From Zero to Hero: Building Serbian NER from Rules to LLMs

Milica Ikonić Nešić
University of Belgrade, Faculty of Philology,
Serbia
milica.ikonic.nesic@fil.bg.ac.rs

Saša Petalinkar
University of Belgrade, Serbia
sasa5linkar@gmail.com

Ranka Stanković
University of Belgrade,
Faculty of Mining and Geology, Serbia
ranka.stankovic@rgf.bg.ac.rs

Ruslan Mitkov
University of Alicante, Spain
ruslan.mitkov@ua.es

Abstract

Named Entity Recognition (NER) presents specific challenges in Serbian, a morphologically rich language. To address these challenges, a comparative evaluation of distinct model paradigms across diverse text genres was conducted. A rule-based system (SrpNER), a traditional deep learning model (Convolutional Neural Network – CNN), fine-tuned transformer architectures (Jerteh and Tesla), and Large Language Models (LLMs), specifically ChatGPT 4.0 Nano and 4.1 Mini, were evaluated and compared. For the LLMs, a one-shot prompt engineering approach was employed, using prompt instructions aligned with the entity type definitions used in the manual annotation guidelines. Evaluation was performed on three Serbian datasets representing varied domains: newspaper articles, history textbook excerpts, and a sample of literary texts from the srpELTeC collection. The highest performance was consistently achieved by the fine-tuned transformer models, with F1 scores ranging from 0.78 on newspaper articles to 0.96 on primary school history textbook sample.

1 Introduction

The task of Named Entity Recognition (NER) involves identifying and classifying key information, such as persons, locations, organisations, dates, and other specific entities, within unstructured text (Krstev et al., 2014; Frontini et al., 2020). Accurate NER is crucial for numerous downstream NLP applications, including information extraction (Feng et al., 2022), question answering (Mollá et al., 2006; Verma et al., 2023), and machine translation (Sulistyo et al., 2025). Historically, NER systems have evolved through various approaches, ranging from rule-based methods to those leveraging machine learning and deep learning. Rule-based systems, exemplified by SrpNER for the Serbian language (Krstev et al., 2014), which utilised extensive lexical resources (Krstev, 2008; Vitas and Krstev, 2012) and local grammars, demonstrated high efficiency, particularly on specific text types like news articles (achieving an F1 score of approximately 96% on newspaper texts). However, their development necessitates significant linguistic expertise, and adapting them to new classes, domains or languages can be resource-intensive.

Moving beyond rule-based systems, machine learning and deep learning approaches, including models like Conditional Random Fields, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs), became prevalent. While demonstrating impressive results for high-resource languages with extensive datasets, these models also showed potential for more specific or challenging contexts. For example, CNN architectures have been successfully employed for NER tasks in specific low-resource domains, such as legal text in Turkish (Çetindağ et al., 2023) or historical literary text in Serbian (Šandrih Todorović et al., 2021), achieving competitive performance (F1 scores aprox 91%) on such datasets.

The advent of transformer-based models marked a significant paradigm shift in NLP. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) emerged as a cornerstone model, setting new benchmarks across a wide array of language understanding tasks. Its architecture enables the learning of deep contextualised representations, leading to superior performance in tasks like NER (Zhang and Zhang, 2023). BERT's capability for transfer learning has proven particularly beneficial for low-shot classification tasks (Garrido-Merchan et al., 2023). The success of BERT spurred the development of multilingual models (Wang et al., 2020) and specialised models for various non-English languages. For South Slavonic languages, dedicated models like Bertić (Lju-

Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models associated with RANLP 2025,
pages 87–96, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-102-8-010

bešić and Lauc, 2021), xlm-r-bertić [1] developed by the CLARIN Knowledge Centre for South Slavonic languages (CLASSLA), SRoBERTa (Cvejić, 2022), and XLM-R based models jerteh-355-tesla (Ikonić Nešić et al., 2024) and TESLA-mini (Škorić, 2024) have been developed, demonstrating the effectiveness of the transformer architecture in this linguistic context.

More recently, Large Language Models (LLMs) (Brown et al., 2020) have demonstrated remarkable zero-shot and few-shot abilities across numerous NLP tasks via prompt engineering (Li and Liang, 2021). This approach allows leveraging the vast knowledge within frozen pre-trained LLMs by crafting specific input prompts, appealing greatly to low-resource scenarios as it avoids resource-intensive training. However, applying prompt learning directly to tasks like NER presents unique challenges (Shen et al., 2023). While LLMs excel at tasks aligned with their pre-training objectives (e.g., text generation or "fill-in-the-blank"), NER is fundamentally a sequence labelling task requiring precise identification of entity spans and types (Ma and Hovy, 2016). Early prompt-based NER approaches, such as span-orientated methods that enumerate all potential spans (Cui et al., 2021) or type-oriented methods that query for specific entity types (Liu et al., 2022), often required multiple inference rounds or relied on complex, hand-crafted prompt templates, limiting their efficiency and practical applicability (Shen et al., 2023). This inherent mismatch means that despite the general capabilities of LLMs, achieving robust and accurate NER performance via prompt engineering is still an active area of research and often requires careful prompt design or specialised techniques.

