big data and cognitive computing

MDPI

Article

Adverse Drug Reaction Concept Normalization in Russian-Language Reviews of Internet Users

Alexander Sboev [1,2,*], Roman Rybka [1], Artem Gryaznov [1], Ivan Moloshnikov [1], Sanna Sboeva [1], Gleb Rylkov [1] and Anton Selivanov [1]

[1] National Research Centre "Kurchatov Institute", 123182 Moscow, Russian Federation
[2] Department of Computer and Engineering Modeling, National Research Nuclear University "MEPhI", 115409 Moscow, Russian Federation
* Correspondence: Sboev_AG@nrcki.ru

Abstract: Mapping the pharmaceutically significant entities on natural language to standardized terms/concepts is a key task in the development of the systems for pharmacovigilance, marketing, and using drugs out of the application scope. This work estimates the accuracy of mapping adverse reaction mentions to the concepts from the Medical Dictionary of Regulatory Activity (MedDRA) in the case of adverse reactions extracted from the reviews on the use of pharmaceutical products by Russian-speaking Internet users (normalization task). The solution we propose is based on a neural network approach using two neural network models: the first one for encoding concepts, and the second one for encoding mentions. Both models are pre-trained language models, but the second one is additionally tuned for the normalization task using both the Russian Drug Reviews (RDRS) corpus and a set of open English-language corpora automatically translated into Russian. Additional tuning of the model during the proposed procedure increases the accuracy of mentions of adverse drug reactions by 3% on the RDRS corpus. The resulting accuracy for the adverse reaction mentions mapping to the preferred terms of MedDRA in RDRS is 70.9% *F*1-micro. The paper analyzes the factors that affect the accuracy of solving the task based on a comparison of the RDRS and the CSIRO Adverse Drug Event Corpus (CADEC) corpora. It is shown that the composition of the concepts of the MedDRA and the number of examples for each concept play a key role in the task solution. The proposed model shows a comparable accuracy of 87.5% *F*1-micro on a subsample of RDRS and CADEC datasets with the same set of MedDRA preferred terms.

Keywords: concept normalization; entity linking; entity disambiguation; Russian drug review corpus; deep learning; language models; natural language processing

Citation: Sboev, A.; Rybka, R.; Gryaznov, A.; Moloshnikov, I.; Sboeva, S.; Rylkov, G.; Selivanov, A. Adverse Drug Reaction Concept Normalization in Russian-Language Reviews of Internet Users. *Big Data Cogn. Comput.* **2022**, *6*, 145. https://doi.org/10.3390/bdcc6040145

Academic Editor: Moulay A. Akhloufi

Received: 19 October 2022
Accepted: 24 November 2022
Published: 29 November 2022

1. Introduction

The limited amount of clinical trials of drugs does not always allow coverage of the full range of adverse drug reactions. In turn, Internet resources contain a lot of drug information useful for pharmacovigilance and for the practice of drug use. Therefore, due to the time-consuming process of manual processing of texts from these sources, there is a need to automatically collect and analyze useful information from texts of Internet resources about drugs. Currently, the problem of extracting pharmaceutically significant information from electronic texts is intensively treated in the literature. It is usually decomposed into several tasks:

- filtering the texts containing pharmaceutically and medically significant information [1,2],
- extracting mentions of significant entities (Named Entity Recognition, NER task) [3–6],
- extracting relations between mentions of entities [7–9],
- searching for a correspondence between the selected pharmaceutically significant expressions (mentions) from texts and terms from dictionaries (concepts) (normalization task) [10–14].

The last task seems especially difficult; it requires establishing relations between mentions written in free form by non-specialists and the huge number of formalized terms in medical dictionaries. Moreover, the first often contains noisy data, jargon, abbreviations, and errors in the grammar and punctuation, along with vocabulary that is not contained in medical dictionaries. The above complicates the process of collecting labeled datasets and does not allow using standard classification methods for solving.

For some languages, primarily English, there are extensive annotated corpora for solving the normalization problem and a number of tested methods for concept normalization, but not for Russian (see Section 2).

The main goal of this work is to establish the accuracy of solving the normalization problem for Russian-language texts about medicines.

We formulate the normalization task as the multiclass text classification: the goal is to find for each mention $M = m_1, m_2, ..., m_{|M|}$ of review i a matching concept $C = c_1, c_2, ..., c_{|C|}$ from MedDRA (Medical Dictionary for Regulatory Activities) [15,16], which describes standard medical terminology and has a dictionary of preferred terms.

As a basis of the normalization method for the Russian language, we use a metric learning approach that was previously proven on the English datasets (see Section 2). The difference between the written styles of user mentions and of medical concepts are taken into account by using different language models as part of the whole neural network solution.

MedDRA has a hierarchical structure and includes more than 80 thousand terms, of which: 80262 low-level terms (LLT) and 23708 preferred terms (PT). A hierarchy makes it possible to determine the correspondence between terms of different levels. In this work, we focus on the PT level as some other works on this topic [17–19]. The neural network is trained and validated on the basis of the Russian Drug Review Corpora that contains the annotations of correspondence among mentions and concepts (see Section 3.1).

Due to the problem of imbalance, i.e., not all concepts are presented in the RDRS corpora at all, we expand the training set for our neural network solution by data from the available English corpora automatically translated into the Russian language (see Section 3.3).

The purpose of this study is not only to build a solution for the normalization task for the Russian language but also to compare the accuracy of the obtained results with the analogical English one. For this, we apply the state-of-the-art approach of automatic normalization to the annotated normalization dataset from Russian corpus RDRS [5] and compare the obtained result accuracy with the closest analog of the dataset in English (see Section 3.2). Moreover, we compare these corpora related to the compositions of included concepts and distributions of phrases among them and investigate reasons for accuracy differences in the results received from both corpora.

