applied sciences

MDPI

Article

Improving Text Recognition Accuracy for Serbian Legal Documents Using BERT

Miloš Bogdanović *, Milena Frtunić Gligorijević, Jelena Kocić and Leonid Stoimenov

Faculty of Electronic Engineering, University of Nis, 18000 Nis, Serbia; milena.frtunic.gligorijevic@elfak.ni.ac.rs (M.F.G.); tosic.jelena@elfak.rs (J.K.); leonid.stoimenov@elfak.ni.ac.rs (L.S.)

* Correspondence: milos.bogdanovic@elfak.ni.ac.rs

Abstract: Producing a new high-quality text corpus is a big challenge due to the required complexity and labor expenses. High-quality datasets, considered a prerequisite for many supervised machine learning algorithms, are often only available in very limited quantities. This in turn limits the capabilities of many advanced technologies when used in a specific field of research and development. This is also the case for the Serbian language, which is considered low-resourced in digitized language resources. In this paper, we address this issue for the Serbian language through a novel approach for generating high-quality text corpora by improving text recognition accuracy for scanned documents belonging to Serbian legal heritage. Our approach integrates three different components to provide high-quality results: a BERT-based large language model built specifically for Serbian legal texts, a high-quality open-source optical character recognition (OCR) model, and a word-level similarity measure for Serbian Cyrillic developed for this research and used for generating necessary correction suggestions. This approach was evaluated manually using scanned legal documents sampled from three different epochs between the years 1970 and 2002 with more than 14,500 test cases. We demonstrate that our approach can correct up to 88% of terms inaccurately extracted by the OCR model in the case of Serbian legal texts.

Keywords: BERT; text recognition; optical character recognition; word similarity

check for updates

Academic Editors: Stefan Fischer and Douglas O'Shaughnessy

Received: 29 November 2024
Revised: 26 December 2024
Accepted: 9 January 2025
Published: 10 January 2025

Citation: Bogdanović, M.; Frtunić Gligorijević, M.; Kocić, J.; Stoimenov, L. Improving Text Recognition Accuracy for Serbian Legal Documents Using BERT. *Appl. Sci.* **2025**, *15*, 615. https://doi.org/10.3390/app15020615

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Today, a vast amount of information is stored in numerous documents and is accessible online in both structured and unstructured forms. These data come from various domains and can be examined and processed to obtain the relevant information for a given task. A significant amount of work and time is required for the manual processing and analysis of a repository consisting of documents often varying in size, structure, and layout. The content of the documents, in the form of raw text, is the essence of the building of tools based on recent advances in artificial intelligence (AI), which demands automated information extraction, processing, and analysis.

Some of the most significant advances in natural language processing (NLP) research, where machine-based techniques and systems are created to automatically extract and analyze texts from conventional humanities, can be found in interdisciplinary studies such as in digital humanities. Due to a limited number of annotated datasets, which are a prerequisite for many supervised machine learning algorithms, many advanced technologies, including neural networks, are still underutilized in digital humanities when compared to general text analysis [1]. For many digital humanities study fields, producing new corpora is a big challenge due to the required complexity and labor expenses [2]. This

task becomes even harder in cases where the language used within the documents is in the position of being low-resourced in the field of digital infrastructure and digitized language resources, such as the Serbian language [3].

The production of new corpora for a single domain starts with a repository constituting a single domain which often contains documents varying in size, but also in structure and layout. Nevertheless, domain experts are capable of extracting the desired information, but this process tends to be time-consuming and error-prone. The domain of legislation is an example of such a domain. Due to the nature of the legal domain, legal documents are often elaborated in terms of structure from a general document and can be too long to analyze and understand. Furthermore, they appear in different forms, such as legal contracts, law commission reports, tribunals, judgments, different acts, etc. Thus, information extraction from legal documents is highly desired [4–6], and it is foreseen to be significant for the development of AI tools in this domain.

Deep neural networks have proven their potential in a range of document-related tasks such as semantic search and document retrieval, question answering, and document generation. The quality and performance of models built using deep neural networks depend on the representation and the quality of the text used for training. If a neural network model is built for a particular field and purpose, it will become more efficient and accurate if it is provided with high-quality text during its training process. Thus, the existence of high-quality text corpora is a prerequisite for building precise and efficient models in any field, including the legal domain [7].

Although it has plenty of resources, the Serbian language is considered to be low-resourced in the field of digitized language resources. Our focus in this research was the development of a BERT-augmented system named LexBERT capable of generating high-quality text corpora by improving text recognition accuracy for scanned documents belonging to Serbian legal heritage. We consider the improved text recognition procedure that we present in this paper to be a fundamental prerequisite to the development of reliable AI tools in the Serbian language legal domain. LexBERT focuses on a problem previously described within the scientific community [8,9] and emerged from a general approach we define and describe in this paper. To demonstrate the usability of our approach, it was implemented for a specific domain—the legal domain. Thus, LexBERT emerged as a combination of three major components, two of which were specially developed for this system:

- SrBERTa v2—BERT-based language model for Serbian legal texts written in Cyrillic;
- Optical character recognition engine—Tesseract OCR;
- Word-level similarity measure for Serbian Cyrillic used to compare OCR results and SrBERTa v2 suggestions in cases of lower OCR accuracy.

The major contribution of this paper is within the approach, its components, and its overall capability. Our approach envisions the development of a BERT-based model for each specific domain that it is applied within. According to the reported OCR error detection confidence [9], we take advantage of this information at the word level to optimize and enhance our approach. Compared to similar approaches [8], we consider our approach to be simpler and more efficient, and it can achieve excellent results when focused on a specific domain.

In a demonstration of the previously described advantages, LexBERT autocorrects the results of the optical character recognition process applied upon scanned legal documents written in Serbian Cyrillic and generate high-quality legal text corpora for the Serbian language. To do so, LexBERT relies on a BERT-based deep neural network model we have developed specially for these purposes, named SrBERTa v2. LexBERT detects lower OCR accuracy and generates correction suggestions through the means of SrBERTa v2. Generated

corrections are compared to OCR output using a similarity measure we have developed for Serbian Cyrillic letters. We present validation results obtained using publicly available scanned legal document corpora gathered from the Official Gazette of the Republic of Serbia [10], originating between the years 1970 and 2002.