This paper presents a comparative analysis of the performance of different generations of models applied to NER in Serbian. The comparison of rule based, traditional deep learning approaches, represented by a trained CNN model, with two fine-tuned BERT models is presented. Furthermore, the potential of leveraging contemporary LLMs for Serbian NER through prompt engineering, utilising the capabilities of the ChatGPT 4.0 mini and ChatGPT 4.0 nano models was investigated. By evaluating and comparing these diverse modelling paradigms which is a major contribution of the this study, we aim to provide insights into their relative strengths, weaknesses, and applicability for Serbian NER, contributing to the understanding of model evolution and resource efficiency in this field.

2 Related Work

While NER has a long research history, its comparative evaluation across model generations remains unevenly distributed across languages. In particular, Serbian has seen a few dedicated surveys or systematic comparisons only. The earliest and most relevant work is by Vitas and Pavlović-Lažetić (2008), who provided a general overview of NER methods and linguistic resources for Serbian, including rule-based systems. Since then, most research has focused on domain-specific applications or evaluated individual systems. For instance, Sandrih et al. (2019) examined NER systems for Serbian personal names, while Todorović et al. (2021) developed models for recognising entities in 19th-century Serbian literature. More recently, Živković et al. (2022) assessed transformer-based models in the clinical domain. However, none of these works offer a broad, comparative evaluation across diverse model paradigms, nor do they address resource-efficiency concerns across domains.

In contrast, surveys on NER for English are abundant and continuously updated. Well-cited foundational works such as Nadeau and Sekine (2007), Marrero et al. (2013), and Shaalan (2014) laid the groundwork. More recently, deep learning–focused surveys like Li et al. (2020), Keraghel et al. (2024), and Warto et al. (2024) have provided extensive reviews of neural and transformer-based NER systems. Domain-specific surveys also exist, such as Ehrmann et al. (2023) for historical texts and Jehangir et al. (2023) for biomedical and multilingual NER. This disparity further motivates our study, which addresses a clear gap in the literature for Serbian and offers a multi-paradigm evaluation from rule-based through deep learning to LLM-based NER systems.

The study by (Affi and Latiri, 2022) addresses NER for the Arabic language, highlighting the challenges posed by its complex morphology which often necessitates extensive handcrafted feature engineering. To overcome this limitation, the authors proposed a novel deep neural network architecture combining CNN, LSTM, and BERT embeddings to generate rich word representations without re-

[1] Classla/xlm-r-bertic · Hugging Face. (2023, December 18). https://huggingface.co/classla/xlm-r-bertic

lying on external knowledge or handcrafted features. Their approach achieved state-of-the-art results on the ANERCorp dataset, with F1-scores of 93.34% and 93.68% using bidirectional LSTM-CRF (BLC) and bidirectional GRU-CRF (BGC) architectures, respectively. This work is relevant to our study as it demonstrates the effectiveness of advanced deep learning architectures (integrating embeddings from models like BERT with sequence models like LSTM/GRU and CRF) for NER, even for morphologically rich languages. Furthermore, it implicitly includes a comparison of the performance between related architectures (LSTM vs. GRU) within their proposed framework, which aligns with our goal of comparing different model types.

Shelar et al. (2020) conduct a comparative analysis of different existing libraries and tools for NER, including Python's spaCy, Apache OpenNLP, and TensorFlow. The comparison was based on key performance metrics such as training accuracy, F-score, prediction time, model size, and ease of training, using the same dataset across all evaluated tools. A key finding was that Python's spaCy generally achieved higher accuracy and better overall results compared to the other tools tested. This paper is highly relevant to our research as it serves as a direct example of a comparative study of different NER systems or implementations. Its methodology of using standard performance metrics to evaluate distinct tools provides a valuable template and context for our own comparison of various NER models, even if our focus might be more on the underlying model architectures rather than solely the libraries used.

The research presented in (Ikonić Nešić et al., 2024) investigates NER for the Serbian, focusing on the integration of BERT models with the spaCy library. The paper presents a comparison of different architectures and techniques for preparing NER models, trained to recognise seven entity types on a diverse Serbian dataset. Specifically, the authors explored various configurations and training pipelines within the spaCy framework, as well as the impact of different BERT versions (varying architectures, sizes, and pre-training corpora containing Serbian). The goal was to evaluate the trade-offs between model complexity and performance. This research is relevant as it addresses NER for a specific language (Serbian), which is the focus of our study. Most importantly, the paper explicitly conducts a comparison of different configurations and variations of a powerful NER approach (BERT+spaCy), analysing their impact on performance, which relates to our objective of comparing different NER models or different configurations/implementations.

The current study provides a comprehensive comparative analysis of NER models for the Serbian, spanning rule-based systems, traditional machine learning approaches, modern deep learning architectures, and LLMs. In contrast to prior work, which has typically focused on specific domains or isolated model types, our evaluation is conducted across multiple real-world text genres including historical textbooks, news articles, and literary prose, allowing for a robust assessment of model generalisation and domain adaptability. By systematically benchmarking diverse NER paradigms on both seen and unseen data, we offer novel insights into the strengths and limitations of each approach, thereby contributing to the advancement of NER in low-resource and morphologically rich languages.