As a result, we set the accuracy for the adverse reaction mentions mapping to the preferred terms of MedDRA in RDRS, along with pointing out factors that affect the accuracy of the task's result. In the context of the conducted study, the main contributions of this work are:

- an estimation of the accuracy of solving the problem of normalizing the adverse reaction mentioned in Russian to MedDRA PT level terms;
- a procedure for training a neural network, including various models for the vectorization of mentions and contexts that is based on metric learning using an extended set of examples from English corpora;
- a comparative analysis of the results obtained on the RDRS corpus with the results of solving the normalization problem on a similar English-language corpus of Internet user reviews.

The rest of the article is organized as follows. In Section 2, we provide a survey of existing and commonly used methods and corpora suitable for the normalization task. Section 3 considers the corpora used in this research. Section 4 describes the neural network model and training procedure. Section 5 presents the experiment descriptions and achieved results. The analysis results and case study are provided in Section 6.

2. Related Works

Traditional approaches to the normalization task are based on a lexical matching of dictionary references and concepts using various heuristics, synonyms, etc. [20–22].

However, such approaches show a low accuracy for the informal language of Internet users. In particular, in the 2017 SMM4H competition, the lexical match solution was much worse than the best approach based on machine learning methods: the accuracy was 63.5% vs. 87.2% [23]. Neural network models demonstrate the best accuracy in the normalization task but require sufficient annotated data for efficient training. Papers [24,25] propose classification methods based on neural network types LSTM and CNN. The methods that represent the words as vectors include word2vec [26] model and trainable neural network layers (Embedding).

The works [12,14] develop the metric learning approach, which implements a learning process to minimize the distance between the mention vector and the corresponding concept. Concept vectors could be the result of joint training with the model for input mentions encoding, algorithms of graph encoding applied to the source of complex structures [14], or an application of the more efficient language models [12]. The quality of these methods could be improved with the preliminary fitting of the model on the collection of augmented samples or by introducing specialized loss functions, in particular, triplet loss [27]. The paper [28] presents a modification of the metric learning approach: the authors proposed to select the ten closest concepts for each mention sequentially to classify the correspondence in pairs of "concept and mention".

In this work, by analog with paper [12], we use a neural network model based on a metric learning approach due to its ability to combine modern language models at different stages of analysis.

For the English language, there are several corpora that exist that include an annotation for the normalization problem: CADEC [17], PsyTAR [29]—based on posts from the askapatient forum (https://www.askapatient.com/ accessed on 28 November 2022), the SMM4H Competition Corpora [19,30–33], TwADR-L [25], and the corpus from [34], where Twitter messages are used as texts.

The most commonly used among them and closest to the Russian RDRS corpus is CADEC. The corpus was created on the base of the texts of reviews of Internet users from the askapatient forum. To create the CADEC dataset, the authors collected reviews on medicines containing Diclofenac and Lipitor as active substances. The annotation was carried out for several types of named entities: Disease, Symptoms, Findings, Drug, and ADR. Mentions of the ADR type in this corpus contain the annotation matched to MedDRA LT level terms. In total, the corpus contains 1253 reviews with 5990 ADR mentions.

The PsyTar corpus is based on 891 reviews of 4 psychiatric drugs: Zoloft, Lexapro, Cymbalta, and Effexor XR. During the corpus collection, each feedback from the forum was divided into sentences, and sentences were classified into four classes based on the content of different types of mentions in them: ADR, Withdrawal Symptoms (WDs), Sign/Symptoms/Illness (SSIs), Drug Indications (DIs), Drug Effectiveness (EF), Drug Ineffectiveness (INF), and Others (not applicable). Further work was carried out to identify and extract mentions from sentences. At the final stage, the extracted mentions of the types ADR, WDs, SSIs, and DIs were compared with the corresponding UMLS metathesaurus concepts (916) and SNOMED CT concepts (755).

The SMM4H 2017 corpus consists of ADR mentions manually extracted from Twitter posts. Twitter posts were pre-selected for 250 keywords of drug trade names and their spellings with popular misspellings. Each reference was matched to a MedDRA PT level term. In total, the corpus contains about 9000 annotated examples.

The TAC-2017 corpus was presented as a part of the 2017 conference on the analysis of texts on adverse drug events [18]. The corpus consisted of 101 texts from packages and inserts of medicines (drug labels) distributed in the United States. The annotation was made for several tasks, including ADR normalization to the MedDRA standard. The corpus included more than 7000 mentions with MedDRA LLT and PT level annotation.

The MedMentions corpus consists of titles and abstracts of articles randomly selected from among the 2016 PubMed articles written in English on biology or medicine topic. The corpus contains annotations of mentions according to the UMLS dictionary. Mentions in a corpus do not have a corresponding named entity class tag.

Recently, work has been actively carried out to annotate Russian-language corpora and create information extraction tools on their basis [9,33,35,36]. Despite this, for the Russian language, there is only one RDRS corpus of texts of Internet user reviews about medicines [5] containing the annotation on the normalization for entities of ADR and Indication types, with references to the MedDRA dictionary (see Section 3). Described annotation allows us to conduct a study on the accuracy estimation for data extraction methods in the Russian segment of Internet reviews.

3. Datasets

3.1. Russian Drug Review Corpus

The development of the corpus began in 2019, including the annotation of named entities' mentions and their classes in the reviews of Internet users from the otzovik.ru website. The annotation contains the following set of classes and their attributes:

- Medication—mentions related to the medicines, including attributes, names of medicines, methods, dosages of their administration, etc.;
- Disease—mentions, including attributes, the name of the disease, symptoms, as well as the dynamics of development (positive, negative, and unchanged);
- ADR—a class of mentions related to adverse reactions described in reviews.