The rest of this paper is organized as follows. In Section 2, we present the state-of-the-art large language models and the possibility of combining them with OCR capabilities. Section 3 contains a detailed description of the approach we present in this paper. It includes a description of the process of generating legal text corpora for the Serbian language along with a description of the key components: a BERT-based language model for Serbian legal texts, a word-level similarity measure, and OCR engine usage. An evaluation of the approach is presented in Section 4 which demonstrates the capabilities of our approach determined based upon publicly available scanned legal documents containing Serbian legislation. In Section 5, we conclude with an outlook on enhancements we will implement in future research and development.

2. Related Work

2.1. Large Language Models—State of the Art and Position in Text Recognition Task

Representation learning techniques offer a chance to bridge the gap between text and vectors. Because character and word embeddings may embed discrete texts into continuous vector space, they are important for a variety of NLP tasks. Many representation learning approaches based on embedding techniques have proven their quality in different downstream tasks [11–14].

The development of a transformer architecture [15] with self-attention mechanisms—BERT [16]—marked the beginning of significant advancements in artificial intelligence. Hajiali et al. [8] proposed a method to detect OCR errors, using the BERT language model and FastText sub-word embeddings, and managed to achieve 70.9% precision when generating correction candidates. Numerous efforts were influenced by BERT's effectiveness in pre-training tasks on large-scale unlabeled corpora. New architectures and models inspired by transformers, including GPT-2 [17] and BART [18], began to appear and demonstrate their effectiveness when using their general-purpose semantic features for a range of natural language processing applications. Various pre-trained language models (PLMs) found their way into many different fields in NLP [17,19]. The scientific community persisted in experimenting with bigger models by increasing the volume of data and the size of the model. When the number of parameters increased to more than 10 billion, the models began to behave and perform differently from the smaller models. In contrast to the 175-billion-parameter GPT-3 and the 540-billion-parameter PaLM, the 330-million-parameter BERT and the 1.5-billion-parameter GPT-2 were shown to be less capable of in-context learning and exhibited surprisingly strong conversational skills. The research community began addressing language models as large language models (LLMs) as a result of the emergence of these models.

Dedicated PLMs have also been trained for the legal domain. PLMs developed for the legal sector offer a more capable foundational system for legal operations. As an example, Zhong et al. [20] suggest a language model pre-trained on Chinese court papers, such as those from civil and criminal cases, to address this problem. Bogdanovic et al. presented SrBERTa [21]—a language model designed to understand the formal language of Serbian legal documents, trained using Cyrillic legal texts contained within a dataset created specifically for this purpose.

2.2. Optical Character Recognition

The goal of optical character recognition research is to create an AI-based model that can automatically extract and process text from text-containing documents. Text-containing documents of interest vary from handwritten text to printed or scanned text images, and their processing is performed with the goal of transforming them into an editable digital format for deeper and further processing. The extraction process is not a trivial one, since it encounters some major obstacles such as the font characteristics of the characters within documents and the quality of images. Unfortunately, OCR technology is still not as sophisticated as human ability.

The quality of input documents directly affects OCR's efficiency and accuracy. Consequently, the information extraction process tends to be noisy and often suffers from blur effects, faded text, scanning quality, and wrinkles. When scanned documents are used, the low accuracy is caused by these noises. Preprocessing operations including binarization and contrast enhancement may occasionally be necessary to improve the scanned documents' visual quality. For OCR to perform at its best, a high degree of accuracy and minimal processing latency are necessary [22].

To improve OCR accuracy, various approaches have been considered and investigated. OCR engine performance analysis was reported in [23], while the significance of OCR accuracy was emphasized in [24]. In the legal domain, OCR has in many cases been used for named entity recognition, coupled with machine learning techniques. This coupling of technologies includes a rule-based approach for data extraction from court decisions [25], CRFs (Conditional Random Fields), and BiLSTM (Bidirectional Long Short-Term Memory) [26]. In many cases, the Tesseract OCR engine was used for different scanned documents ranging from small scanned bill documents to sample images [27–29]. The accuracy of Tesseract measured in these cases was based on string matching and varied from 90% to 97%. In addition to string matching, a study reported by Ramdhani et al. [30] uses conversion time, NER (named entity recognition) time, string match accuracy for precision, and a recall measure for the number of acquired entities to measure OCR engine effectiveness—not only for Tesseract OCR but also for the Foxit and PDF2GO OCR engines. Although the differences between the engines studied in [30] are not very large, Tesseract OCR has still shown the highest accuracy level and F1-Score.

The previously described state of OCR engines has led us to decide to rely on Tesseract OCR for the research we present in this paper. Although there are OCR adjustments for the Serbian language [3,31], these adjustments usually refer to restoring texts written using Serbian Cyrillic or Latin with diacritics, which is not the case in our research. Cyrillic is not unique to the Serbian language; thus, we have decided to rely on the capabilities of the engine that has proven its usability—Tesseract OCR.

3. A BERT-Augmented Text Recognition Approach

Building precise and efficient models in any field requires high-quality text corpora. As mentioned in the Introduction of this paper, the Serbian language is low-resourced in the field of digital infrastructure and digitized language resources, especially in Serbian legal heritage documents. Therefore, within this research, we present LexBERT, an approach for generating high-quality legal text corpora by improving text recognition accuracy for Serbian legal heritage documents. LexBERT is designed to recognize text from scanned legal documents written in the Serbian Cyrillic language and return high-quality text corpora. The whole system, presented in Figure 1, combines three main components:

- Tesseract OCR—optical character recognition engine;
- SrBERTa v2—BERT-based language model for Serbian legal texts;
- Word-level similarity measure.