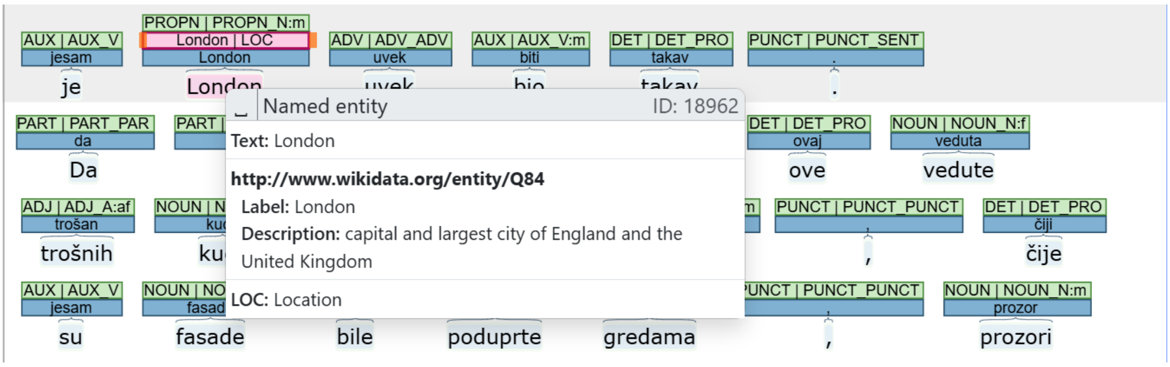
3 Methodology

This section outlines the dataset preparation process for model training, the training of CNN, two BERT-based models, as well as the one-shot prompting approach applied to LLMs.

3.1 Data Preparation

The preparation of the training dataset has been ongoing for an extended period and constitutes a part of the TESLA-NER-NEL corpus (Ikonić Nešić and Utvić, 2024), which, in its final version, will contain 150,000 sentences annotated with named entities linked to Wikidata entries, as well as part-of-speech (POS) tags and lemmatisation.

The training dataset was compiled through two distinct annotation strategies: a semi-automated procedure (*srTESLA-SA*) and a fully automated one (*srTESLA-FA*). Within the semi-automated workflow, a total of 53,417 sentences were initially labelled automatically using SrpNER (Krstev et al., 2014) and jerteh-355-tesla (Ikonić Nešić et al., 2024) model. For manual correction of pre-annotated dataset, INCEpTION tool (Figure 1) was used. The sentences were post-annotated by multiple trained annotators, and all annotations were cross-checked by an expert. These sentences were selected from (1) novels from SrpELTeC (Stanković et al., 2024) and SrpKor (Vitas et al., 2024)



Named entity ID: 18962

Text: London

http://www.wikidata.org/entity/Q84

Label: London

Description: capital and largest city of England and the United Kingdom

LOC: Location

Figure 1: An example of annotation in INCEpTION

(23,273 sentences), (2) newspapers from SrpKor (8,737 sentences), (3) legal documents from Intera (Stanković et al., 2017) (19,383 sentences) and (4) wikipedia from srELEXIS (Krstev et al., 2024) (2024 sentences).

The fully automated approach relied on two techniques: the first utilised sentence templates and structured lexical resources, including gazetteers such as Leximirka (Stanković et al., 2018; Lazic and Škoric, 2019), to generate annotated examples (*srTESLA-lex*); the second employed ChatGPT 4.0 for automatic annotation generation (*srTESLA-chat*). This approach provided context-rich sentences, facilitating disambiguation for NEL task, with 20,076 sentences in total.

The named entity tagset used in this study is aligned with categories commonly applied in the annotation of literary and historical texts, such as those developed within the European Literary Text Collection (ELTeC) (Stanković et al., 2024; Frontini et al., 2020). It includes the following entity types: personal names (PERS), geographical locations (LOC), organizations (ORG), professional roles and titles (ROLE), demonyms (DEMO), cultural and artistic works (WORK), and events (EVENT). Among these, locations (LOC) are the most frequent, with approximately 36,654 instances, followed by personal names (PERS) with 13,636, and organization names (ORG) with around 11,060 occurrences (Table 1).

3.2 NER Modelling Approaches and Configuration

The configuration and implementation of the NER task across different modelling approaches is presented in this subsection.

One-Shot LLM-based NER with `spacy-llm`

To explore the capabilities of modern LLMs for Serbian NER, we employed a one-shot learning strategy facilitated by the `spacy-llm` [2] library. This approach avoids resource-intensive fine-tuning by leveraging the model's existing knowledge through carefully crafted prompts.

Recognising that prompt performance is often enhanced in multilingual contexts and by providing concrete examples, we designed a custom prompt template using the `Jinja` [3] templating language. This allowed for a dynamic and structured input for the LLMs. The core of our prompt is designed to instruct the model to act as an expert in NER and to identify entities within a given text according to a specified set of categories.

