computers

MDPI

Article

PharmKE: Knowledge Extraction Platform for Pharmaceutical Texts Using Transfer Learning

Nasi Jofche, Kostadin Mishev, Riste Stojanov *, Milos Jovanovik, Eftim Zdravevski and Dimitar Trajanov

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, 1000 Skopje, North Macedonia
* Correspondence: riste.stojanov@finki.ukim.mk

Abstract: Even though named entity recognition (NER) has seen tremendous development in recent years, some domain-specific use-cases still require tagging of unique entities, which is not well handled by pre-trained models. Solutions based on enhancing pre-trained models or creating new ones are efficient, but creating reliable labeled training for them to learn on is still challenging. In this paper, we introduce PharmKE, a text analysis platform tailored to the pharmaceutical industry that uses deep learning at several stages to perform an in-depth semantic analysis of relevant publications. The proposed methodology is used to produce reliably labeled datasets leveraging cutting-edge transfer learning, which are later used to train models for specific entity labeling tasks. By building models for the well-known text-processing libraries spaCy and AllenNLP, this technique is used to find *Pharmaceutical Organizations* and *Drugs* in texts from the pharmaceutical domain. The PharmKE platform also incorporates the NER findings to resolve co-references of entities and examine the semantic linkages in each phrase, creating a foundation for further text analysis tasks, such as fact extraction and question answering. Additionally, the knowledge graph created by DBpedia Spotlight for a specific pharmaceutical text is expanded using the identified entities. The obtained results with the proposed methodology result in about a 96% F1-score on the NER tasks, which is up to 2% better than those of the fine-tuned BERT and BioBERT models developed using the same dataset. The ultimate benefits of the platform are that pharmaceutical domain specialists may more easily identify the knowledge extracted from the input texts thanks to the platform's visualization of the model findings. Likewise, the proposed techniques can be integrated into mobile and pervasive systems to give patients more relevant and comprehensive information from scanned medication guides. Similarly, it can provide preliminary insights to patients and even medical personnel on whether a drug from a different vendor is compatible with the patient's prescription medication.

Keywords: knowledge extraction; natural language processing; named entity recognition; knowledge graphs; drugs

Citation: Jofche N.; Mishev K.; Stojanov R.; Jovanovik M.; Zdravevski E.; Trajanov D. PharmKE: Knowledge Extraction Platform for Pharmaceutical Texts Using Transfer Learning. *Computers* **2023**, *12*, 17. https://doi.org/10.3390/computers12010017

Academic Editors: Lucia Maddalena, Robertas Damaševičius and Paolo Bellavista

Received: 10 November 2022
Revised: 30 December 2022
Accepted: 5 January 2023
Published: 9 January 2023

1. Introduction

Huge volumes of data are being created regularly, such as content generated on online social media networks and news portals. To facilitate the processing, selection, and comprehension of the huge data volumes that are available, we propose a methodology that uses intelligent knowledge extraction (KE) using natural language processing (NLP) technologies. In this work, we focus on named entity extraction from the pharmaceutical domain, specifically entities that represent *Pharmaceutical Organizations* and *Drugs*. According to [1,2], this NLP job is known as named entity recognition (NER). Its goal is to find entities of a specific kind inside text corpora. NER occupies a prominent position in many NLP systems as a foundational task in information extraction, question answering, and other processes.

Our interest in this subject stems from a challenge we are currently working on with our LinkedDrugs dataset [3], where the manufacturers (*Pharmaceutical Organization*) and active ingredients (*Drug* entities) of the collected drug products can be expressed in varying forms, depending on the data source, country of registration, language, etc. Encouraged by the results of our preliminary research [4], we wish to build on it. Given the ambiguity in entity naming in our drug products dataset, we aimed to greatly enhance the dataset's quality, as well as the outcomes of any downstream analytical work by utilizing NER to normalize these name values for the active components and manufacturers.

Recently, NER accuracy has been improving due to advances in neural network architectures, particularly due to bidirectional long short-term memory (LSTM) networks [5,6], convolutional networks [7], and recently, transformer architectures [8]. Over the years, several language-processing libraries from academia and business have been made accessible to the public [9]. These libraries are fit with incredibly precise pre-trained models for the extraction of common entity classes, such as *Person*, *Date*, *Location*, *Organization*, etc. These models should either be improved or re-trained using relevant datasets for the desired entity types, since a particular business may need to recognize more specific entities in text.

In order to train a model with a high level of accuracy, a significant amount of labeled training data must be obtained. Although several carefully labeled, extremely precise, and general datasets are available online [10], their use may not be practical for the purpose at hand. It may not be possible to manually classify relevant data, or relevant data may not be available on the Internet.

To solve this issue, we provide a way to automatically generate labeled datasets for unique entity types, as seen in texts from the pharmaceutical industry. In our instance, this tactic is used on texts from the pharmaceutical domain, i.e., in news articles from the domain. The research described in [11], where the basic findings about named entity recognition and knowledge extraction from pharmaceutical texts were reported, is expanded in this paper.

By labeling *Drug* entities and analyzing the results, we demonstrate that it may be expanded to tagging additional custom entities in other texts in the pharmaceutical domain. The primary focus is on the automatic application of common language-processing tasks, including tokenization, handling stop words and punctuation, lemmatization, and the potential application of custom, business-specific text-processing functions, such as performing text similarity calculations or tagging multi-token entities by joining consecutive tokens.

Two well-known language-processing libraries, spaCy [12] and AllenNLP [13], come with a pre-trained model based on convolutional layers with residual connections and a pre-trained model based on ELMo embeddings [14], respectively. We used them as the baseline to assess the overall applicability and accuracy of the proposed methodology. The custom-trained models exhibit high tagging accuracy in tagging the custom entity *Pharmaceutical Organization* when compared to the initial pre-trained models' accuracy while tagging the more generic *Organization* entity over the same testing dataset. Likewise, a model trained on fine-tuned BERT is used for obtaining a more in-depth insight into the results. Finally, a fine-tuned BioBERT [15], a model based on the BERT architecture and pre-trained on biomedical text corpora, was used to further benchmark the proposed methodology.

The contributions of the work presented in this paper are:

- The extension of our prior work [3,4] based on the existing BioBERT model to be able to extract two new entity types, namely *Drug* and *Pharmaceutical Organization*, and proposing a technique for automatically building the training set, which is beneficial to multiple downstream tasks.
- We show how to create the labeled dataset, if we know many class representatives that we want to learn, which can be considered as a semi-supervised method.
- We show that we can optimize the performance of the whole learning process, especially in low-resource domains. This includes reducing the effort of time-consuming tasks such as manual labeling owing to the visualization tool. This time is multiple orders of magnitude lower than the performance of any of the current models.