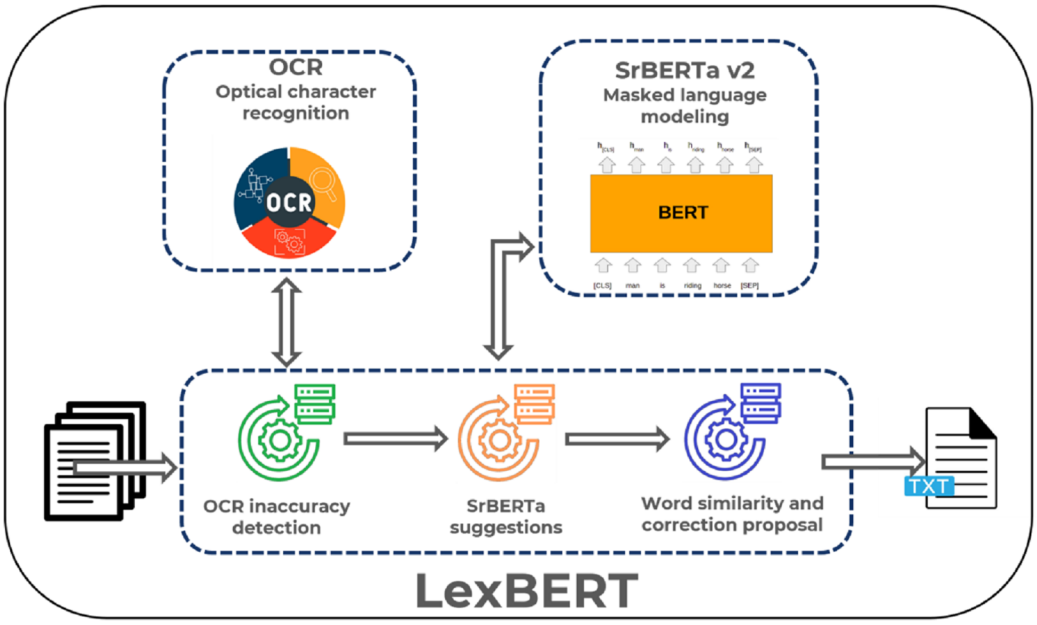


Figure 1. LexBERT architecture—a BERT-augmented text recognition approach.

The LexBERT system receives an image of the document as input in the system and extracts the text using the OCR engine. Within our approach, we rely on the Tesseract OCR engine for the Serbian Cyrillic language for extracting the text from an image. The OCR engine returns all the recognized words coupled with information about the word's position in the image, position in the block, paragraph, and line, and its confidence level given in a percentage. Based on the obtained results and information about the block, paragraph, and line, the system reconstructs the text from the image into paragraphs and executes the merging of words that have been split across new lines due to their length.

Since OCR accuracy may vary, the core of the LexBERT system is implemented as an autocorrection mechanism. The diagram for this autocorrection mechanism is presented in Figure 2. While reconstructing the text, for each of the recognized words, the system analyzes the confidence level and evaluates whether it is necessary to apply the autocorrection mechanism for that word. The LexBERT system is set to verify and autocorrect each word with an OCR confidence level lower than 90% and each split word that was merged by the system. The threshold of 90% was chosen after a manual check of the OCR output for several pages, during which it was noticed that for words with confidence levels lower than 90%, OCR tended to make errors.

This mechanism uses a masked language modeling approach to create suggestions for a particular place in a paragraph. The system uses the developed SrBERTa v2, a BERT-based model for Serbian legal texts, explained in detail in the next section of this paper. For each word that needs to be verified, the system acquires the top 20 suggestions for that word by passing the context that contains the paragraph where that word is masked into the SrBERTa v2 model. Although more than one word in a paragraph can be a candidate for verification, the suggestions are acquired for each word separately. This way, the system obtains a list of the top 20 possible mask values ranked according to score, for the verification and autocorrection of each word.

After receiving suggestions from the model, the system calculates the similarities between suggested tokens and the word recognized by OCR and proposes the final value for that position in the paragraph. Since our approach prioritizes words' content and tries to find an exact match for the one in the text, our goal is to determine the suggested token that is most similar to the OCR output with or without the expected punctuation marks.

Therefore, the system relies on the Levenshtein distance for similarity calculation between terms and uses it to compare tokens with original OCR output and OCR output trimmed from all non-alphanumeric characters at the beginning and at the end. Furthermore, before the Levenshtein distance is calculated with two combinations of OCR output, each token is trimmed from space characters at the beginning and at the end. Afterward, the similarity is calculated with the trimmed token and the trimmed token with the first character converted to lowercase. The final similarity is the minimal Levenshtein distance calculated for all combinations of the token and the OCR output.

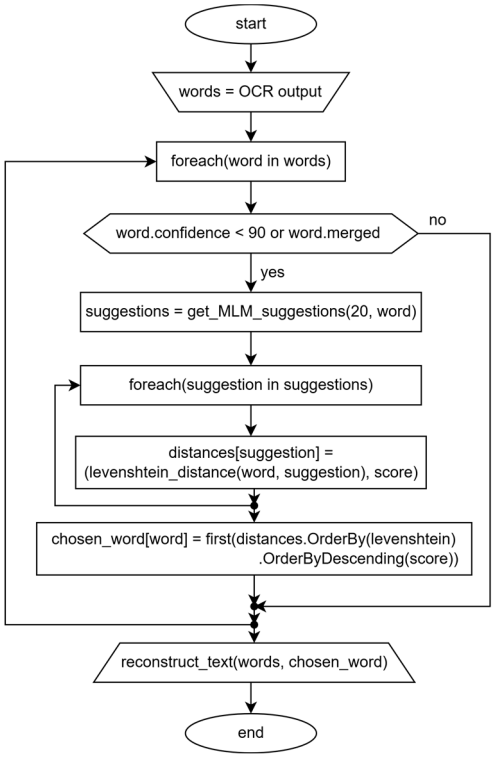


Figure 2. Diagram for autocorrection mechanism.

Next, after all similarities are calculated, the final suggestion for a particular place in the paragraph is determined based on the calculated Levenshtein distances for all proposed tokens and their scores returned by the model. The highest priority is given to the token with the lowest Levenshtein distances. If only one token has the lowest value for Levenshtein distances, that token is suggested. However, if more than one token has the same low value, the decision is made based on the score value returned by the model, and the token with a higher score is suggested to be placed at the given position in the paragraph. Finally, the suggestions are inserted in the appropriate places in the paragraphs, and thus, recreated text is passed to the output of the system.

