	0	4941
	100%	T-B-DEMO	0	8	1	0	1	0	30
		T-B-LOC	0	13	0	0	2	0	45
		T-B-PER	0	2	16	0	0	0	47
		T-I-DEMO	0	0	0	0	0	0	2
		T-I-LOC	0	0	0	0	0	0	1
		T-I-PER	3	1	4	0	1	8	25
		T-O	0	50	12	0	11	0	4916
Old-Serb-XLM-RoBERTa	0%	T-B-DEMO	1	3	6	0	0	0	30
		T-B-LOC	0	18	0	0	0	1	41
		T-B-PER	0	2	25	0	0	0	38
		T-I-DEMO	0	0	0	0	0	0	2
		T-I-LOC	0	0	0	0	0	0	1
		T-I-PER	0	0	10	0	0	19	13
		T-O	0	23	38	0	2	7	4919
	30%	T-B-DEMO	16	3	4	0	0	0	17
		T-B-LOC	3	28	0	0	5	1	23
		T-B-PER	0	4	27	0	0	0	34
		T-I-DEMO	0	0	0	0	0	2	0
		T-I-LOC	0	0	0	0	0	0	1
		T-I-PER	2	0	13	0	0	23	4
		T-O	4	31	28	0	9	4	4913
	50%	T-B-DEMO	12	4	1	0	0	0	23
		T-B-LOC	2	29	0	0	5	1	23
		T-B-PER	0	3	28	0	0	0	34
		T-I-DEMO	0	0	0	0	0	0	2
		T-I-LOC	0	0	0	0	0	0	1
		T-I-PER	0	0	11	0	0	22	9
		T-O	6	18	21	0	15	5	4924
	100%	T-B-DEMO	20	2	4	0	0	0	14
		T-B-LOC	2	31	0	0	7	1	19
		T-B-PER	0	4	31	0	0	2	28
		T-I-DEMO	0	0	0	0	0	1	1
		T-I-LOC	0	0	0	0	0	0	1
		T-I-PER	0	0	13	0	0	22	7
		T-O	9	22	41	0	8	6	4903

		P-DEMO	P-LOC	P-PER	P-O
CNN-based NER	T-P-DEMO	1	0	0	41
	T-P-LOC	0	21	0	40
	T-P-PER	0	0	50	57
	T-P-O	2	229	223	0