The `Jinja` template is structured as follows:

```
Vi ste stručnjak za prepoznavanje
    imenovanih entiteta (NER).
Vaš zadatak je da primite tekst i iz
    njega izdvojite imenovane entitete.
Svaki entitet mora pripadati jednoj od
    sledećih kategorija:
    {{ ', '.join(labels) }}.
Ako neki deo teksta nije entitet, označ
    ite ga kao: `==NONE==`.

{%- if label_definitions %}
Ispod su definicije svake kategorije
    koje će vam pomoći da tačno
prepoznate vrste imenovanih entiteta.
{%- endif %}
...
Pasus: {{ text }}
Odgovor:
```

The NER task was configured within the `spacy-llm` framework by defining the set of entity labels and providing their detailed description specified below.

[2] https://spacy.io/usage/large-language-models
[3] https://jinja.palletsprojects.com/en/stable

Class	Description for annotators	Count
LOC	Names of continents, countries, populated places, geographic features, celestial bodies, etc.	36,654
ROLE	Professional titles, functions, or social roles, such as doctor, director, king, or teacher.	15,170
PERS	Personal names of individuals, including given names, surnames, and aliases of real or fictional figures.	13,636
ORG	Names of institutions, companies, political bodies, schools, hospitals, and other formal organizations.	11,060
DEMO	Demonyms indicating origin, nationality, or ethnic background, including adjectival forms derived from locations.	7,559
WORK	Titles of creative works such as books, poems, artworks, theatrical plays, and periodicals.	3,319
EVENT	Specific historical or recurring events such as wars, revolutions, natural disasters, or commemorations.	464

Table 1: Entity types with descriptions and frequency in the dataset.

```
[components.llm.task.label_definitions]
PERS = "Vlastita imena stvarnih ili izmi
    šljenih pojedinaca  li  čna imena,
    prezimena, nadimci, bogovi, sveci i
    imenovane životinje."
ROLE = "Zanimanja, činovi, titule i
    funkcije koje ljudi obavljaju, sa
    ili bez ličnog imena; uključuje viš
    erečna zvanja."
DEMO = "Nazivi naroda, etničkih grupa i
    stanovnika mesta, kao i pridevi
    izvedeni iz geografskih imena."
ORG = "Imena organizacija, institucija i
     udruženja: kompanije, partije, š
    kole, muzeji, kafane, crkve,
    sportski klubovi "
LOC = "Vlastita imena geografskih
    lokacija: kontinenti, države,
    regioni, gradovi, sela, planine,
    reke, jezera, ulice, trgovi."
WORK = "Naslovi umetničkih i kulturnih
    dela: knjige, pesme, filmovi, slike,
     skulpture, spomenici, novine, video
    -igre."
EVENT = "Nazivi događaja: praznici,
    revolucije, ratovi, bitke,
    demonstracije, festivali, sportski
    događaji, prirodne katastrofe."
```

For these experiments, two specific models were employed: gpt-4.1-mini and gpt-4.1-nano. These models were tasked with performing NER on our evaluation datasets using the described one-shot prompting configuration.

CNN and Fine-Tuned Transformer Models

In addition to the LLM-based prompting method, we trained a CNN and two transformer-based models to serve as comparative baselines. These experiments were conducted within the spaCy framework, making use of the core library for the CNN and the spacy-transformers extension for the transformer models.

The **CNN model** was configured using spaCy's standard multi-layer tok2vec architecture. This component generates context-sensitive token vectors which are then passed to the Named Entity Recognition (ner) layer for classification.

The two **transformer models** leverage the spacy-transformers library to integrate pretrained language models into the spaCy pipeline. The transformer's contextual word embeddings are fed into the ner component. The specific pretrained models used as a base for fine-tuning were:

- te-sla/TeslaXLM[4] model is derived from the large multilingual architecture, FacebookAI/xlm-roberta-large, having been further fine-tuned for the nuances of Serbian and Serbo-Croatian. Comprising 561 million parameters, its adaptation involved training on a substantial 20-billion-token corpus encompassing both Latin and Cyrillic scripts commonly used in Serbian. This comprehensive fine-tuning process results in a model that demonstrates high proficiency and robustness across varying scripts and dialectal forms (Škorić and Petalinkar, 2024).

- jerteh/Jerteh-355[5] is based on the RoBERTa-large architecture but was trained from scratch exclusively on a monolingual Serbian corpus of 4 billion tokens. With 355 million parameters, this model is specifically tailored to generate high-quality, context-aware embeddings optimized for the Serbian language environment (Škorić, 2024).

To ensure a fair comparison, all three models were trained on the same dataset for a total of 10 epochs for transformers and 5 epochs for CNN.

[4] https://huggingface.co/te-sla/TeslaXLM

[5] https://huggingface.co/jerteh/Jerteh-355

4 Evaluation Results and Discussion

The trained models were evaluated for their performance on the test set, as shown in Table 2. To assess performance across different text types, the models were evaluated on data from three distinct sources: **newspaper articles** from *Politika* newspapers (841 sentences), a sample from corpus of history textbook for elementary school named **srHistory** (331 sentences), and a sample of three novels from the srpELTeC collection (SRP19070[6], SRP19121[7], SRP19180[8]) (Table 3) named **srpELTeC sample** (544 sentences), were the literary texts used for evaluation were explicitly excluded from the training set. The distribution of named entities across these datasets is shown in Figure 2.