3.1. SrBERTa v2—BERT-Based Language Model for Serbian Legal Texts Written in Cyrillic

As part of this research, we present a follow-up study on the SrBERTa model [21]. The primary goal of SrBERTa was to develop a model capable of understanding the formal language used in Serbian legislation. Initially, SrBERTa was trained to comprehend the Serbian language and then fine-tuned for masked language modeling using Cyrillic legal texts from a specially created dataset [32]. Masked language modeling is a downstream task which predicts a masked token in a sequence, and the model performing the task can attend to tokens bidirectionally. This task predicts only the masked words rather than reconstructing the entire input [16].

In this second phase of model development, our objective was to enhance the previously achieved results by using the following approaches:

- Creating an expanded dataset of legal texts;
- Training a more extensive tokenizer for the Serbian language, optimized for better tokenization of legislative terms;
- Training a larger model on the task of masked language modeling using an increased amount of training data, including a new, legal dataset.

For the purpose of training the newer version of the SrBERTa network, we utilized two new datasets. The primary dataset selected for the training phase was the OSCAR [33] dataset, which represents a comprehensive collection of open data, generated through linguistic classification from the Common Crawl corpus [34]. Additionally, we utilized a more recent version of OSCAR, specifically OSCAR 23.01, which includes a greater volume of Serbian language data, comprising 1,677,896 documents and 632,781,822 words and occupying 7.7 GB of memory.

The second dataset used within this research was created specifically for this purpose and comprised only legal texts. These texts were gathered from the Legal Information System of the Republic of Serbia, resulting in a total of 1.6 GB of data. Each legal text underwent preprocessing in order to minimize data loss due to truncation, with the maximum input sequence size set to 512 tokens. Consequently, all texts were split into smaller units of 512 tokens and then concatenated using the newline character. This preprocessing ensured the generation of inputs that could be optimally utilized by the network.

For the tokenizer training phase, we utilized the Byte-Pair Encoding (BPE) tokenization algorithm provided by the HuggingFace library [35]. Our goal was to train a network specialized in modeling not only the Serbian language but also Serbian legislative language. Therefore, we adopted a novel training approach in comparison to the previous version of the SrBERTa network.

Initially, we used a tokenizer trained only on texts collected from the Internet (OSCAR dataset). However, this approach had limitations in tokenizing domain-specific terms found in legal texts. Consequently, during the masked token prediction process, the proposed words often did not represent the best possible options. This issue arose because the network, despite being fine-tuned on legal texts, did not tokenize many legal terms as complete words but rather as sub-words, due to the fact that the tokenizer was not trained using legal texts. To address the previously described issue, we devised a new training approach for the tokenizer. Now, during the tokenizer training phase, we select an equal number of legal texts and texts from the Internet. This strategy aims to create a tokenizer capable of effectively handling both standard spoken language and the specialized terminology used in legal texts. Moreover, in the new version of SrBERTa, we incorporate a larger vocabulary, consisting of 50,256 tokens. In the end, we prove that the network trained with this kind of domain-adapted vocabulary is able to encode more legal terms as whole words and effectively utilize them in masked language modeling tasks.

The configuration of the newly trained SrBERTa tokenizer is shown in Table 1.

Table 1. SrBERTa tokenizer configuration.

Vocabulary Size	Minimum Frequency	Special Tokens
50,256	2	<s> </s> <pad> <unk> <mask>

The following example shows the output obtained when we apply the improved version of the SrBERTa tokenizer to a random input sample, selected from the legal corpus

generated out of publicly available legal texts from the Legal Information System of the Republic of Serbia:

Example input (in Serbian): "Чланом правилника регулисано је да вршилац техничке контроле појединог дела пројекта за грађевинску дозволу потврђује исправност тог дела пројекта тако што, на посебној страници даје изјаву уз навођење података о правном лицу, односно предузетнику који је извршио техничку контролу".

Example input translated into English: "The article of the rulebook stipulates that the person performing the technical control of a particular part of the project for the building permit confirms the correctness of that part of the project by making a statement on a separate page with information about the legal entity, that is, the entrepreneur who performed the technical control".

The main methods of any tokenizer are tokenizing (splitting strings in sub-word token strings), converting token strings to ids and back, encoding/decoding (tokenizing and converting to integers), adding new tokens to the vocabulary in a way that is independent of the underlying structure, and managing special tokens. The output of a tokenization process is a tensor of IDs belonging to the tokenizer vocabulary generated for a given input sentence. Let us observe the generated tokenized output for the previously shown input sample, shown using two Pytorch tensors: one for the input IDs and the other for the attention masks. The first tensor of the input stores corresponding token IDs from the dictionary and the second stores a mask, which is used by the network to avoid padding token indices.

'input_ids': tensor([[0, 5429, 304, 6295, 14584, 313, 330, 37782, 6904, 3942, 35052, 1190, 3569, 350, 12057, 6564, 7764, 30902, 1883, 1190, 3569, 1068, 627, 16, 339, 18757, 6738, 3409, 7946, 827, 11355, 1718, 280, 4413, 2963, 16, 786, 36924, 472, 313, 4726, 18525, 5813, 18, 2]]),

'attention_mask': tensor([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]])

In the end, to compare the results, we loaded the same input into the previous version of the tokenizer (publicly available on HuggingFace Hub [36]) and measured the length of the generated tokenized output. The results show that the previous version produced a tensor with a length of 53, whereas the newer version of the tokenizer generated only 45 tokens for the same input sample. It is worth noting that the input sentence can consist of at least 44 tokens, including words, punctuation, start-of-sequence, and end-of-sequence tokens. This indicates that our new approach allows for more legal terms to be tokenized as whole words than was the case before.