The remainder of the paper is structured as follows. In Section 2, we review relevant works in the NER domain. Then, in Section 3, we describe the proposed methodology, and afterwards, in Section 4, we illustrate how it is applied and fine-tuned in the pharmaceutical domain. Section 5 explains how the knowledge graph can be generated and enriched. Finally, Sections 6 and 7 discuss the main contributions of the paper and the limitations of our study and conclude the work, while providing ideas for future work.

2. Related Work

By adding hierarchical identification, named entity recognition (NER), a crucial part of NLP systems for tagging entities with their appropriate classes, enhances the semantic context of the words. There is much new research being performed in this area right now, particularly in the area of neural network label sequencing optimization, which outperforms earlier NER systems based on domain dictionaries, lexicons, orthographic feature extraction, and semantic rules. Neural network NER systems with minimum feature engineering have gained popularity since [16], because of the results they provide. They do this by proposing unified n-dimensional word representations and convolutional-neural-network (CNN)-based neural-sequence-labeling models.

Character-level models treat text as distributions over characters, and they can generate embeddings for any string of characters within any textual context. With this, they improve the model's generalization on both frequent and unseen words, making them popular in the biomedical domain. A model based on stacked bidirectional long short-term memory (LSTM) was introduced in [17]. This model inputs characters and outputs tag probabilities for each character, achieving state-of-the-art NER performance in seven languages without using additional lexicons and hand-engineered features. In [18], the authors presented a language model composed of a CNN and LSTM, where they used characters as the input to form a word representation for each token in the sentence; thus, it outperformed word/morpheme-level LSTM baselines.

The authors of [19] proposed a biomedical named entity recognition (Bio-NER) method that is based on a deep neural network architecture, which utilizes word representations pretrained on unlabeled data collected from the PubMed database with a skip-gram language model. In [20], the authors developed a general model based on the long short-term memory network-conditional random field (LSTM-CRF), which outperforms cutting-edge entity-specific NER technologies. Word embedding techniques were used to capture the semantics of the terms in the phrase.

T5 [21] and XLNet [22] are state-of-the-art natural language processing (NLP) models that have been developed by Google. Text-to-text transfer transformer (T5) is based on the transformer architecture and is a general-purpose model that can be fine-tuned for various NLP tasks. XLNet, on the other hand, is a generalized autoregressive pretraining method that uses permutation language modeling to learn bidirectional representations from unlabeled text data. One key difference between the two models is that T5 is trained on a single task or direction of text, while XLNet can be trained on multiple tasks and can handle text in either direction. In terms of performance, both models have achieved state-of-the-art results on a range of benchmarks, but XLNet has shown particularly strong performance on natural language understanding tasks.

Since 2018, sequence-to-sequence (Seq2Seq) architectures that work with text have become a popular topic in NLP, due to their powerful ability to transform a given sequence of elements into another sequence. This concept fits well in machine translation. Transformers are models that implement the Seq2Seq architecture by using an encoder–decoder structure.

The launch of Google's BERT [8], which is built on a transformer architecture and incorporates an attention mechanism, is one of the most-recent achievements in this development. Due to its capacity to recognize contextual relationships between words (or sub-words) in a text, it excels in various NLP tasks, including NER, and is thus useful in the biomedical and pharmaceutical industries. For the recognition of biomedical named entities for content in Spanish, Hakala and Pyysalo [23] proposed a method based on conditional

random fields (CRFs) and multilingual BERT. The authors investigate feature-based and fine-tuning training methods for the BERT model for NER in Portuguese in [24]. A method for question answering in the biomedical sector was presented by Lamurias and Couto in their paper [25]. It was based on a transformer architecture.

A domain-specific language representation called BioBERT [15] was pre-trained on sizable biomedical corpora. Using the BERT architecture, it was pre-trained on large general domain datasets (English books, Wikipedia, etc.) and biomedical domain corpora (PubMed abstracts, PMC full-text articles). This language model offers better outcomes for NER and other biological text-mining applications.

The problem of co-reference resolution [14] was also discussed in [26], further stressing the application of it in downstream tasks, as well as the challenges associated with it, particularly with rarer and under-resourced languages. Therefore, the authors proposed a method to overcome this and apply it to process e-health records from the reception at a Lithuanian hospital.

In neural networks, transfer learning as a machine learning technique introduces the idea of re-usability, where a model created for one task may be utilized as the starting point for training a different problem with a much smaller training set. Transfer learning has been one of the most widely used methods for computer vision and NLP applications in recent years, because it consistently outperforms state-of-the-art models while using much less computational power and training data.

Over the past few years, transfer learning has helped the F1-score for co-reference resolution tasks rise, enabling it to attain a gratifying average of 73%. This assignment aims to group textual mentions of the same underlying real-world objects into groups. Different methods employ biLSTM and attention processes to calculate span representations, and then, a softmax mention ranking model [27] is used to locate co-reference chains. The F1-score significantly improved with the addition of ELMo and coarse-to-fine and second-order inference, obtaining the aforementioned average of 73%. This task was evaluated with the OntoNotes co-reference annotations from the CONLL2012 shared task [28], which involved predicting co-references in English, Chinese, and Arabic, using the final version (5.0) of the OntoNotes corpus. It provides an accurate and integrated annotation of multiple levels of the shallow semantic structure in text in multiple languages.

On the other hand, using transfer learning for semantic role labeling demonstrates that using a straightforward BERT-based model can produce state-of-the-art results compared to earlier neural models that included lexical and syntactic features such as parts-of-speech tags and dependency trees [29]. The reason is that, out of the four tasks that make up semantic role labeling, predicate detection, predicate sense disambiguation, argument identification, and argument classification, the predicate disambiguation task, which can be formulated as a sequence labeling task and is where BERT really excels, is focused on determining the correct meaning of a predicate in a given context.

There are multiple ways to construct an RDF-based knowledge graph (KG), which generally depend on the source data. In our case, we worked with extracted and labeled data to utilize existing solutions that recognize and match the entities in our data with their corresponding version in other publicly available KGs. One such tool is DBpedia Spotlight, an open-source solution for automatic annotation of DBpedia entities in natural language text [30]. It provides phrase spotting and disambiguation, i.e., entity linking, for the provided input. Its disambiguation algorithm is based on cosine similarities and a modification of the TF-IDF weights. The main phrase spotting algorithm is exact string matching, which uses LingPipe's (http://alias-i.com/lingpipe, accessed on 1 November 2022) Aho–Corasick implementation.

Many systems, such as AllenNLP [13] and Spacy [12], attempt to provide demo sites for NLP model testing, as well as code snippets for machine learning experts to use more conveniently. On the other hand, libraries such as Hugging Face Transformers [31] and Deep Pavlov AI [32], considerably speed up prototyping and make it easier to develop new solutions based on the NLP models that already exist.