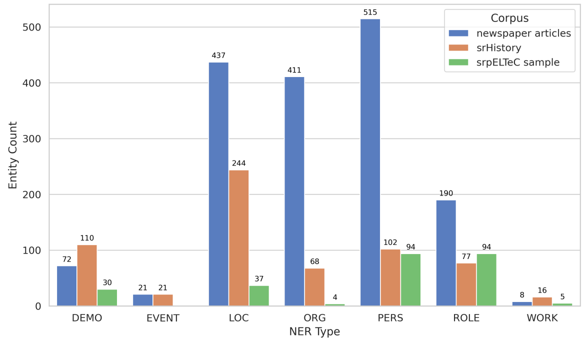


Figure 2: Distribution of NE types across corpora

The models included in the comparison represent distinct paradigms: the rule-based `SrpNER`, a traditional `CNN`, fine-tuned BERT models (`Jerteh` and `Tesla`), and prompt-based LLMs (`ChatGPT 4.0 Nano` and `ChatGPT 4.1 Mini`). As anticipated, the fine-tuned transformer models `Jerteh` and `Tesla` generally achieved the highest overall F1 scores across the evaluation sets (Table 3). Their strong performance stems from their ability to learn complex contextual patterns from the large and diverse training dataset as TESLA-NER-NEL and generalise these patterns. The rule-based SrpNER system demonstrated robustness, performing strongly on domains it was specifically designed for, such as newspaper articles, and maintaining solid performance on other domains due to its reliance on linguistic rules and lexicons, although it exhibits less flexibility than machine learning models when encountering entirely new patterns and types (e.g. EVENT). The `CNN`, a traditional deep learning approach, proved more susceptible to domain and style shifts; while capable of learning effective local features, it is less adept at capturing long-range context compared to transformers, which can hinder performance on complex sentences or subtle entity mentions. The prompt-based `LLMs` approach show a notable decrease in performance on the more challenging or domain-shifted datasets (Table 3). A key aspect of LLM evaluation approach involved providing the `ChatGPT` models with detailed instructions and definitions for each entity type directly in the prompt. These instructions were identical to those provided to human annotators who created the gold standard dataset used for training and evaluation. This consistency ensures that both the LLMs and the supervised models are attempting to solve the exact same NER task definition. By leveraging the LLMs' strong instruction-following capabilities with the annotation guidelines, we aimed to facilitate a direct and fair comparison between the performance achieved via prompt engineering and that of models explicitly trained on data annotated according to those same guidelines. Despite this, the inherent nature of prompt-based generation, as opposed to fine-tuned sequence labelling, appears less optimal for achieving high precision and recall on this specific task without further adaptation, as detailed by their class-level results (Table 4).

Analysing performance by dataset and entity type reveals the impact of textual characteristics (Table 4). The *srHistory* dataset has very strong results from BERT models (`Tesla` F1 0.958, `Jerteh` F1 0.884 overall), and exceptional per-class performance for `Tesla`, achieving F1 scores of 0.94 or higher for most entity types, including PERS, LOC, ORG, DEMO, and EVENT (Table 4b). This high performance, combined with Tesla's low FP count on this dataset (Table 3), suggests the well known named entities and clear style of a textbook, with is well suited to its capabilities. `SrpNER` and `CNN` also showed solid per-class results on this dataset, outperforming the `LLMs` on many entity types. On the *newspapers* dataset, `Tesla`, `Jerteh` and `SrpNER` models performed well on common entity types like PERS, LOC, and ORG (Table 4a), likely because this domain aligns closely with the newspaper portion of the training data. `SrpNER` showed particularly high precision for these classes, consistent with its rule-based na-

[6] Jelena Dimitrijević, Fati-Sultan (ELTeC edition)

[7] Veljko M. Milićević, *Bespuće* (Wasteland ELTeC edition)

[8] Milica Jankovic, *Pre sreće* (Before happiness ELTeC edition)

Table 2: Precision, recall, and F1 for each entity type and model on a test dataset, including entity counts.

Class	Count	Tesla			Jerteh			CNN		
		P	R	F1	P	R	F1	P	R	F1
PERS	7,471	0.972	0.972	**0.972**	0.955	0.976	0.965	0.918	0.890	0.904
LOC	2,910	0.966	0.975	**0.971**	0.966	0.973	0.970	0.940	0.944	0.942
ROLE	2,719	0.844	0.820	**0.832**	0.825	0.837	0.831	0.795	0.756	0.775
DEMO	2,053	0.934	0.966	**0.950**	0.931	0.959	0.945	0.903	0.902	0.902
ORG	1,458	0.817	0.817	**0.817**	0.807	0.802	0.804	0.708	0.742	0.725
WORK	699	0.698	0.645	0.671	0.659	0.718	**0.687**	0.582	0.476	0.524
EVENT	100	0.736	0.670	**0.702**	0.685	0.630	0.656	0.786	0.330	0.465