Further, the annotation has been extended to determine the matching among mentions and concepts of MedDRA PT level. Two types of mentions were selected for annotation:

- ADR: adverse effects of the drug mentioned in the text. For example, in the sentence: "I have very dry mucosa after the Snoop drops", the mention "very dry mucosa" is the entity of type ADR,
- Indication (class disease): symptoms of the disease. Therefore, in the sentence "During the coronavirus, I had a temperature of 40", "temperature of 40" is annotated as a symptom of the disease.

In total, 3837 reviews are currently annotated, highlighting more than 147,000 mentions of named entities across more than 20 entity types, of which 4941 mentions of the ADR class and 6967 mentions of the Indication type. More details about the annotation scheme are described in our previous works [5,9,37–39].

3.2. Comparison of the RDRS corpus with CADEC

The comparison of the quantitative characteristics of these corpora is presented in Table 1.

Table 1. Comparison of corpora datasets of Internet-user reviews in RDRS and in CADEC.

Corpora	Number of Texts	Mean Length of Texts (in Words)	Number of Mentions (incl. Unique)	Mean Length of Mentions (in Words)	Number of Unique PT
RDRS	3837	130	11,908 (5427)	2.2	936
RDRS (ADR)	1609	136	4941 (3117)	2.7	615
RDRS (Indication)	2525	136	6967 (2608)	1.9	627
CADEC	1244	103	5985 (2981)	2.6	442

Figure 1 shows the representativity of PT codes in mentions, i.e., the number of the PT codes that have an equal number of mentions in the dataset.

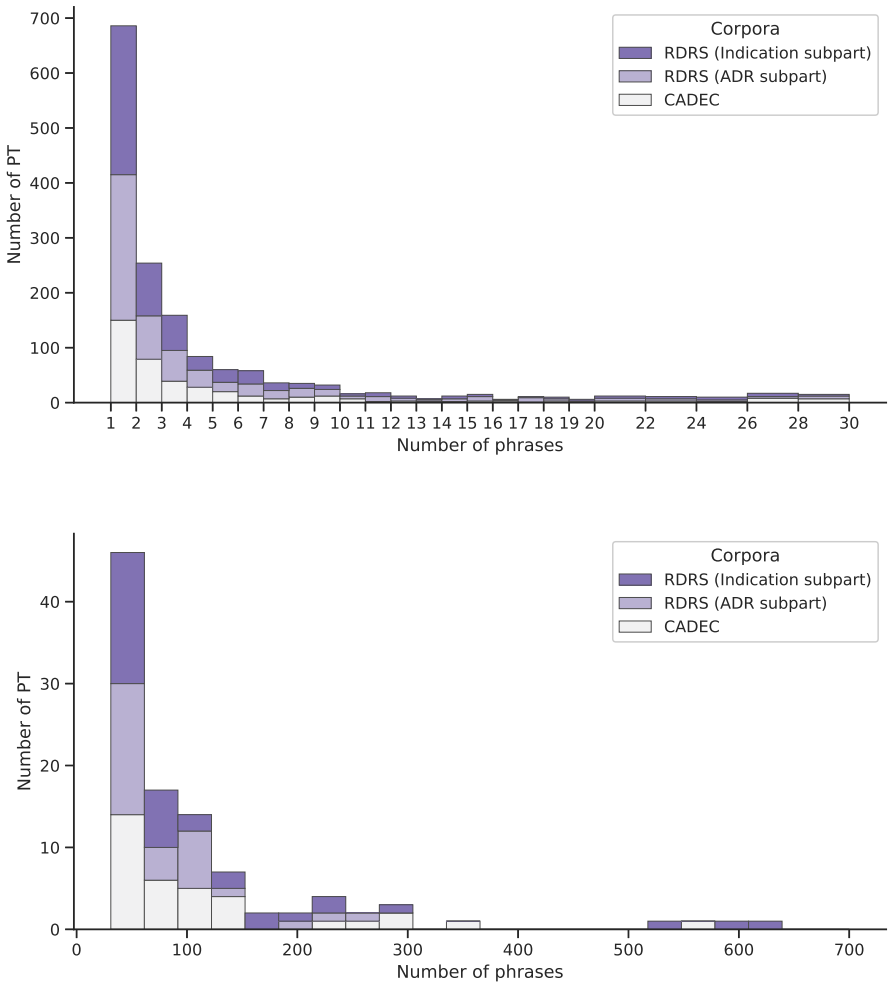

Figure 1. Histograms of "Phrase counts on PT" distribution for CADEC and RDRS corpora ((**Top picture**) number of phrases from 1 to 30; (**Bottom picture**) number of phrases from 31 to 700;

The analysis of Figure 1 and Table 1 demonstrates that the RDRS and CADEC corpora have a set of PT codes with the highest representativeness, and this set is in order of magnitude greater in the number of phrases per code than for the PT codes with the less representativeness. The RDRS, compared to the CADEC, has a large gap between codes that have only a single unique mention in the corpus and the rest of the codes that have two or more unique mentions in the corpus. This fact characterizes the RDRS corpus as more difficult for training neural network models: most of the unique codes in RDRS are single-representative codes. Those codes that have unique phrases in RDRS still have fewer of them than most of the representative codes in CADEC. Even so, the CADEC corpus is the closest corpus to RDRS in the source of texts, the type of annotation, and in the number of marked examples (see Table 2).

Due to the lack of accuracy in the solution of the ADR normalization problem for the Russian language in the literature, the results obtained from the RDRS are compared with the results from CADEC. For the subsequent comparison, we also select subsamples that are identical in terms of the set of unique PT codes in the RDRS corpus and in the CADEC (see Table 2).

Table 2. Comparison of the CADEC and RDRS datasets on the common PTs set.