In the past year, we have seen several models developed that try to solve a similar problem to the one we target. Ruijie et al. [33] developed an entity-recognition model, which they used on abstracts of biomedical scientific papers. However, even though their model showed very good F1-score and accuracy values on the two datasets they used, their approach is focused on general pharmaceutical entities: genes, diseases, metabolic processes, etc., in contrast to our focus on pharmaceutical organizations and drugs. Colombo and Oliveira [34] developed a system that extracts information from pharmaceutical package inserts, to help health professionals guide their patients. Their approach targets the extraction of drugs, diseases, and people, and even though the overall F1 value was comparable to ours, their F1 value for detecting drugs was only 59.14%.

Currently, the main challenge is not only building an architecture of a model, but also obtaining a labeled training dataset. Therefore, the novelty of our work is that we provide the methodology and source code to crawl a large dataset of drugs and diseases, which can be later used and fine-tuned to obtain even larger labeled datasets. To the best of our knowledge, there is no full solution for knowledge extraction in the pharmaceutical domain that is focused on the needs of professionals and allows for the visualization of the outcomes in a manner that they can comprehend. We provide a solution in the form of a platform that aims to close this gap.

The next sections provide a full overview of the labeled-dataset-generation process, followed by an assessment of the custom model training. The extracted entities can aid in the documents and news filtering, but this is insufficient in the age of "data overload". Consequently, we take things a step further and include these findings in a platform that later extracts and presents the knowledge associated with these entities. Currently, this platform integrates state-of-the-art NLP models for co-reference resolution [14] and semantic role labeling (SRL) [29] to extract the context in which the entities of interest appear. This platform additionally offers convenient visualization of the obtained findings, which brings the relevant concepts closer to the people who use the platform.

The Resource Description Framework (RDF) [35] is then used to create a knowledge graph (KG), which is a graph-oriented knowledge representation of the entities and their relations. This provides two main advantages: the RDF graph data model enables the platform to seamlessly integrate the results of multiple knowledge extraction processes from a variety of news sources and, at the same time, links the extracted entities to their counterparts in DBpedia [36] and the rest of the linked data on the web [37]. This gives platform users uniform access to all knowledge collected from the platform and relevant connected knowledge already existing in knowledge graphs that are open to the general public.

3. PharmKE Knowledge Extraction Platform

In this section, we describe the proposed PharmKE platform [38,39], which, in addition to identifying *Drugs* and *Pharmaceutical Organizations*, also extracts relations in the mentioned context and constructs a knowledge graph based on them.

As shown in Figure 1, the platform includes the full process of understanding a document and its content, from categorization and filtering (i.e., whether it fall inside the pharmaceutical domain) through visualization of the entities and their semantic relationships. In this part, each stage is explained in further depth.

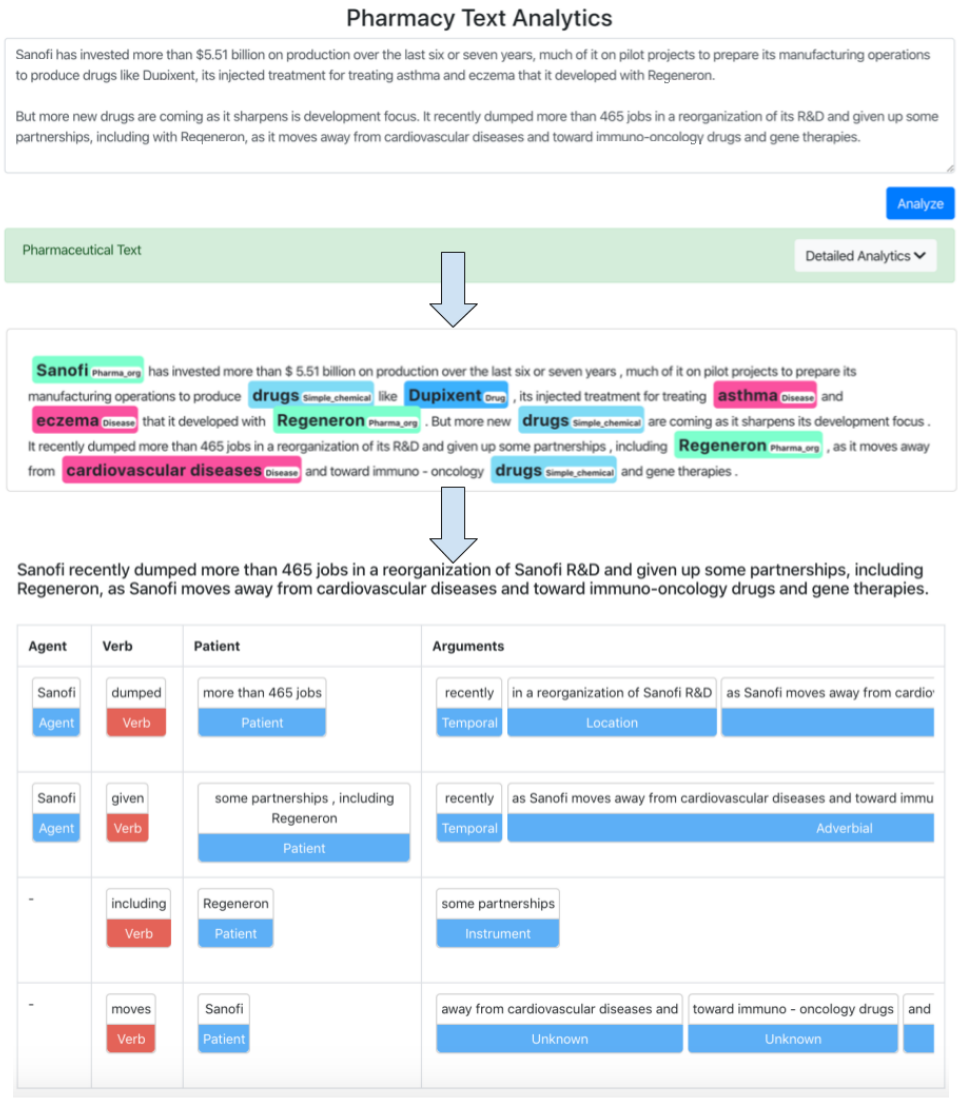

Figure 1. Platform workflow, available via the public instance of the platform [38].

The PharmKE platform can be formally represented with the following functional expression:

$$
\begin{aligned}
&bestKE(\\
&\quad bestSRL(\\
&\quad\quad bestCRR(\\
&\quad\quad\quad fineTunedPharmaNER(\\
&\quad\quad\quad\quad pharmaText(text)\\
&\quad\quad\quad)\\
&\quad\quad)\\
&\quad)\\
&)
\end{aligned}
\tag{1}
$$

The functional expression (1) shows that the platform is designed to combine the best of the available models in each of the steps, while also enabling us to fine-tune some of the models, as is the case with the $fineTunedPharmaNER$ model, which is explained in more detail in Section 4.