3.2. SrBERTa v2 Training and Evaluation

Given that our ultimate goal was to enhance the performance of the SrBERTa model, the next logical step was to train a new version of the model, from scratch, using the new tokenizer and a larger embedding layer, now consisting of 50,256 vectors. The primary objective during this part of the training process was to use a larger network architecture and train the model to understand the Serbian language as well as the language used in Serbian legislation. As was the case before, we used a BERT-type neural network architecture, but now specifically a RoBERTa base, which is available as a ready-made model architecture at HuggingFace [37]. The SrBERTa-base model architecture can be summarized as follows: the network begins with an embedding layer comprising word-, position-, and token-type embeddings, followed by normalization and dropout layers connected to 12 encoder layers, with each of them containing a self-attention mechanism, and outputs results to a final language modeling head. The language modeling head

maps previously generated contextualized embeddings to the size of the vocabulary, while simultaneously, it generates a score for each possible token in the vocabulary for each position in the input sequence.

Another crucial aspect was configuring and setting the training parameters for the new version of the SrBERTa network. During this process, we compared the previously used hyperparameters with the recommended values from the original RoBERTa paper, selecting valid value ranges that aligned with our requirements and hardware capabilities. Notably, for this development phase, we utilized enhanced GPU hardware in comparison to the previous SrBERTa version, specifically the NVIDIA QUADRO RTX 5000, which allowed us to employ a network size twice as large, corresponding to the RoBERTa-base architecture.

We continued to use the AdamW optimizer to update the network weights, due to its improved weight decay methods, with the weight decay parameter set to 0.01. Additionally, the mini-batch size was set to 8 because of hardware constraints, but we effectively increased it to 64 using the gradient accumulation method.

Table 2 shows the most important configuration parameters of the SrBERTa v2 model in comparison to the previous version.

Table 2. Comparison of SrBERTa versions.

	SrBERTa v1	**SrBERTa v2**
Vocabulary size	30 K	50 K
Hidden layers	6	12
Attention heads	12	12
Hidden size	768	768
Mini-batch size	8	64
Training epochs	19	45

Additionally, during the training phase, we adopted a similar approach for setting the learning rate to that described in the RoBERTa paper. Specifically, the learning rate gradually increased over the first 600 steps to a peak value of 1×10^{-4}, after which it was linearly decayed.

Given that the task in question was masked language modeling, the SrBERTa network was implemented as a HuggingFace RobertaForMaskedLM class, which utilizes three types of input tensors: the token ID tensor, the attention mask tensor, and the label tensor. Initially, for the purpose of creating token ID tensors, we applied masks to 15% of the tokens in each input sequence. It is important to note that the loss calculation was performed only on the masked tokens, while the others were ignored.

The initial stage of training the SrBERTa v2 network was conducted over 45 epochs and took a total of 31 days on an NVIDIA QUADRO RTX 5000 GPU. For this phase, we utilized 90% of the data from the OSCAR dataset and 90% of the data from the Serbian legal corpus, which we previously created. The rationale was to pre-train the network to understand natural language while also incorporating legal terminology since the vocabulary now included many legal terms as well. By using this approach, we ensured that all embedding layer vectors were trained and not just those for tokens appearing in the OSCAR dataset, which is usually the case during the pre-training process. Figure 3 depicts the previously described training process of the SrBERTa v2 model.

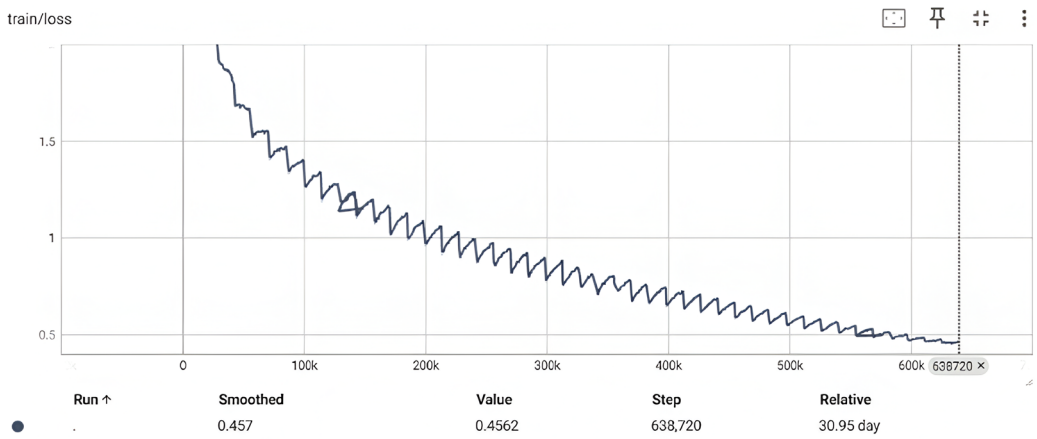


Figure 3. SrBERTa-base training graph—the value of the loss function across 45 epochs.

During the network evaluation phase, we utilized 10% of the input data previously extracted from the legal dataset, as our focus was solely on assessing the model's domain-specific masked language modeling capabilities. The model's performance was quantified using a modified accuracy metric, implemented as follows:

- For each masked word, the top_k predictions generated by the network were considered, with k = 10;
- Only those predictions that exactly matched the true label of the masked word were considered correct.

After the previously described adapted accuracy metric was applied to the test dataset, the SrBERTa v2 model achieved a score of 90.114%, thereby concluding this phase of development.

4. Evaluation

The approach presented in this paper was evaluated using scanned legal document corpora gathered from the Official Gazette of the Republic of Serbia. In order to evaluate the model in different domains, we chose issues that do not contain pure legal provisions, but legal provisions from various specific domains. Therefore, we used 99 scanned pages from six issues published in three different epochs between 1970 and 2002. In order to evaluate this approach for different qualities of scanned documents, we chose two issues from each epoch.

The worst scan quality was in issues published in the oldest epoch. Within this epoch, we chose 14 pages from an issue originating from the year 1970 that contained the Law on procedure before the Constitutional Court, the Law on Educational Inspection, the Law on realization of special social interest in the field of communal activities, Amendments to the law on republican administration, and a Rulebook on the technical inspection of vehicles. The second issue from this epoch was published in 1983, and we chose 20 pages that contained text regarding the Resolution on socio-economic development and economic policy.