Corpora	Number of Unique MedDRA PT	Number of PT MedDRA Intersecting with RDRS (ADR)	Number of PT MedDRA Intersecting with RDRS (Indication)
RDRS (ADR)	615	615	306
RDRS (Indication)	627	306	627
CADEC	441	221	187

When forming such comparative samples, we use MedDRA-tagged RDRS examples of both the ADR and Indication parts to expand the dataset since the mentions of these entities do not differ in wording (see Section 5).

3.3. Additional Corpora

We used: PsyTAR, SMM4H 2017, TAC, and MedMentions corpora. To use their data as pre-training material, we needed to translate them and, in some cases, to modify their annotations. PsyTAR does not have MedDRA annotation. Instead, its mentions are linked to UMLS metathesaurus concepts and SNOMED CT concepts. The corpus annotation was reduced to PT concepts using UMLS. As a result of this procedure, 1114 references were excluded from the comparison due to no suitable match. The MedMention corpus is also annotated with UMLS codes. They were automatically converted to MedDRA PT. At the same time, out of 350 thousand mentions, 316 thousand were excluded, for which no correspondence was found between UMLS CUI and PT MedDRA.

Besides English corpora, we also used low-level terms (LLT) from Russian MedDRA. The hierarchical structure of MedDRA defines the correspondence between the PT term and several of its LLT concepts. Therefore, LLT can be interpreted as variations of PT and be used as an extended set of examples for preliminary training.

As a result, the extended version of the annotated set includes:

- The English-language corpora, an annotation is converted to the MedDRA of PT level, and the mentions themselves are automatically translated into Russian using the Google API;
- Additional examples from Russian-language MedDRA.

The general statistics of the generated set are presented in Table 3.

Table 3. Features of the extended corpus.

Corpora	Total Number of Mentions with Normalization Markup	Number of Selected and Translated Mentions	Number of MedDRA PT Codes in Translated Subparts
TAC	7045	2482	1355
SMM4H 2017	9149	2863	440
PsyTAR	7414	2313	350
MedMention	352,496	7402	2818
MedDRA	84,139	80,377	24,999
Total:	460,243	97,793	25,071

The main disadvantage of the first set is the imperfection of automatic translation. For the translation, all mentions were used without their context of being used in the considered corpora, which could lead to an incorrect translation. Automatic translation of abbreviations leads to the transliteration of English letters into Russian, which is not correct. There were also cases of the same translation of different terms. All repeats in the

extended corpus were removed. The disadvantage of the second set is the strict formal language of MedDRA terms, which differs from the spoken language of review texts.

4. Methods

4.1. Normalization Model

The transformer-based architecture of the proposed model is inspired by the paper [12]. It includes two encoders: Mention Vectorization Model (MVM) and Concept Vectorization Model (CVM) (see Figure 2).

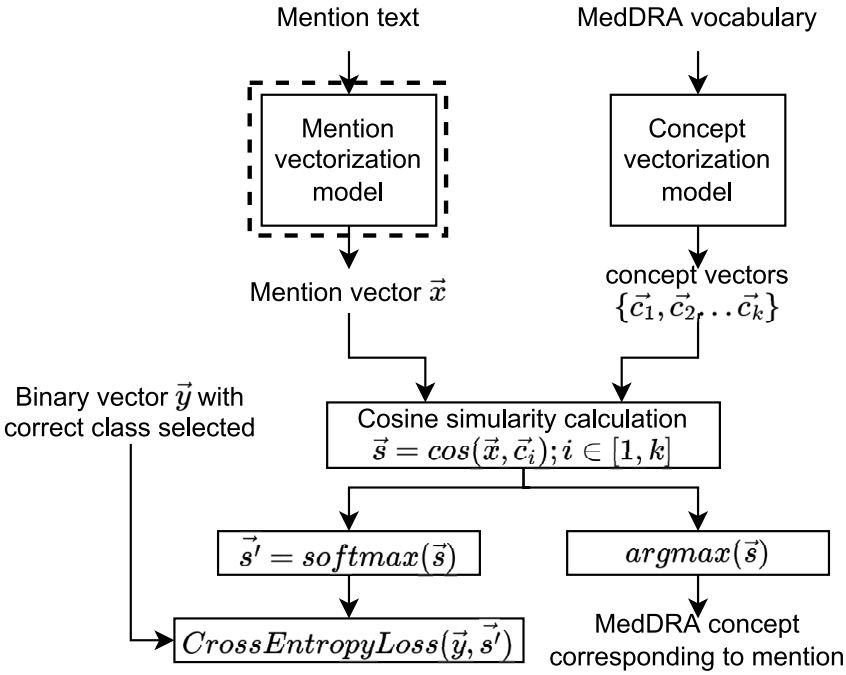

Figure 2. Architecture of the proposed model. Dotted line highlights the model with weights that tune during the fitting.

The difference in the models is in the tuning process: only the model weights of the MVM block change during the fine-tuning, and the concept vectors obtained as a result of the CVM block operation remain unchanged and are used as the target for calculating the loss function in the learning process.

Learning is aimed at the minimization of the cosine distance between vectors of input mentions and the concept vectors. Training affects only the weights of the layers that form the mention vector.

The vector of concept or mention is the average vector of tokens of the concept's/mention's text.

The most relevant concept for the analyzed mention is the closest concept vector by the cosine measure. The model output is the vector of the distance between the target mention and every concept in the dictionary, normalized using the softmax function to calculate the loss (categorical cross-entropy).

Figure 3 shows the input mention coding schemes. If the context is unused, the mention vector is the result of averaging the vectors of tokens of this mention text. In the case of context usage, the vector of the special token [CLS] at the beginning of the sentence is concatenated with the vector of the token. The language model was pre-trained to form a sentence representation vector for the [CLS] token. Therefore, it is used as a vector that encodes the context of the mention. The resulting concatenated vector is an input of the additional layer that reduces the dimensionality of the vector to the same size as the concept vector.