3.1. Pharmaceutical Text Detection

In the beginning, the platform classifies whether a given text is from the pharmaceutical domain, and only the positively classified texts are accepted for further analysis. For the documents in the pharmaceutical domain, we are certain about the assigned classes, because in the dataset, we include documents from scientific journals in this area. The negative label is assigned to documents representing the news that do not belong to this class. In terms of named entities' annotation, our goal was to show that we can bootstrap the dataset automatically, if we have large enough elements that represent the classes of interest. In this case, we manually checked that the annotated elements are correct.

The classification model used in this step is a transferred BERT model, fine-tuned with a corpus of ∼5000 documents from the pharmaceutical domain as positive samples (the documents were extracted from https://www.fiercepharma.com/, https://www.pharmacist.com/ and https://www.pharmaceutical-journal.com/, all accessed on 1 November 2022) and general news documents as negative samples (https://www.kaggle.com/snapcrack/all-the-news, accessed on 1 November 2022). Seventy percent of these documents were used for fine-tuning BERT's and XLNet's models, and their precision, recall, and F1-score were evaluated with the remaining thirty percent of the documents. Table 1 shows the results obtained by the fine-tuned models.

Table 1. Pharmaceutical text classification.

Model	Precision	Recall	F1
BERT	0.9633	0.9528	0.9580
XLNet	0.9983	0.9871	0.9926

3.2. Pharmaceutical Named Entity Recognition

Each correctly classified pharmaceutical text was further analyzed by recognizing combined entities through the proposed models, as well as by using BioBERT for the detection of BC5CDR (https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/, accessed on 1 November 2022) and BioNLP13CG (https://github.com/cambridgeltl/MTL-Bioinformatics-2016/tree/master/data, accessed on 1 November 2022) tags [40], which include *Disease*, *Chemical*, *Cell*, *Organ*, *Organism*, *Gene*, etc. Additionally, we used a fine-tuned BioBERT model in order to detect *Pharmaceutical Organizations* and *Drugs*, entity classes that are not covered by the standard NER tasks. We explain the fine-tuning process in more detail in Section 4. Tag collisions when combining the results from both models were avoided by applying the precedence of the tags recognized by our fine-tuned model over the tags recognized by BioBERT's model (*Simple Chemical*). All of the recognized entities are visualized in the sentence, along with their respective tags.

3.3. Co-Reference Resolution and Semantic Role Labeling

By using co-reference resolution in the background and swapping out each mention ("it" "it's", "his", etc.) with its appropriate entity, the identified entities serve as a baseline for detecting all of their mentions across the full text. Libraries such as AllenNLP, StanfordNLP [41], and NeuralCoref (https://github.com/huggingface/neuralcoref, accessed on 1 November 2022) provide implementations of the algorithms for co-reference resolution, focused on the CONLL2012 shared task [28]. Due to its high accuracy, simplicity of integration compared to StanfordNLP, and flexibility to include user-specific data and speakers in a conversation, our platform uses the NeuralCoref library for co-reference resolution.

The last step is categorizing the semantic roles in each phrase once the mentions in the text have been replaced with their appropriate entities. This is performed by using the BERT-based algorithm for semantic role labeling [29]. Then, for easy comprehension, the platform sequentially visualizes the concrete arguments, such as subject and object, as well as the modifier arguments, such as temporal, location, instrument, etc.

The end result is a modular framework for pharmaceutical text analysis that employs both fine-tuned models for detecting unique entities such as *Pharmaceutical Organization* and *Drug*, in addition to current state-of-the-art entity recognition models. The platform's modular architecture allows for the merging of findings from many models that may identify a wide variety of entities. Utilizing cutting-edge algorithms built by well-known libraries also enables semantic role labeling and visualization for each item and its corresponding mentions in the text. The full analysis may be exported in JSON format, making it usable for further processing, such as question answering, text summarizing, fact extraction, etc.

3.4. Knowledge Graph Generation

For the final step, we used the state-of-the-art knowledge extraction method DBpedia Spotlight to annotate the entire article [42]. Following that, we constructed and added further RDF facts to the acquired results using the detected *Pharmaceutical Organization* and *Drug* entities. After that, this enhanced knowledge graph is accessible for usage both inside and outside the platform.

4. Entity Recognition for Pharmaceutical Organizations and Drugs

Our methodology starts with a text corpora from the pharmaceutical domain and a closed collection of things that are members of a particular class. In our case, we utilized entities that stand for *Pharmaceutical Organizations* and *Drugs*. We demonstrate that we can train models that can extract even unseen entities from the class of interest using just these two preconditions. Figure 2 visualizes the whole process.

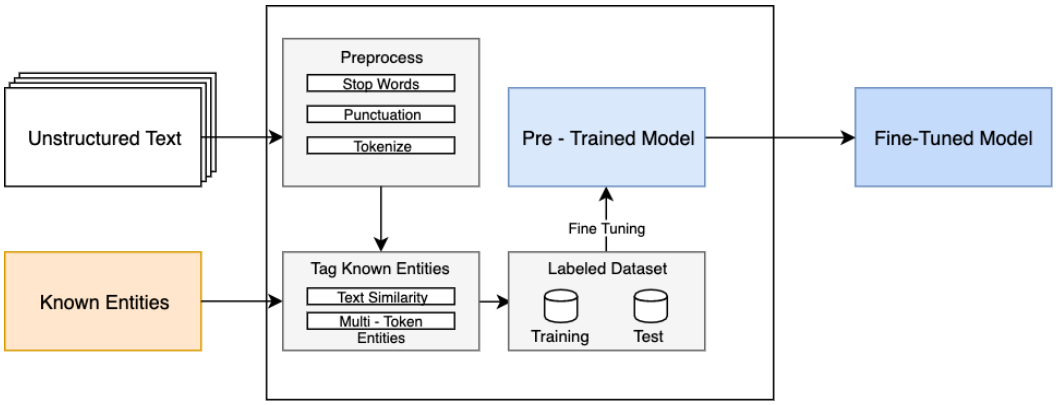

Figure 2. Named entity recognition pipeline.

The text corpora from the pharmaceutical domain that may contain the items from the class of interest are where we begin. The news in this text corpus was gathered from the following websites that are relevant to pharmacies: *FiercePharma* (https://www.fiercepharma.com/, accessed on 1 November 2022), *Pharmacist* (https://www.pharmacist.com/, accessed on 1 November 2022), and *Pharmaceutical Journal* (https://www.pharmaceutical-journal.com/, accessed on 1 November 2022). Next, we tokenized the text such that we extracted the words and, then, attempted to annotate each word in relation to the collection of entities from the needed type. We used the cosine similarity and Levenshtein distance to determine whether the word is comparable to some of the entities [43]. Each token in the text is given a start position and an end position throughout the annotation process. After finishing this stage, we will have created a labeled dataset, denoted as *MD*.