Table 3: Summary of NER model performance across three distinct datasets

Dataset	Total Entities	Model	TP	FP	FN	P	R	F1
Newspaper articles	1,654	SrpNER	965	140	689	0.873	0.583	0.699
		CNN	834	613	820	0.576	0.504	0.538
		Jerteh	1,339	563	315	0.704	0.810	0.753
		Tesla	1,375	476	279	0.743	0.831	**0.785**
		Chat 4.0 Nano	1,027	735	627	0.583	0.621	0.601
		Chat 4.1 Mini	1295	944	359	0.578	0.783	0.665
srHistory	638	SrpNER	472	117	166	0.801	0.740	0.769
		CNN	487	98	151	0.833	0.763	0.796
		Jerteh	559	68	79	0.892	0.876	0.884
		Tesla	601	15	37	0.976	0.942	**0.958**
		Chat 4.0 Nano	329	251	309	0.567	0.516	0.540
		Chat 4.1 Mini	540	253	98	0.681	0.846	0.755
sprELTeC Sample	264	SrpNER	209	54	55	0.795	0.799	0.793
		CNN	98	56	166	0.636	0.371	0.469
		Jerteh	163	58	101	0.738	0.617	0.672
		Tesla	200	29	64	0.873	0.758	**0.811**
		Chat 4.0 Nano	135	338	129	0.285	0.511	0.366
		Chat 4.1 Mini	221	179	43	0.553	0.837	0.666

ture, though with lower recall on some rarer types (e.g., WORK, EVENT).

In contrast, the *srpELTeC sample* dataset presented the greatest challenge for most models, leading to a significant overall performance drop, particularly for the CNN (F1 0.469) and ChatGPT 4.0 Nano (F1 0.366). This is primarily attributable to a substantial domain and style shift (19th-century literary language vs. modern training data). A specific instance of the style shift impacting LLMs was observed with the term *Arnautin*. This archaic and historical term, originating from Turkish, was not recognised by the ChatGPT 4.1 Nano model, while other models successfully identified it, illustrating the potential sensitivity of LLMs to lexical and historical variations not prominent in their pre-training. The class-level results for ELTeC (Table 4c) clearly show this difficulty across multiple entity types for these models. For example, ChatGPT 4.0 Nano exhibited very low precision across several common classes (PERS, LOC, DEMO). The CNN also showed low F1 scores on most classes. Interestingly, the rule-based SrpNER demonstrated relatively more stable per-class performance on some types (e.g., ROLE F1 0.80) compared to some data-driven models (Jerteh ROLE F1 0.54, ChatGPT 4.1 Mini ROLE F1 0.60), suggesting its rules were less affected by stylistic nuances than statistical patterns learned by CNN or LLMs. Transformer models (Tesla F1 0.811, Jerteh F1 0.672 overall), while still the best performers on this challenging set, showed reduced per-class scores compared to other datasets for types like ORG, ROLE, and WORK. The ChatGPT 4.1 Mini model on *srpELTeC sample* showed a pattern of higher recall but lower precision compared to Tesla for some classes (e.g., PERS, LOC, DEMO), indicating it retrieved more potential entities but with more false positives.

5 Conclusion/Future Work

In this study, a comparative evaluation of diverse NER model paradigms for Serbian was conducted across distinct text genres: newspa-

Table 4: Evaluation Results by Entity Type (Precision / Recall / F1)

(a) newspaper articles

Class	SrpNER			CNN			Jerteh			Tesla			Nano			Mini		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PERS	0.92	0.67	0.77	0.76	0.55	0.64	0.89	0.92	0.90	0.91	0.91	0.91	0.65	0.78	0.71	0.89	0.87	0.88
LOC	0.85	0.86	0.86	0.67	0.67	0.67	0.83	0.87	0.85	0.86	0.90	**0.88**	0.65	0.78	0.71	0.84	0.78	0.81
ORG	0.89	0.36	0.51	0.46	0.28	0.35	0.76	0.70	0.73	0.75	0.75	**0.75**	0.60	0.46	0.52	0.58	0.73	0.65
ROLE	0.78	0.47	**0.59**	0.35	0.51	0.42	0.31	0.62	0.41	0.38	0.63	0.48	0.45	0.36	0.40	0.34	0.63	0.46
WORK	1.00	0.13	0.22	0.31	0.50	**0.38**	0.23	0.38	0.29	0.13	0.13	0.13	0.02	0.13	0.03	0.61	0.38	0.11
DEMO	1.00	0.07	0.13	0.46	0.56	0.50	0.58	0.93	0.71	0.58	0.94	**0.72**	0.29	0.31	0.30	0.20	0.96	0.33
EVENT	0.00	0.00	0.00	0.14	0.05	0.07	0.37	0.48	0.42	0.52	0.67	**0.58**	0.27	0.14	0.19	0.18	0.62	0.28