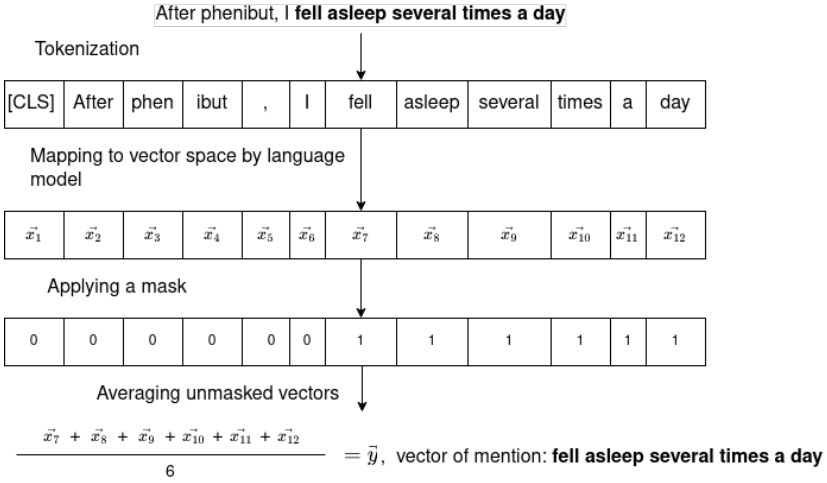

Figure 3. An example of mention vector preparation.

Experiments on the RDRS dataset were carried out on the basis of two language models:

- RuBERT [40], trained on the Russian part of Wikipedia and news data. It contains 12 layers of the transformer architecture, a 768-dimensional vector encoding one token in the model, 12 attention heads, and ∼110 million parameters;
- xlm-roberta-sag [5], trained on ∼1.2 million Russian drug reviews, contains 24 transformer architecture layers, 1024-dimensional vector encoding one token in the model, 16 attention heads, and ∼340 million parameters.

4.2. CADEC Normalization Model

Similar to the work [12] for the CADEC corpus, roberta and sroberta were chosen as the models to encode MedDRA mentions and concepts, respectively.

Roberta (base, large) [41]—a language model with a transformer architecture, analogous to BERT, was trained on the task of masked tokens prediction and the next sentence prediction. Roberta had a significantly bigger training dataset and modified training hyperparameters. The roberta training set had 160 GB of texts, including the following corpora: BookCorpus [42], English Wikipedia (https://en.wikipedia.org/wiki/, accessed on 28 November 2022), CC-News [43], OpenWebText [44], and Stories [45].

Sroberta is a version of the roberta model, additionally trained to form more informative vectors for the entire sentence. The learning process was carried out using the technology of Siamese networks on datasets annotated for the task of sentence pair classification: Stanford Natural Language Inference (SNLI) [46] and Multi-Genre NLI [47].

The obtained solution for the CADEC dataset was used in this work, on the one hand, to validate the normalization procedure and on the other hand, to compare the obtained accuracy on subsamples of English and Russian corpora with similar representativity of the PT codes (see Section 5).

These models are compared in various combinations when implementing the MVM and CVM blocks to select the best configuration.

4.3. Evaluation Metrics

Micro-averaged and macro-averaged F1-scores were used to estimate the accuracy of the normalization task solution. For each class i (PT MedDRA code in the analyzed corpus, N in total), F1 is calculated by the following formulae:

$$P_i = TP_i/(TP_i + FP_i),$$
$$R_i = TP_i/(TP_i + FN_i),$$
$$F1_i = 2 \cdot P_i \cdot R_i/(P_i + R_i)$$

$$F1_{macro} = \frac{\sum_{i=1}^{N} F1_i}{N}$$

Here TP_i is the number of correctly-classified examples of class i, FP_i is the number of examples of other classes that were determined as class i; FN_i is the number of examples of class i that were attributed to other classes.

The following formula was used to calculate $F1$-micro:

$$TP = \sum_{i=1}^{N} TP_i, FP = \sum_{i=1}^{N} FP_i, FN = \sum_{i=1}^{N} FN_i$$

$$P = TP/(TP + FP),$$

$$R = TP/(TP + FN),$$

$$F1_{micro} = 2 \cdot P \cdot R/(P + R),$$

Thus, $F1$-micro shows the accuracy of the classification with respect to all test cases, and $F1$-macro estimates the accuracy with respect to the representativeness of the classes in the analyzed corpus.

5. Experiments and Results

The paper shows four series of experiments to achieve the following goals:

1. Choosing an efficient language model for text encoding as part of a neural network model to solve the normalization problem. The goal was to find the best language models (see Section 4.1) for: (a) MVM block and (b) CVM block. Accuracy was estimated with a 5-fold cross-validation on the RDRS that includes texts containing mentions of ADR and Indication types.
2. Estimating the accuracy improvement by the preliminary training of the model on an additional dataset (see Section 3.3). The most efficient language models were used during this set of experiments.
3. Comparing the results of the normalization task on CADEC and RDRS corpora using the MedDRA dictionary. The experiments used the full set of ADR mentions from the texts.
4. Establishing the causes of differences in the accuracy of CADEC and RDRS normalization. The analysis was carried out on subsets of these corpora, including examples for MedDRA PT codes presented in both CADEC and RDRS—in both parts of ADR and Indication.

Table 4 represents the results of language model choice and the impact of pre-training.

Table 4. Comparison of language models used in the normalization solution for RDRS.

Pre-Training on Additional Corpora	Mention Vectorization Model	Concept Vectorization Model	*F*1-micro	*F*1-macro
False	xlm-roberta-sag	xlm-roberta-sag	67.9	23.8
	xlm-roberta-sag	RuBERT	75.4	28.5
	RuBERT	xlm-roberta-sag	66.7	21.6
	RuBERT	RuBERT	73.3	26.7
True	xlm-roberta-sag	RuBERT	76.5	30.5
	RuBERT	RuBERT	76.2	32.6

Table 5 represents the results of testing the developed model on the CADEC dataset in comparison with the literature data.