4.1. Creating a Labeled Dataset

One of the main challenges is that the *Pharmaceutical Organization* entity type can be found in a given text as multi-word phrases, such as "Sanofi Pharmaceuticals Ltd. Spain", or as a single word: "Sanofi". Additionally, the name of the *Pharmaceutical Organization* can contain pharmacy-related keywords, such as "Pharmaceuticals", "Pharma", "Medical", "Biotech", etc., which are not part of the core name of the organization and can either be

found along with it in the sentence or not at all. This means that we should not classify the countries, legal entities, and pharmacy-related words as parts of the *Pharmaceutical Organization* type. Therefore, the annotation process sequentially performs use-case-specific token filtering during the creation of the *MD* dataset.

A non-entity list, which comprises all tokens that need to be ignored, is used to do this. In our instance, the list includes all nations, as well as business legal forms (such as "Ltd.", "Inc.", "GmbH", "Corp.", etc.) and terms related to pharmacies. In our scenario, after removing the tokens from the non-entity list, only "Sanofi" will be left, and we can be sure that the core name has been fully extracted. The same lists are then used to identify any neighbor tokens for multi-token names that may be present as parts of the organization name using text similarity metrics once the core name has been matched in the text.

After the application of the custom, use-case-related filtering, the *MD* dataset consists of the core entities that have high text similarity. Only the entities that have similarity above the customized threshold are labeled as members of the target class. In our experiments, we used a similarity threshold of 0.9. Some *Pharmaceutical Organization* entities consist of multiple, consecutive tokens, such as "J & J". We solved this by token concatenation of consecutive relevant tokens, using a custom function applied on the *MD*.

After applying all custom text-processing functions, the state of the *MD* is as shown in Table 2.

Table 2. State of the *MD* after the application of the custom text processing functions.

Token	Range	Entity
Sanofi	0:14	PH_ORG
GlaxoSmithKline	258:272	PH_ORG
Regeneron	3436:3440	PH_ORG
Regeneron	3649:3654	PH_ORG
Gilead	3660:3668	PH_ORG
Sanofi	3699:3704	PH_ORG
J & J	3801:3806	PH_ORG

4.2. Model Fine-Tuning

Next, a model that can extract named entities from the specified class was trained using the *MD* dataset. The training dataset does not need to contain a large number of varied entities since NER models take into account the context in which the entities appear in a phrase. Here, we used small to moderate quantities of labeled data to enhance the general knowledge language model for the more particular job.

In our case, we fine-tuned the spaCy, AllenNLP, BERT, and BioBERT models. However, each of these models requires a different data format. SpaCy requires an array of sentences with respective tagged entities for each sentence and their start and end positions. AllenNLP requires a dataset in BIOUL or BIO notations (https://natural-language-understanding.fandom.com/wiki/Named_entity_recognition, accessed on 1 November 2022), which differentiate the following token annotations:

- Multi-word entity beginning token: *(B)*;
- Multi-word entity inside tokens: *(I)*;
- Multi-word entity ending token: *(L)*;
- Single-token entities: *(U)*;
- Non-entity tokens: *(O)*.

Regardless of the number of tokens, the dataset customized for BERT and BioBERT labels the entities with *I–PH_ORG*, while all other tokens are tagged with *O*. As a result, we exported the training and test datasets for the fine-tuning procedure in the needed format using various dataset serializers. For the *Drug* entity type, labeled datasets are produced using the same process. In this instance, we made use of the same text corpora, but they were annotated with a somewhat bigger collection of *Drug* entities. After finishing

the process of fine-tuning, we have named entity recognition models that can extract the entities from a given type.

4.3. Evaluation

Approximately 5000 news items from a collection of pharmacy-related news were used to gauge the accuracy of our suggested method. The *Drug* entities set had 20,266 distinct drug brand names, whereas the *Pharmaceutical Organization* entities set had 3633 unique values. As a part of our earlier effort [3,4], these sets were already extracted and released.

Both entity types were subjected to two separate assessment situations. Without taking into account the distribution of the entities within them, we divided the news articles from the dataset into training and test parts with sizes of 70% and 30%, respectively. To mitigate the risk of bias and randomness, the whole process was repeated 10 times, and the results presented in this paper are the averages from the 10 repetitions. This approach helped us evaluate the refined model's overall accuracy.

We assessed our approach's generalizability in the second evaluation scenario. In this case, we divided the training and test parts according to the entities they each include, ensuring that there was no entity overlap between the two. For testing purposes, we extracted the news articles that included 30% of the entities, while the remaining news was used for training. However, with this, more than 30% of the entire news document set was in the testing part. Therefore, the test component was decreased to comprise exactly 30% of the news articles, while the entities in the remaining documents were changed to other entities that did not belong to the entity set used in the testing portion in order to create a 70% to 30% ratio between the training and test portions.

4.3.1. Entity Recognition for Pharmaceutical Organizations

The obtained fine-tuned models for detecting *Pharmaceutical Organization* entities using spaCy, AllenNLP, BERT, and BioBERT were evaluated accordingly. The results were compared to the initial models prior to their fine-tuning, where the task was the extraction of *Organization* entities. The results are given in Table 3, indicating that the fine-tuned models are able to achieve significantly higher F1-scores compared to the original models. Furthermore, we can outline that AllenNLP outperforms spaCy in this NER task, a result that can be attributed to the different neural architectures used by both libraries, while the BERT model is able to outperform both. However, the pre-trained BioBERT on biomedical text is able to slightly outperform BERT in every evaluation.

Table 3. Evaluation of models trained on a dataset that contains known entities, for recognizing Pharmaceutical Organizations. The best result is denoted with an asterisk (*).

	PH_ORG		**Organization**	
Library	**Precision**	**F1**	**Precision**	**F1**
AllenNLP	95.57	90.3	49.41	48.26
spaCy	91.36	91.54	22.22	29.10
BERT	97.65	96.66	51.65	53.18
BioBERT	98.35 *	96.86 *	52.12 *	53.38 *

The sentence context in which the entities occur is taken into account by the pre-trained models, but we can assess the enhanced model generalization capabilities by producing a test dataset that only comprises the entities that were not present during the training. In order to do this, we selected a sample of entities at random from the joint dataset of pharmacy-related news in order to establish a split of 70% to 30% between the training and test datasets, where the test dataset comprises entities that were absent from the training dataset.

The SpaCy, AllenNLP, BERT, and BioBERT models were also trained using these datasets, and the results are given in Table 4. To better visualize the accuracy, Figure 3

denotes a sentence extracted from pharmacy-related news where the *Pharmaceutical Organization* entities are recognized as expected.

Table 4. Evaluation of the models on previously unseen entities, for recognizing Pharmaceutical Organizations. The best result is denoted with an asterisk (*).