For the second epoch, we chose 35 pages from two issues originating from the year 1990. These issues had a better scan quality and contained text regarding the Law on social activities, Rulebook on reimbursement of costs in court proceedings, Law on public information, Law on cultural heritage, Law on library activity, and Law on cultural financing funds.

The best scan quality was for 30 pages selected from two issues published in 2001. The selected pages contained text of the rules of procedure and regulations regarding

the National Assembly of the Republic of Serbia, excise taxes, tax administration, public enterprises, and children's social care.

The evaluation was focused on the reconstruction of words written in the Serbian Cyrillic language with low-level accuracy returned by the OCR engine and merged words. Within this evaluation, the system was not tested for the reconstruction of enumeration, numbers, and words written in all capital letters. For each test page, the test examples for evaluation were created based on the output received from the OCR engine. Consequently, we selected 14,604 test examples for all three epochs that were verified manually. The distribution of the number of examples by epoch is shown in Table 3.

Table 3. The structure of test examples.

Issues	Number of Pages	Total Number of Examples	Distinct Number of Paragraphs	Total Number of Words
2001	30	2569	984	24,624
1990	35	5498	1297	28,568
1975, 1983	34	6537	799	25,935

It can be noticed that all three epochs had a similar total number of words. However, the oldest epoch had the most test cases due to it having the worst scan quality. Furthermore, it was noticed that usually, multiple words needed to be verified within one paragraph. Moreover, the lengths of the paragraphs were different, and there were many cases of contexts that contained only a few words. One example of a scanned paragraph with multiple words that needed verification from an issue published in 1990 is presented in Figure 4. In this paragraph, all words framed in green had lower accuracy, and words underlined in orange had to be merged and verified because of the merge. The context returned by OCR for this paragraph is presented in Figure 5. It can be noticed that only one word was wrongly recognized by the OCR, with recognition of the word *упране* instead of *управе*. In this paragraph, the model accurately predicted values for seven out of eight tests, giving the incorrect value for the first word in the paragraph.

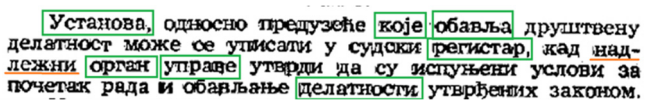

Установа, односно предузеће које обавља друштвену делатност може се уписати у судски регистар, кад надлежни орган управе утврди да су испуњени услови за почетак рада и обављање делатности утврђених законом.

Figure 4. Example of a scanned paragraph, year 1990.

Установа, односно предузеће које обавља друштвену делатност може се уписати у судски регистар, кад надлежни орган упране утврди да су испуњени услови за почетак рада и обављање делатности утврђених законом.

Figure 5. Merged OCR output for scanned paragraph in Figure 4.

For the purpose of this evaluation, we defined an accuracy measure that was used for manual verification. The accuracy measure considered a test case to be successful only if the output was the same word that matched the expected value in gender, number, and case. Based on this accuracy measure, the success rate was calculated for all test pages. The summary evaluation results per test issue, containing average success rates and the highest success rates per issue, are presented in Table 4. The obtained evaluation results differed between epochs. The best results were obtained from the issues published in 2001, and the average success rate for all pages in this epoch was 71.49%. The pages published in 1990 had an average success rate for all pages of 67.7%, and the lowest average success

rate was 59.8% for all pages published in 1975 and 1983 together. The highest average success rate per issue was for an issue published in 2001 with 18 selected pages, with this value being 74.96%. When analyzed individually, seven pages had a success rate higher than 80%, and the highest value per page was 88.3%. These results are very promising and confirm that our approach corrects most of the OCR inaccuracies. At this point, we would like to emphasize that a manual analysis was performed for all situations where the approach did not provide an appropriate correction proposal. The analysis results and the reasons for incorrect correction proposals are presented in the rest of this section. This additional analysis provides a deeper understanding of the potential improvements that we also present.

Table 4. Summarized evaluation results.

Issue	Average Success Rate	Highest Success Rate
2001 test issue 1	74.96	88.31
2001 test issue 2	66.29	79.24
1990 test issue 1	67.34	82.14
1990 test issue 2	69.22	77.55
1982 test issue	53.85	66.37
1973 test issue	68.26	76.16

During the analysis of the results, it was noticed that of all the tests that were not matched, in 41% of tests, the model suggested a synonym to the expected word, a correct word expressed in a different form or a different case, or the expected word consisting of multiple tokens, like words that consist of two words separated by a hyphen. Also, it was noticed that within unmatched terms, some of the domain values were repeated multiple times in the same and different form and case. Moreover, 4.7% of all tests were words consisting of two words separated by a hyphen, and in many cases, the model suggested one of the combined words in the adequate form and case. An example of such a case is presented in a scanned paragraph from an issue published in 1975 in Figure 6. The depicted paragraph contains 18 words that needed to be verified (underlined in green) due to the lower accuracy returned by OCR or the merging of words. The corresponding context generated based on the OCR output is presented in Figure 7, and all words that were not correctly recognized are highlighted in yellow.

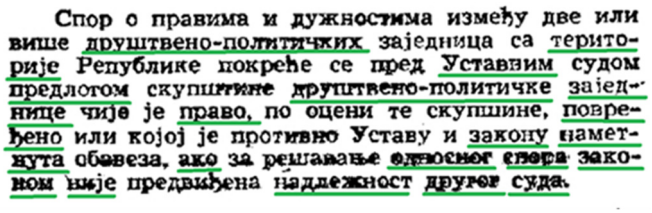

Спор о правима и дужностима између две или више друштвено-политичких заједница са територије Републике покреће се пред Уставним судом предлогом скупштине друштвено-политичке заједнице чије је право, по оцени те скупшине, повређено или којој је противно Уставу и закону наметнута обавеза, ако за решавање односног спора законом није предвиђена надлежност другог суда.

Figure 6. Example of a scanned paragraph, year 1975.

Спор о правима и дужностима између две или више друштвено-полнтичких заједница са територије Републике покреће се пред Уставзим судом предлотом скупштине друштвено-политичке заjед— нице чије је права, по оцени те скупшине, повреЂено или којој је противно Уставу и закону наметнута обавеза, ако за решавање едпоснег свера закоом није предвиђена надлежност друтет суда.