Table 5. Accuracy of normalization of ADR mentions for CADEC and RDRS (ADR subpart) corpora.

Approach	Mention and Concept Vectorization Models	CORPORA (Labels)	*F*1-micro	*F*1-macro
This work	RuBERT with pre-training, RuBERT	RDRS (PT)	70.9	39.9
This work	roberta-base, sroberta	CADEC (PT)	84.8	52.9
kNN method [11]	BERT	CADEC (LLT)	57.1	-
Metric learning [12]	roberta-base, sroberta	SNOMED CT	84.5	-

Table 6 shows the results of comparative experiments on subsets of the CADEC and the RDRS, including ones on subsets containing mentions related to the same set of PT (common PT*).

Table 6. Comparison of the accuracy of normalization on RDRS and CADEC corpora, * common PT: $PT \in \{PT(CADEC) \cap PT(RDRS(ADR\ subpart)) \cap PT(RDRS(Indication\ subpart))\}$.

Mention and Context Vectorization Models	Corpora	*F*1-micro	*F*1-macro
RuBERT with pre-training, RuBERT	RDRS (common PT*)	87.5	64.3
roberta-base, sroberta	Cadec (common PT*)	89.4	66.5

The best result on the RDRS dataset was achieved using domain-specific model xlm-roberta-sag to vectorize the mentions and RuBERT to vectorize the concepts. The pre-training on the formed corpus of additional mentions raises the accuracy by ∼1% for *F*1-micro and ∼2% for *F*1-macro for the vectorization model xlm-roberta-sag. At the same time, the highest *F*1-macro is achieved when concepts and mentions both vectorized with the RuBERT language model are trained on texts of general vocabulary.

6. Discussion

There are no direct analogies to the solution of ADR normalization in our formulation for the CADEC corpus in the literature. Therefore, it is difficult to compare the results of our approach with the ones of other researchers. Most of the works [11,12,14] use Snomed CT or MedDRA LLT annotation system for the ADR normalization, which differs from MedDRA PT. However, a comparison with the results of other methods on CADEC (see Table 5) confirms the successful implementation of the chosen neural network model for normalization in MedDRA PT.

The accuracy loss in the results of ADR normalization experiments on the RDRS in relation to the obtained ones on CADEC is 13.9%. Three core factors affect the results :

- primarily, the difference in the sets of PT codes: 615 codes in RDRS versus 442 in CADEC (see Table 1);
- a larger number of non-unique mentions in the CADEC in comparison with the RDRS (4941 in the RDRS vs. 5985 in the CADEC), in the presence of comparable numbers of unique mentions (3117 in the RDRS vs. 2981 in the CADEC);
- differences in the distributions of both types of mentions for PT codes between RDRS and CADEC (see Figure 4).

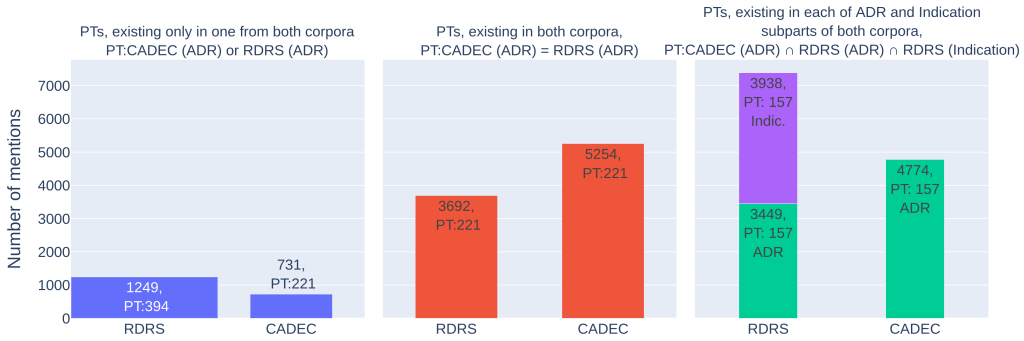

Figure 4. Similarities and differences between RDRS and CADEC corpora.

A result of experiments on subparts of RDRS and Cadec corpora with the same set of PT, in which the above-mentioned otherness of full versions of corpora is partly smoothed, demonstrating the closer coincidence of MedDRA normalization accuracies (the difference is 2% for *F*1-micro) that grounds the above-mentioned statement.

Table 7 presents the most common errors of the developed model of mention normalization according to the MedDRA PT annotation scheme. They can be divided into three main groups:

1. Inaccuracies from the 1st, 2nd, and 3rd sentences are associated with the annotation process when Experts matched references to several PT codes at once. In the cases of multivalued gold normalization, we only used the most common labeling option as a true value for model training.
2. Mistakes from the 4th, 5th, and 6th sentences are caused by no accounting contexts mentioned in training. In most cases, the context is redundant; however, in some examples, using only the mention text can be ambiguous.
3. Errors from the 7th, 8th, and 9th sentences were due to unbalanced numbers of examples of terms in the corpus. This leads to a more likely selection of terms with greater representativeness.

Eliminating these shortcomings by including the possibility of determining several PT codes for the analyzed mention and including information of the mention context in the feature space of the model, as well as expanding the corpus to reduce the imbalance, is the direction of our further work.

Table 7. Examples of the normalization for RDRS corpus with translation to English.