	PH_ORG		Organization	
Library	**Precision**	**F1**	**Precision**	**F1**
AllenNLP	94.76	89.98	47.12	46.44
spaCy	90.95	88.51	21.98	28.01
BERT	97.45	97.68	51.51	55.68
BioBERT	97.52 *	97.86 *	52.42 *	55.70 *

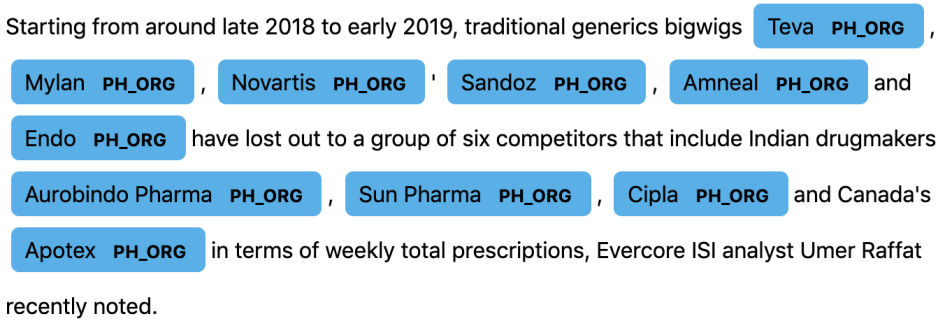

Figure 3. Detecting *Pharmaceutical Organization* entities in text.

4.3.2. Entity Recognition for Drugs

The SpaCy, AllenNLP, BERT, and BioBERT models were also created for recognizing *Drug* entities in texts. The evaluation results are given in Table 5 for the scenario where the same *Drug* entity can be present in both the training and the test dataset, while Table 6 shows the results when the test dataset does not contain any of the entities used in the training phase. Again, the training–test dataset ratio is 70–30%. To better visualize the accuracy, Figure 4 denotes a sentence extracted from pharmacy-related news, where the *Drug* entity is recognized as expected.

Table 5. Evaluation of models trained on a dataset that contains known entities, for recognizing Drugs. The best result is denoted with an asterisk (*).

Library	**Precision**	**F1**
AllenNLP	96.24	95.12
spaCy	90.95	94.87
BERT	98.86	95.98
BioBERT	98.92 *	96.14 *

Table 6. Evaluation of the models on previously unseen entities, for recognizing Drugs. The best result is denoted with an asterisk (*).

Library	**Precision**	**F1**
AllenNLP	92.65	89.85
spaCy	88.16	89.25
BERT	98.12	95.01
BioBERT	98.65 *	95.14 *

Drugs from the burgeoning JAK group have so far run into their fair share of safety liabilities, and if the label for new AbbVie med Rinvoq DRUG is any indication, the FDA now believes those issues are class-wide.

Figure 4. Detecting *Drug* entities in text.

5. Knowledge Graph Generation and Enrichment

As a final step in the pipeline, we wanted to generate an RDF knowledge graph (KG) with the knowledge extracted from the previous steps. One way to create a general-purpose knowledge graph is to use a tool such as DBpedia Spotlight [42], which performs recognition of interlinked entities in the DBpedia knowledge graph. Therefore, in theory, it can be used to recognize the drugs and pharmaceutical organizations in the texts of interest and correctly annotate them with their semantic type. However, our experiments showed that the annotated entities were of more general types, such as schema:Organization (http://schema.org/Organization, accessed on 1 November 2022) or dbpedia:Company (http://dbpedia.org/ontology/company, accessed on 1 November 2022). In addition to that, most drug entities referenced by their brand names were not annotated at all. Therefore, we decided to use the results obtained so far by the pipeline described in the previous sections, to expand the knowledge graph generated by DBpedia Spotlight with specific types: schema:MedicalOrganization (http://schema.org/MedicalOrganization, accessed on 1 November 2022) for the recognized Pharmaceutical Organizations and schema:Drug (http://schema.org/Drug, accessed on 1 November 2022), dbpedia:Drug (http://dbpedia.org/ontology/Drug, accessed on 1 November 2022) for the recognized Drugs.

To properly test the benefits of this knowledge graph enrichment, we decided to apply the technique on the test set that contains texts with previously unknown entities while training the named entity recognition models. The results show an average expansion of 47.69% on the originally generated knowledge graph by DBpedia Spotlight. Figure 5 shows an example knowledge graph for a given input text, extracted using the DBpedia Spotlight annotation tool (left), and the enriched knowledge graph with additional knowledge about *MedicalOrganization* and *Drug* entities (right).

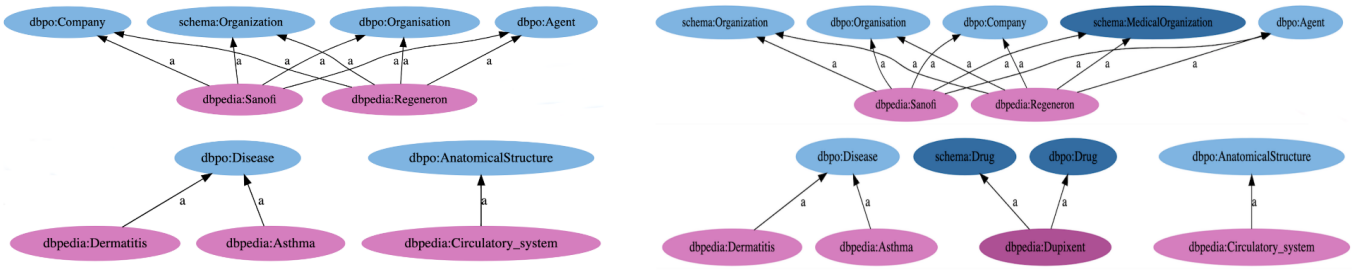

Figure 5. Original knowledge graph generated by DBpedia Spotlight (**left**) and the expanded knowledge graph (**right**). The additional RDF triples are highlighted.

Figure 6 shows the overall knowledge enrichment obtained by our system for the test dataset. It presents the ratio between the number of texts and the percentage of knowledge enrichment. This overview indicates a normal distribution of the enrichment over the test set.

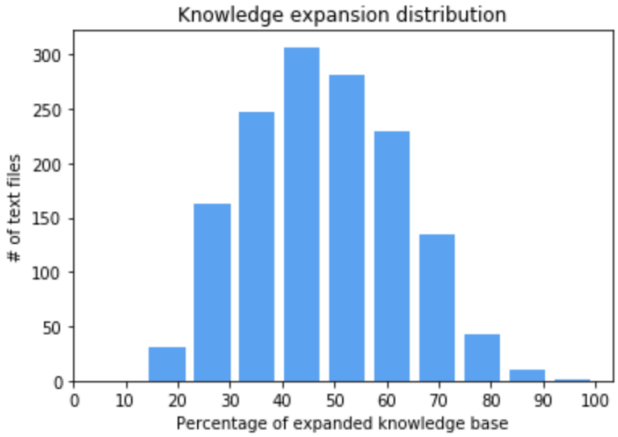

Figure 6. Distribution of knowledge graph enrichment among the texts from the test set.

The knowledge graph generated and enriched as part of the pipeline can then be used for other purposes within or outside the platform. We are currently providing an RDF output in Turtle syntax (https://www.w3.org/TR/turtle/, accessed on 1 November 2022).