Figure 7. Merged OCR output for scanned paragraph in Figure 6.

Within this paragraph, there were two cases of words consisting of two words separated by a hyphen—*друштвено-политичких* and *друштвено-политичке*—both instances of the usage of an adequate case of the term *socio-political*. Both examples could not be presented with a single token, and for both examples, the model returned the term *social* in the Serbian Cyrillic language in the correct form and case. Furthermore, within this paragraph, there were three more cases in which the model returned incorrect values. However, one of these words could not be represented by a single token, and the model proposed a word with a similar meaning. For the remaining tests, the model returned correct values.

An analysis of the obtained results showed that for some specific domains, the model returned incorrect results more often due to the lack of particular words or their correct form for the specific usage. Additionally, it was noticed that in cases of short contexts, like cases where one item in the enumeration is recognized as a separate paragraph with only a few words, the model tended to return incorrect values for specific domains. For example, take the following two paragraphs:

- "Архив, паред послова из члана 98. овог закона:" (*The archive, in addition to the tasks referred to in Article 98 of this law:*);
- "Музеј, поред послова из члана 98. овог закона:" (*The museum, in addition to the tasks referred to in Article 98 of this law:*).

For the first words in these paragraphs, the model returned more general suggestions in both cases.

Additionally, it was noticed that there were test cases where, due to the lower scan quality, the context returned by the OCR engine had a larger number of incorrectly recognized words close to each other, and in those cases, the model returned incorrect results more often. The example presented in Figure 8 contains a scanned paragraph from an issue published in 1983, and the context in Figure 9 represents a merged output from the OCR engine. It can be noticed that in this example, the OCR engine failed to recognize the year 1984. Furthermore, the words underlined in yellow were completely misrecognized and interpreted as the words highlighted in yellow in Figure 9. Therefore, the created context was incomplete, with some of the original words missing, and from four test cases, only one had the correct output.

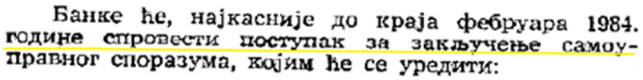

Банке ће, најкасније до краја фебруара 1984. године спровести поступак за закључење самоуправног споразума, којим ће се уредити:

Figure 8. Example of a scanned paragraph, year 1983.

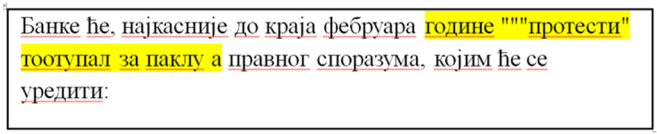

Банке ће, најкасније до краја фебруара године """протести" тоотупал за паклу а правног споразума, којим ће се уредити:

Figure 9. Merged OCR output for scanned paragraph in Figure 8.

5. Conclusions

The results presented in this paper confirm that producing new corpora in any field of digital humanities is a big challenge. Even with the help of tooling and approaches such as the one presented in this paper, producing a new high-quality corpus for low-resourced languages comes with complexities, labor, and obstacles. Nevertheless, we are convinced that the approach we presented in this paper offers a step forward in terms of efficiency and quality when generating a new corpus out of legacy documents written in the Serbian language.

Starting with an image, through text extraction using OCR and word reconstruction augmented with a specially designed BERT-base deep neural network, our approach has shown very promising results. Although it is still quite dependent on OCR text extraction quality and capability, our approach is currently capable of correcting up to 88.3% of incorrectly extracted words in a single image. Furthermore, the evaluation we have conducted proves that there are many situations that we can use to improve accuracy and provide more accurate results. For example, we have determined that of all tests in which the approach did not provide an appropriate correction proposal, in over 41% of these tests, the model we developed suggested a synonym to the expected word, a correct word expressed in a different form or a different case, or the expected word consisting of multiple tokens.

As previously stated, the evaluation we performed offers possibilities for improving our approach. We would like to provide an outlook on the first step we will implement to improve the quality of word reconstruction. The most obvious approach is an improvement in the quality of suggestions given by the SrBERTa model that we developed for these purposes. Currently, SrBERTa efficiency can be improved through the usage of higher-quality datasets originating from different epochs of Serbian language usage. Our model would benefit from different datasets during the initial training phase, but also from the usage of a larger dataset containing Serbian legislation. Moreover, case sensitivity is one of the aspects that should be improved during dataset cleaning and the SrBERTa training process. Furthermore, the SrBERTa architecture is sensitive to the size of the paragraph used during the word reconstruction process, since the paragraph plays the role of providing context for the SrBERTa model. This is a consequence of the RoBERTa architecture we used as a foundation, the masked language modeling problem definition, and the fact that OCR paragraph detection varies in terms of quality and size. By augmenting the size and the quality of the context given to the model, combined with previously stated improvements, we expect the quality of suggestions to improve, giving us the ability to improve reconstruction precision.

Author Contributions: M.B.: Conceptualization, Methodology, Software, Writing; M.F.G.: Software, Data Curation, Methodology, Writing; J.K.: Software, Writing; L.S.: Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia [grant number: 451-03-66/2024-03/200102].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The model presented in this study is openly available at https://huggingface.co/JelenaTosic/SRBerta (accessed on 26 December 2024). The Corpus of Legislation texts of Republic of Serbia 1.0 used within this study is available at https://www.clarin.si/repository/xmlui/handle/11356/1754 (accessed on 26 December 2024). Documents used during the evaluation can be found at https://pravno-informacioni-sistem.rs/reg-overview (accessed on 26 December 2024).

Acknowledgments: The authors would like to thank the Ministry of Science, Technological Development and Innovation of the Republic of Serbia for funding this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, G.; Nulty, P.; Lillis, D. Enhancing Legal Argument Mining with Domain Pre-training and Neural Networks. *arXiv* **2022**, arXiv:2202.13457.