Sentence 1	The first reaction was as instructed: [malaise]$_{ADR\#1}$, [muscle pain]$_{ADR\#2}$, [joint pain]$_{ADR\#3}$, [chills]$_{ADR\#4}$, [temperature rise]$_{ADR\#5}$
Gold normalization	*ADR*#1:Malaise, *ADR*#2:Myalgia, *ADR*#3:Arthralgia, *ADR*#4:Chills, *ADR*#5:Pyrexia
Predicted normalization	*ADR*#1:Malaise, *ADR*#2:Myalgia, *ADR*#3:Arthralgia, *ADR*#4:Chills, *ADR*#5:Body temperature increased
Sentence 2	[Terrible vomiting]$_{ADR\#1}$ after these pills
Gold normalization	*ADR*#1:Vomiting projectile
Predicted normalization	*ADR*#1:Vomiting
Sentence 3	Polysaccharides, essential oils, marshmallow, yarrow, oak bark tannins help reduce [swelling of the mucous membrane of the respiratory tract]$_{Indication\#1}$
Gold normalization	*Indication*#1:Oedema mucosal
Predicted normalization	*Indication*#1:Respiratory tract oedema
Sentence 4	All the same [gastric juice comes out]$_{ADR\#1}$ of you and then [everything inside hurts]$_{ADR\#2}$
Gold normalization	*ADR*#1:Vomiting, *ADR*#2:Abdominal pain
Predicted normalization	*ADR*#1:Vomiting, *ADR*#2:Pain
Sentence 5	It says that after application, there may be a [burning sensation]$_{ADR\#1}$, that supposedly this drug began to act on the [fungus]$_{Indication\#1}$
Gold normalization	*ADR*#1:Burning sensation, *Indication*#1:Fungal skin infection
Predicted normalization	*ADR*#1:Burning sensation, *Indication*#1:Fungal infection
Sentence 6	I feel [dizzy]$_{ADR\#1}$, [nauseous]$_{ADR\#1}$, [my vision loses focus]$_{ADR\#3}$, I have been lying down for 3 days and [I can't walk]$_{ADR\#4}$
Gold normalization	*ADR*#1:Dizziness, *ADR*#2:Nausea, *ADR*#3:Vision blurred, *ADR*#4:Asthenia
Predicted normalization	*ADR*#1:Dizziness, *ADR*#2:Nausea, *ADR*#3:Vision blurred, *ADR*#4:Walking disability
Sentence 7	On the 4th day, [itching on the skin of the face and around the eyes]$_{ADR\#1}$ appeared.
Gold normalization	*ADR*#1:Eye pruritus
Predicted normalization	*ADR*#1:Pruritus
Sentence 8	[left side hurt]$_{ADR\#1}$ after taking the pill
Gold normalization	*ADR*#1:Flank pain
Predicted normalization	*ADR*#1:Abdominal pain upper
Sentence 9	On the second day of admission, a [very severe headache began]$_{ADR\#1}$, [pain in the eyes]$_{ADR\#2}$, [nausea]$_{ADR\#3}$, [ripples in the eyes]$_{ADR\#4}$, [dizziness]$_{ADR\#5}$
Gold normalization	*ADR*#1:Headache, *ADR*#2:Eye pain, *ADR*#3:Nausea, *ADR*#4:Visual snow syndrome, *ADR*#5:Dizziness
Predicted normalization	*ADR*#1:Headache, *ADR*#2:Eye pain, *ADR*#3:Nausea, *ADR*#4:Visual impairment, *ADR*#5:Dizziness

7. Conclusions

This paper presents the current level of accuracy of normalizing adverse reactions mentions written in free form to concepts of the MedDRA dictionary of regulatory activity. Its estimation was obtained on the corpus of Russian-language reviews of Internet users — Russian Drug Review Corpora — which contains expert annotations for a wide range of tasks for extracting significant information about the use of medicines, including normalization tasks.

The evaluation is based on a neural network, which was built using the metrics learning approach. The highest accuracy of solving the problem is achieved on the basis of a combination of language models to vectorize mentions of adverse reactions and MedDRA concepts in the overall solution composition. We proposed a training procedure for the target task based on pre-training on a set of examples from English corpora, which had been automatically translated into Russian.

It is shown that the level of accuracy essentially depends on the dataset formation, primarily the composition of PTs. Thus, on the initial ADR dataset from RDRS, the obtained accuracy is 70.9% for *F*1-micro and 84.8% on the dataset from CADEC.

However, on the subpart of RDRS with the same subset of MedDRA PT concepts as in CADEC, the accuracy reaches 87.5% for *F*1-micro, which is close to the accuracy of 89.4% for *F*1-micro, reached on the CADEC subset with the same PT concepts as in RDRS. This conducted investigation points out the strong necessity to control the consistency of compared datasets on characteristics. It seems necessary to expand a part of the RDRS corpus with ADR normalization by mentions of the given PT concept set to achieve a higher level of accuracy.

Author Contributions: Conceptualization, A.S. (Alexander Sboev) and R.R.; methodology, A.S. (Alexander Sboev) and I.M.; software, A.S. (Anton Selivanov), G.R., I.M. and A.G.; validation, G.R. and A.S. (Anton Selivanov); investigation, A.S. (Alexander Sboev), A.S. (Anton Selivanov), G.R. and I.M.; resources, A.S. (Alexander Sboev) and R.R.; data curation, A.S. (Alexander Sboev), S.S. and A.G.; writing—original draft preparation, R.R., A.S. (Anton Selivanov) and A.S. (Alexander Sboev); writing—review and editing, A.S. (Alexander Sboev), R.R. and A.S. (Anton Selivanov); visualization, A.G., I.M. and G.R.; supervision, A.S. (Alexander Sboev); project administration, R.R.; funding acquisition, A.S. (Alexander Sboev). All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Russian Science Foundation grant No. 20-11-20246.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Trained models are presented on the page of our team on the huggingface repository: https://huggingface.co/sagteam (accessed on 12 November 2022). The code is available at https://github.com/sag111/MedNorm (accessed on 12 November 2022). RDRS corpora can be obtained through sending a request from the website of our project: https://sagteam.ru/en/med-corpus/ (accessed on 12 November 2022).