6. Discussion

The platform outlined in this work places a strong emphasis on a strategy for integrating the top NLP models and applying them to a new domain. We employed a modular strategy, wherein each model is a distinct stage in the information extraction pipeline, enabling a simple upgrade with new and potentially better models, ultimately enhancing platform performance.

In contrast to [12,13,31,32], our platform's objective is to offer a knowledge extraction solution for the pharmaceutical industry that makes cutting-edge NLP accomplishments more accessible to those who examine enormous volumes of text. Because the PharmKE platform is human-centric, it is primarily intended for users who need to extract knowledge. Users can better grasp the procedure for capturing and connecting this knowledge since each phase's results are shown. We are also releasing an application programming interface (API) that exposes the outcomes from our platform to other applications because the web browser might not be the most-practical tool for domain experts to use in the process of knowledge extraction, particularly when they analyze texts from various sources. By doing this, we make it possible to create editor plugins that might one day extract and display the knowledge contained in the tools that professionals now work with regularly.

In the most-recent release of the PharmKE platform described in this paper, we improved the named entity recognition module to extract two new entity types in addition to those previously recognized by the better BioBERT model, namely *Drug* and *Pharmaceutical Organization*. Using a text corpus from the pharmaceutical domain and a closed set of entity instances from the kinds of interest, we demonstrate a technique for automatically building the training set for the recognition of *Pharmaceutical Organization* and *Drug* during the fine-tuning phase. This technology allows for the recognition of entities that are not included in the training set, which is a promising outcome, according to the evaluation of the fine-tuned model.

The goal of our paper was not to optimize the model performances for a specific task, such as the text analysis, but to show that we can optimize the performance of the whole process, which includes time-consuming tasks such as manual labeling, especially in low-resource domains. This time is multiple orders of magnitude lower than the performance of any of the current models. Therefore, in this paper, we focused on how to create the labeled dataset, if we know many class representatives that we want to learn. The speedup that the proposed methodology introduces, in combination with the visualization tool, is quite significant, especially in the data-labeling phase.

The knowledge graph that we constructed and enhanced at the last step of the pipeline aims to demonstrate the potential for packaging and reusing the knowledge produced by the pipeline in other software solutions. Because an RDF knowledge graph is created as the last stage in the platform's process, even if it is human-centric, the outcomes may be saved, shared, merged with other RDF knowledge graphs, and (re)used programmatically outside of the platform. Due to the nature of RDF and knowledge graphs, it is possible to practically seamlessly combine platform findings with additional RDF data that are available externally or internally in the user environment.

The PharmKE platform is receptive to ongoing developments in the NLP industry. The coupling of the relations acquired by the SRL model with the relevant attributes in the knowledge graph is one of the essential steps in the knowledge extraction process that is not addressed by the present models. In addition to adding any model that will produce better outcomes in some of the present jobs, our team will attempt to address this difficulty in its future study. The platform's modular construction makes all of this feasible. Cleaning up the knowledge graph from incorrect inferences produced by the pipeline, which is a common and anticipated issue with NLP, would be another obstacle.

One limitation of the study is that we do not redistribute the dataset because we do not have such a license. However, we do provide the source code (see [38,39]) so that interested readers can execute it and recreate the dataset by themselves. With this code, all experiments performed after the creation of the dataset can also be reproduced.

7. Conclusions

Using cutting-edge models for text categorization, pharmaceutical domain named entity recognition (NER), co-reference resolution (CRR), semantic role labeling (SRL), and knowledge extraction, we built our modular PharmKE platform [38,39]. The platform is primarily intended for human users. Pharmaceutical domain specialists may easily identify the information extracted from the input texts thanks to PharmKE's visualization of the findings from each of the integrated models.

The PharmKE platform's modular architecture makes it simple to incorporate additional and potentially improved models, which is one of our strategic goals. One such move in this manner was our addition of the *Pharmaceutical Organization* and *Drug* entity type identification to the more modern BioBERT model for NER.

Additionally, the platform is open-source and publicly accessible [38,39], providing the reproducibility of our findings. This also implies that, as a result of the platform's modular architecture, other researchers can alter their own copies of it, use them to run their own instances, and even re-purpose it.

The proposed methodology could be used in mobile and pervasive systems since it enables patients to scan the medication instructions for their prescriptions, which can give them more pertinent and understandable information. The proposed methodology may also be used to check whether a drug from a different vendor is compatible with the patient's prescription medication. The potential of patient empowerment lies on such methods.

The absence of labeled datasets for testing and training custom models for language comprehension tasks in text is a prevalent problem. We offer a way to automate the process of producing labeled datasets for training models for custom entity tagging in order to address this problem. SpaCy, AllenNLP, BERT, and BioBERT were used to train custom models for named entity identification in order to evaluate the technique. The findings show that the newly trained models perform better at identifying custom entities than the pre-trained models.

8. Future Work

The results from testing the suggested methodology's effectiveness on texts related to pharmaceuticals are satisfactory. To have a better understanding, the approach should be tested on a variety of texts with diverse contexts that may or may not contain elements from the pharmaceutical domain. With this, we might assess the methodology's effectiveness

more broadly and contrast the findings with the outcomes of the existing, task-specific assessment. This would make it possible to use it for training various models in a range of areas.

By turning our attention to the platform, we may further parse the retrieved semantic roles into RDF triples, which make up a knowledge graph. As part of the upcoming work, a platform optimization is planned. This will allow the background maintenance of the knowledge graph, which will be continually improved with each text analysis the platform performs. By using queries over the knowledge graph, the system will make it simple to access and extract data. Further, it may be coupled to other relevant knowledge graphs of the user, or ones that are publicly available.

Author Contributions: Conceptualization: N.J., R.S., M.J., D.T., and K.M., methodology: N.J., M.J., D.T., R.S., and E.Z., software: N.J. and R.S., validation: E.Z., M.J., D.T., R.S., and K.M., formal analysis: N.J., M.J., D.T., R.S., and K.M., investigation: N.J., M.J., D.T., R.S., E.Z., and K.M., supervision: D.T., writing—original draft preparation: N.J., M.J., D.T., K.M., R.S., and E.Z., writing—review: N.J., M.J., D.T., K.M., R.S., and E.Z.; and editing: N.J., M.J., D.T., K.M., R.S., and E.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The work in this paper was partially financed by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code developed in this study is available at [39]. An instance of the PharmKE platform is publicly available at [38]. The source data used in this study was obtained from multiple websites, listed in Section 3.

Acknowledgments: The work presented in this article was partially funded by Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering. We also acknowledge the support of NVIDIA through the donation of a Titan V GPU.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; nor in the decision to publish the results.