(b) srHistory

Class	SrpNER			CNN			Jerteh			Tesla			Nano			Mini		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PERS	0.77	0.83	0.80	0.85	0.89	0.87	0.99	0.98	0.99	1.00	1.00	**1.00**	0.53	0.62	0.57	0.98	0.98	0.98
LOC	0.85	0.90	0.87	0.87	0.89	0.88	0.93	0.89	0.91	0.97	0.97	**0.97**	0.71	0.76	0.73	0.89	0.90	0.89
ORG	0.61	0.34	0.43	0.61	0.37	0.46	0.80	0.77	0.78	0.94	0.94	**0.94**	0.54	0.38	0.45	0.70	0.88	0.78
ROLE	0.69	0.56	0.62	0.76	0.65	0.70	0.84	0.86	0.85	0.99	0.86	**0.92**	0.50	0.10	0.17	0.62	0.49	0.55
WORK	0.00	0.00	0.00	1.00	0.06	0.12	0.29	0.31	0.30	1.00	0.50	**0.67**	0.08	0.25	0.13	0.47	0.50	0.48
DEMO	0.86	0.82	0.84	0.86	0.88	0.87	0.95	0.94	0.94	0.98	0.98	**0.98**	0.54	0.36	0.43	0.46	0.95	0.62
EVENT	0.71	0.57	0.63	0.86	0.29	0.43	0.73	0.76	0.74	1.00	0.81	**0.89**	0.17	0.10	0.12	0.19	0.48	0.27

(c) srpELTeC Sample (Literature)

Class	SrpNER			CNN			Jerteh			Tesla			Nano			Mini		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PERS	0.65	0.73	0.69	0.72	0.40	0.52	0.65	0.72	0.69	0.82	0.91	**0.86**	0.27	0.87	0.41	0.62	0.91	0.74
LOC	0.87	0.70	0.78	0.63	0.60	0.61	0.73	0.87	0.79	0.81	0.95	**0.88**	0.29	0.81	0.43	0.64	0.97	0.77
ORG	0.00	0.00	0.00	0.30	0.75	0.43	0.25	0.25	0.25	1.00	0.50	**0.67**	0.00	0.00	0.00	0.20	0.25	0.22
ROLE	0.77	0.83	0.80	0.61	0.20	0.30	0.88	0.39	0.54	0.96	0.54	**0.69**	0.83	0.21	0.34	0.51	0.73	0.60
WORK	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.60	0.60	1.00	0.40	**0.57**	0.10	0.40	0.16	0.36	0.80	0.50
DEMO	1.00	0.80	0.89	0.67	0.53	0.59	1.00	0.73	0.85	1.00	0.80	**0.89**	0.06	0.03	0.04	0.52	0.83	0.64

per articles, history textbook excerpts, and a literary sample. Evaluated models included a rule-based system (SrpNER), a CNN, fine-tuned transformers (Jerteh, Tesla), and prompt-based LLMs (ChatGPT 4.o Nano and ChatGPT 4.1 Mini). Fine-tuned BERT based models generally achieved the highest performance, demonstrating strong generalisation from a diverse training corpus, with Tesla showing exceptionally high results on the history data. The rule-based SrpNER proved robust, performing well on news and showing resilience to stylistic shifts in literary texts. The CNN was more susceptible to domain variations. Prompt-based LLMs exhibited lower performance for precise NER, particularly on the challenging literary dataset, suggesting limitations of prompting alone for complex sequence labelling tasks despite using human annotation guidelines. This analysis highlights the critical influence of both model architecture and target domain characteristics on NER performance in Serbian.

Based on these findings, our future research will focus on enhancing LLM performance for Serbian NER through refined prompting strategies (e.g., few-shot, PEFT) and exploring their potential in hybrid systems. Addressing the challenges of domain and style shifts, notably for historical/literary texts, is also crucial, potentially via dedicated domain adaptation techniques or advanced hybrid approaches. Further evaluation on a broader spectrum of Serbian text types, including lower-resource domains, is warranted. Finally, conducting detailed qualitative error analysis and exploring few-shot learning paradigms within supervised frameworks are valuable avenues for improving NER performance and reducing annotation effort in new domains.

Acknowledgments

This work was supported by the Science Fund of the Republic of Serbia (#7276, Text Embeddings – Serbian Language Applications, TESLA) and partially by the CIDEXG/2023/12 project, funded by the Generalitat Valenciana.

References

Manel Affi and Chiraz Latiri. 2022. Arabic named entity recognition using variant deep neural network architectures and combinatorial feature embedding based on cnn, lstm and bert. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 302–312.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language models are few-shot learners.

Can Çetindağ, Berkay Yazıcıoğlu, and Aykut Koç. 2023. Named-entity recognition in turkish legal texts. *Natural Language Engineering*, 29(3):615–642.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the ACL: ACL-IJCNLP 2021*, pages 1835–1845, Online. ACL.

Andrija Cvejić. 2022. Prepoznavanje imenovanih entiteta u sprskom jeziku pomoću transformer arhitekture. *Zbornik radova Fakulteta tehničkih nauka u Novom Sadu*, 37(02):310–315.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Matteo Romanello, and Amir Zeldes. 2023. Named entity recognition and classification in historical document collections. *Natural Language Engineering*.

Xin Feng, Yingrui Li, Zhang Hang, Zhang Fan, Qiong Yu, and Ruihao Xin. 2022. Tbr-ner: Research on covid-19 text information extraction based on joint learning of topic recognition and named entity recognition. *Journal of Sensors*, 2022(1):3967171.

Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. Named entity recognition for distant reading in ELTeC. In *CLARIN Annual Conference 2020*.

Eduardo C Garrido-Merchan, Roberto Gozalo-Brizuela, and Santiago Gonzalez-Carvajal. 2023. Comparing bert against traditional machine learning models in text classification. *Journal of Computational and Cognitive Engineering*, 2(4):352–356.