2. Zhang, G.; Lillis, D.; Nulty, P. Can Domain Pre-training Help Interdisciplinary Researchers from Data Annotation Poverty? A Case Study of Legal Argument Mining with BERT-based Transformers. In Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), Silchar, India, 16–19 December 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 121–130.
3. Ilić, V.; Bajčetić, L.; Petrović, S.; Španović, A. SCyDia–OCR for Serbian Cyrillic with Diacritics in Dictionaries and Society. In Proceedings of the XX EURALEX International Congress, Mannheim, Germany, 12–16 July 2022; pp. 387–400.
4. Zadgaonkar, A.V.; Agrawal, A.J. An overview of information extraction techniques for legal document analysis and processing. *Int. J. Electr. Comput. Eng.* **2021**, *11*, 5450–5457. [CrossRef]
5. Turtle, H. Text retrieval in the legal world. *Artif Intell Law* **1995**, *3*, 5–54. [CrossRef]
6. Jain, D.; Borah, M.D.; Biswas, A. Summarization of legal documents: Where are we now and the way forward. *Comput. Sci. Rev.* **2021**, *40*, 100388. [CrossRef]
7. Nguyen, H.-T.; Phi, M.-K.; Ngo, X.-B.; Tran, V.; Nguyen, L.-M.; Tu, M.-P. Attentive deep neural networks for legal document retrieval. *Artif Intell Law* **2024**, *32*, 57–86. [CrossRef]
8. Hajiali, M.; Cacho, J.R.F.; Taghva, K. Generating correction candidates for ocr errors using bert language model and fasttext subword embeddings. In *Intelligent Computing: Proceedings of the 2021 Computing Conference*; Springer International Publishing: Cham, Switzerland, 2022; Volume 1, pp. 1045–1053.
9. Hemmer, A.; Coustaty, M.; Bartolo, N.; Ogier, J.M. Confidence-Aware Document OCR Error Detection. In *Document Analysis Systems*; Sfikas, G., Retsinas, G., Eds.; DAS 2024. Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; Volume 14994. [CrossRef]
10. Official Gazette of the Republic of Serbia. 2024. Available online: https://op.europa.eu/en/web/forum/srbija-serbia (accessed on 26 December 2024).
11. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
12. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jegou, H.; Mikolov, T. Fasttext.zip: Compressing text classification models. *arXiv* **2016**, arXiv:1612.03651.
13. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
14. Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
16. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. 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, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Cedarville, OH, USA, 2019; Volume 1, pp. 4171–4186.
17. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. OpenAI Blog 2019. Available online: https://paperswithcode.com/paper/language-models-are-unsupervised-multitask (accessed on 26 December 2024).
18. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020; pp. 7871–7880.
19. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv* **2019**, arXiv:1906.08237.
20. Zhong, H.; Zhang, Z.; Liu, Z.; Sun, M. *Open Chinese Language Pretrained Model Zoo*; Technical Report; Tsinghua University: Beijing, China, 2019.
21. Bogdanović, M.; Kocić, J.; Stoimenov, L. SRBerta—A Transformer Language Model for Serbian Cyrillic Legal Texts. *Information* **2024**, *15*, 74. [CrossRef]
22. Patel, C.; Patel, A.; Patel, D. Optical character recognition by open source OCR tool tesseract: A case study. *Int. J. Comput. Appl.* **2012**, *55*, 50–56. [CrossRef]
23. Vijayarani, S.; Sakila, A. Performance comparison of OCR tools. *Int. J. UbiComp (IJU)* **2015**, *6*, 19–30.
24. Taghva, K.; Beckleyy, R.; Coombs, J. The effects of OCR error on the extraction of private information. In *Document Analysis Systems VII*; DAS; Bunke, H., Spitz, A.L., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3872.

25. Solihin, F.; Budi, I. Recording of law enforcement based on court decision document using rule-based information extraction. In Proceedings of the International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018, Yogyakarta, Indonesia, 27–28 October 2018.
26. Leitner, E.; Rehm, G.; Moreno-Schneider, J. Fine-grained named entity recognition in legal documents. In Proceedings of the Semantic Systems. The Power of AI and Knowledge Graphs, 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, 9–12 September 2019.
27. Kumar, V.; Kaware, P.; Singh, P.; Sonkusare, R. Extraction of information from bill receipts using optical character recognition. In Proceedings of the International Conference on Smart Electronics and Communication, ICOSEC 2020, Trichy, India, 10–12 September 2020.
28. Akinbade, D.; Ogunde, A.O.; Odim, M.O.; Oguntunde, B.O. An adaptive thresholding algorithm-based optical character recognition system for information extraction in complex images. *J. Comput. Sci.* **2020**, *16*, 784–801. [CrossRef]
29. Harraj, A.E.; Raaissouni, N. OCR accuracy improvement on document images through a novel preprocessing approach. *Signal Image Process. Int. J. (SIPIJ)* **2015**, *6*, 1–18.
30. Ramdhani, T.W.; Budi, I.; Purwandari, B. Optical Character Recognition Engines Performance Comparison in Information Extraction. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 1–8. [CrossRef]
31. Krstev, C.; Stankovic, R.; Vitas, D. Knowledge and rule-based diacritic restoration in Serbian. In Proceedings of the Computational Linguistics in Bulgaria, Third International Conference (CLIB), Sofia, Bulgaria, 28–29 May 2018; pp. 41–51.
32. Bogdanović, M.; Kocić, J. Corpus of Legislation texts of Republic of Serbia 1.0, Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. 2022. Available online: http://hdl.handle.net/11356/1754 (accessed on 26 December 2024).
33. OSCAR Project. Available online: https://oscar-project.org/ (accessed on 26 December 2024).
34. Common Crawl. Available online: https://commoncrawl.org/ (accessed on 26 December 2024).
35. Hugging Face. Summary of the Tokenizers. Available online: https://huggingface.co/docs/transformers/tokenizer_summary#bytepairencoding (accessed on 3 June 2024).
36. Bogdanović, M.; Kocic, J. SRBerta. Available online: https://huggingface.co/JelenaTosic/SRBerta (accessed on 3 June 2024).
37. Hugging Face. RoBERTa Model Documentation. Available online: https://huggingface.co/docs/transformers/model_doc/roberta (accessed on 3 June 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.