References

1. Krishnan, V.; Ganapathy, V. Named Entity Recognition. 2005. Available online: https://cs229.stanford.edu/proj2005/KrishnanGanapathy-NamedEntityRecognition.pdf (accessed on 1 November 2022).
2. Sang, E.F.; De Meulder, F. Introduction to the CoNLL-2003 Shared task: Language-independent named entity recognition. *arXiv* **2003**, arXiv:cs/0306050.
3. Jovanovik, M.; Trajanov, D. Consolidating Drug Data on a Global Scale Using Linked Data. *J. Biomed. Semant.* **2017**, *8*, 3. [CrossRef] [PubMed]
4. Jofche, N.; Jovanovik, M.; Trajanov, D. Named Entity Discovery for the Drug Domain. In Proceedings of the 16th International Conference on Informatics and Information Technologies, Prague, Czech Republic, 29–31 July 2019.
5. Sundermeyer, M.; Schlüter, R.; Ney, H. LSTM Neural Networks for Language Modeling. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, ON, USA, 9–13 September 2012.
6. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. *arXiv* **2016**, arXiv:1603.01360.
7. Chiu, J.P.; Nichols, E. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
8. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
9. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *arXiv* **2018**, arXiv:1812.09449.
10. Balasuriya, D.; Ringland, N.; Nothman, J.; Murphy, T.; Curran, J.R. Named Entity Recognition in Wikipedia. In Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web), Singapore, 7 August 2009; pp. 10–18.
11. Jofche, N.; Mishev, K.; Stojanov, R.; Jovanovik, M.; Zdravevski, E.; Trajanov, D. Named Entity Recognition and Knowledge Extraction from Pharmaceutical Texts using Transfer Learning. *Procedia Comput. Sci.* **2022**, *203*, 721–726. [CrossRef]

12. Honnibal, M.; Montani, I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *Appear* **2017**, *7*, 411–420.
13. Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N.F.; Peters, M.; Schmitz, M.; Zettlemoyer, L.S. AllenNLP: A Deep Semantic Natural Language Processing Platform. *arXiv* **2017**, arXiv:1803.07640.
14. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. *arXiv* **2018**, arXiv:1802.05365.
15. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef] [PubMed]
16. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
17. Kuru, O.; Can, O.A.; Yuret, D. Charner: Character-Level Named Entity Recognition. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 911–921.
18. Kim, Y.; Jernite, Y.; Sontag, D.; Rush, A.M. Character-Aware Neural Language Models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
19. Yao, L.; Liu, H.; Liu, Y.; Li, X.; Anwar, M.W. Biomedical Named Entity Recognition Based on Deep Neutral Network. *Int. J. Hybrid Inf. Technol* **2015**, *8*, 279–288. [CrossRef]
20. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep Learning With Word Embeddings Improves Biomedical Named Entity Recognition. *Bioinformatics* **2017**, *33*, i37–i48. [CrossRef] [PubMed]
21. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J.; et al. Exploring the Limits of Transfer Learning With a Unified Text-To-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
22. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–11.
23. Hakala, K.; Pyysalo, S. Biomedical Named Entity Recognition with Multilingual BERT. In Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Hong Kong, China, 4 November 2019; pp. 56–61.
24. Souza, F.; Nogueira, R.; Lotufo, R. Portuguese Named Entity Recognition using BERT-CRF. *arXiv* **2019**, arXiv:1909.10649.
25. Lamurias, A.; Couto, F.M. LasigeBioTM at MEDIQA 2019: Biomedical Question Answering using Bidirectional Transformers and Named Entity Recognition. In Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 1 August 2019; pp. 523–527.
26. Žitkus, V.; Butkienė, R.; Butleris, R.; Maskeliūnas, R.; Damaševičius, R.; Woźniak, M. Minimalistic approach to coreference resolution in Lithuanian medical records. *Comput. Math. Methods Med.* **2019**, *2019*, 9079840. [CrossRef]
27. Lee, K.; He, L.; Lewis, M.; Zettlemoyer, L. End-to-End Neural Coreference Resolution. *arXiv* **2017**, arXiv:1707.07045.
28. Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; Zhang, Y. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task, Jeju Island, Korea, 12–14 July 2012; pp. 1–40.
29. Shi, P.; Lin, J. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv* **2019**, arXiv:1904.05255.
30. Daiber, J.; Jakob, M.; Hokamp, C.; Mendes, P.N. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In Proceedings of the 9th International Conference on Semantic Systems (I-Semantics), Graz, Austria, 4–6 September 2013.
31. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arXiv:abs/1910.03771.
32. Burtsev, M.; Seliverstov, A.; Airapetyan, R.; Arkhipov, M.; Baymurzina, D.; Bushkov, N.; Gureenkova, O.; Khakhulin, T.; Kuratov, Y.; Kuznetsov, D.; et al. DeepPavlov: Open-Source Library for Dialogue Systems. In Proceedings of the ACL 2018, System Demonstrations, Melbourne, Australia, 15–20 July 2018; pp. 122–127.
33. Ruijie, Z.; Xinyu, T.; Xiaohua, L.; Yonghe, L. Entity Recognition and Labeling for Medical Literature Based on Neural Network. *Data Anal. Knowl. Discov.* **2022**, *6*, 100–112.
34. Colombo, C.d.S.; Oliveira, E.S.d. Intelligent Information System for Extracting Knowledge from Pharmaceutical Package Inserts. In Proceedings of the XVIII Brazilian Symposium on Information Systems, Curitiba, Brazil, 16–19 May 2022. [CrossRef]
35. Lassila, O.; Swick, R.R.; Wide, W.; Consortium, W. *Resource Description Framework (RDF) Model and Syntax Specification*; World Wide Web Consortium: Cambridge, MA, USA, 1998.
36. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
37. Bizer, C.; Heath, T.; Idehen, K.; Berners-Lee, T. Linked Data on the Web (LDOW2008). In Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 21–25 April 2008; pp. 1265–1266.
38. PharmKE Platform: Public Instance. Available online: http://pharmke.env4health.finki.ukim.mk (accessed on 1 November 2022).
39. PharmKE Platform: Source Code. Available online: https://gitlab.com/jofce.nasi/pharma-text-analytics (accessed on 1 November 2022).
40. Wang, X.; Zhang, Y.; Ren, X.; Zhang, Y.; Zitnik, M.; Shang, J.; Langlotz, C.; Han, J. Cross-type Biomedical Named Entity Recognition with Deep Multi-task Learning. *Bioinformatics* **2018**, *35*, 1745–1752. [CrossRef] [PubMed]

41. Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MA, USA, 23–25 June 2014; pp. 55–60.
42. Mendes, P.N.; Jakob, M.; García-Silva, A.; Bizer, C. DBpedia Spotlight: Shedding Light on the Web of Documents. In Proceedings of the 7th International Conference on Semantic Systems, Graz, Austria, 7–9 September 2011; pp. 1–8.
43. Gomaa, W.H.; Fahmy, A.A. A Survey of Text Similarity Approaches. *Int. J. Comput. Appl.* **2013**, *68*, 13–18.