Milica Ikonić Nešić, Sasšа Petalinkar, Stanković Ranka, and Škorić Mihailo. 2024. BERT downstream task analysis: Named Entity Recognition in Serbian. In *14th International Conference on Information Society and Technology – ICIST 2024*. unpublished.

Milica Ikonić Nešić and Miloš Utvić. 2024. Overview of the Tesla-Ner-Nel-Gold Dataset: Showcase on Serbian-English Parallel Corpus. *Technical editors*, page 57.

Asim Jehangir, Muhammad Asad Aslam, et al. 2023. A comprehensive survey on named entity recognition: Recent advances and challenges. *Artificial Intelligence Review*.

Achraf Keraghel, Amine Abdaoui, and Abdelghani Bouramoul. 2024. Recent advances in named entity recognition: A survey. *Information Processing & Management*.

Cvetana Krstev. 2008. *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade.

Cvetana Krstev, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24(2):473–489.

Cvetana Krstev, Ranka Stanković, Aleksandra M. Marković, and Teodora Sofija Mihajlov. 2024. Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 106–114, Torino, Italia. ELRA and ICCL.

Biljana Lazic and Mihailo Škoric. 2019. From DELA based dictionary to Leximirka lexical database. *Infotheca–Journal for Digital Humanities*, 19(2):00–00. `https://doi.org/10.18485/infotheca.2019.19.2.4`.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. ACL.

Yanan Li, Wenjie Li, Bing Qin, and Ting Liu. 2020. A survey of deep learning approaches for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Andy T. Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. 2022. Qaner: Prompting question answering models for few-shot named entity recognition.

Nikola Ljubešić and Davor Lauc. 2021. Berti\'c–the transformer language model for bosnian, croatian, montenegrin and serbian. *arXiv preprint arXiv:2104.09243*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. A. Moreiro. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Australasian Language Technology Association Workshop*, pages 51–58. Australasian Language Technology Association.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Ivana Sandrih, Cvetana Krstev, and Duško Vitas. 2019. A hybrid method for serbian personal name recognition in historical literary texts. In *RANLP*.

Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510.

Hemlata Shelar, Gagandeep Kaur, Neha Heda, and Poorva Agrawal. 2020. Named entity recognition approaches and their comparison for custom ner model. *Science & Technology Libraries*, 39(3):324–337.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. Promptner: Prompt locating and typing for named entity recognition.

Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. Electronic dictionaries–from file system to lemon based lexical database. In *6th Workshop on Linked Data in Linguistic (LDL-2018), Towards Linguistic Data Science*.

Ranka Stanković, Cvetana Krstev, and Duško Vitas. 2024. SrpELTeC: A Serbian Literary Corpus for Distant Reading. *Primerjalna književnost*, 47(2).

Ranka Stanković, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. 2017. Keyword-based search on bilingual digital libraries. In *Semantic Keyword-Based Search on Structured Data Sources: COST Action IC1302 Second International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania*, pages 112–123. Springer.

Danang Arbian Sulistyo, Didik Dwi Prasetya, Fadhli Almu'iini Ahda, and Aji Prasetya Wibawa. 2025. Pivoted low resource multilingual translation with ner optimization. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(5):1–16.

Branislava Šandrih Todorović, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. Serbian NER& Beyond: The Archaic and the Modern Intertwinned. In *Deep Learning Natural Language Processing Methods and Applications – Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1252–1260.

Jovana Todorović, Cvetana Krstev, and Duško Vitas. 2021. Ner in serbian novels from the 19th century using contextual embeddings. In *INFOTEH*.

Devika Verma, Ramprasad S. Joshi, Aiman A. Shivani, and Rohan D. Gupta. 2023. Kāraka-based answer retrieval for question answering in Indic languages. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1216–1224, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, 63:279–292.

Duško Vitas and Gordana Pavlović-Lažetić. 2008. Resources and methods for named entity recognition in serbian. *Infotheca*.

Duško Vitas, Ranka Stanković, and Cvetana Krstev. 2024. The Many Faces of SrpKor. In *South Slavic Languages in the Digital Environment JuDig Book of Abstracts*, Belgrade, Serbia. University of Belgrade - Faculty of Philology.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Arif Warto, Septi Handayani, et al. 2024. A systematic literature review on named entity recognition research (2011–2020). *Journal of Theoretical and Applied Information Technology*.

Yuzhe Zhang and Hong Zhang. 2023. Finbert–mrc: financial named entity recognition using bert under the machine reading comprehension paradigm. *Neural Processing Letters*, 55(6):7393–7413.

Mihailo Škorić. 2024. New Language Models for Serbian. *Infotheca – Journal for Digital Humanities*, 24(1). `https://doi.org/10.18485/2024.24.1.1`.

Mihailo Škorić and Saša Petalinkar. 2024. New XLM-R-based Language Models for Serbian and Serbo-Croatian. In *Artificial Intelligence Conference*, Belgrade. SASA.

Marko Živković, Milan Samardžić, and Ranka Stanković. 2022. Clinical named entity recognition in serbian using bert and ensemble learning. *Biomedical Signal Processing and Control*, 72.