Acknowledgments: The study was supported by a grant from the Russian Science Foundation (project no. 20-11-20246). This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC "Kurchatov Institute", http://ckp.nrcki.ru/ (accessed on 28 November 2022).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Rezaei, Z.; Ebrahimpour-Komleh, H.; Eslami, B.; Chavoshinejad, R.; Totonchi, M. Adverse drug reaction detection in social media by deep learning methods. *Cell J.* **2020**, *22*, 319. [PubMed]
2. Huynh, T.; He, Y.; Willis, A.; Rüger, S. *Adverse Drug Reaction Classification with Deep Neural Networks*; Coling: Gyeongju, Republic of Korea, 2016.
3. Fan, B.; Fan, W.; Smith, C.; Garber, H. Adverse drug event detection and extraction from open data: A deep learning approach. *Inf. Process. Manag.* **2020**, *57*, 102131. [CrossRef]
4. El-allaly, E.d.; Sarrouti, M.; En-Nahnahi, N.; El Alaoui, S.O. MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. *Inf. Process. Manag.* **2021**, *58*, 102473. [CrossRef]
5. Sboev, A.; Sboeva, S.; Moloshnikov, I.; Gryaznov, A.; Rybka, R.; Naumov, A.; Selivanov, A.; Rylkov, G.; Ilyin, V. Analysis of the Full-Size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models. *Appl. Sci.* **2022**, *12*, 491. [CrossRef]
6. Nishioka, S.; Watanabe, T.; Asano, M.; Yamamoto, T.; Kawakami, K.; Yada, S.; Aramaki, E.; Yajima, H.; Kizaki, H.; Hori, S. Identification of hand-foot syndrome from cancer patients' blog posts: BERT-based deep-learning approach to detect potential adverse drug reaction symptoms. *PLoS ONE* **2022**, *17*, e0267901. [CrossRef] [PubMed]
7. Bekoulis, G.; Deleu, J.; Demeester, T.; Develder, C. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.* **2018**, *114*, 34–45. [CrossRef]
8. Yang, X.; Bian, J.; Fang, R.; Bjarnadottir, R.I.; Hogan, W.R.; Wu, Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 65–72. [CrossRef] [PubMed]
9. Sboev, A.; Selivanov, A.; Moloshnikov, I.; Rybka, R.; Gryaznov, A.; Sboeva, S.; Rylkov, G. Extraction of the Relations among Significant Pharmacological Entities in Russian-Language Reviews of Internet Users on Medications. *Big Data Cogn. Comput.* **2022**, *6*, 10. [CrossRef]
10. Mohan, S.; Angell, R.; Monath, N.; McCallum, A. Low resource recognition and linking of biomedical concepts from a large ontology. In Proceedings of the Proceedings of the 12th ACM conference on Bioinformatics, Computational Biology, and Health Informatics, Gainesville, FL, USA, 1–4 August 2021; pp. 1–10.
11. Manousogiannis, E.; Mesbah, S.; Bozzon, A.; Sips, R.J.; Szlanik, Z.; Baez, S. Normalization of Long-tail Adverse Drug Reactions in Social Media. In Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Online, 20 November 2020 ; pp. 49–58.
12. Kalyan, K.S.; Sangeetha, S. Target concept guided medical concept normalization in noisy user-generated texts. In Proceedings of the Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Online, 19 November 2020 ; pp. 64–73.
13. Kalyan, K.S.; Sangeetha, S. Medical concept normalization in user generated texts by learning target concept embeddings. *arXiv* **2020**, arXiv:2006.04014.
14. Pattisapu, N.; Patil, S.; Palshikar, G.; Varma, V. Medical concept normalization by encoding target knowledge. In Proceedings of the Machine Learning for Health Workshop, Online, 11 December 2020 ; pp. 246–259.
15. Segura-Bedmar, I.; Martínez, P. Pharmacovigilance through the development of text mining and natural language processing techniques. *J. Biomed. Inform.* **2015**, *58*, 288–291. [CrossRef]
16. MedDRA, M. *Introductory Guide for Standardised MedDRA Queries (SMQs), Version 21.0*; International Federation of Pharmaceutical Manufacturers and associations (IFPMA): Geneva, Switzerland, 2018 .
17. Karimi, S.; Metke-Jimenez, A.; Kemp, M.; Wang, C. Cadec: A corpus of adverse drug event annotations. *J. Biomed. Inform.* **2015**, *55*, 73–81. [CrossRef] [PubMed]
18. Roberts, K.; Demner-Fushman, D.; Tonning, J.M. Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. In Proceedings of the Text Analysis Conference, Gaithersburg, MD, USA, 17–18 November 2017.
19. Sarker, A.; Gonzalez-Hernandez, G. Overview of the second social media mining for health (SMM4H) shared tasks at AMIA 2017. *Training* **2017**, *1*, 1239.
20. Ristad, E.S.; Yianilos, P.N. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 522–532. [CrossRef]
21. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium, Washington, DC, USA, 3–7 November 2001; American Medical Informatics Association: Bethesda, MA, USA, 2001; p. 17.
22. McCallum, A.; Bellare, K.; Pereira, F. A conditional random field for discriminatively-trained finite-state string edit distance. *arXiv* **2012**, arXiv:1207.1406.
23. Sarker, A.; Belousov, M.; Friedrichs, J.; Hakala, K.; Kiritchenko, S.; Mehryary, F.; Han, S.; Tran, T.; Rios, A.; Kavuluru, R.; et al. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1274–1283. [CrossRef] [PubMed]
24. Limsopatham, N.; Collier, N. Normalising medical concepts in social media texts by learning semantic representation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016 ; pp. 1014–1023.

25. Lee, K.; Hasan, S.A.; Farri, O.; Choudhary, A.; Agrawal, A. Medical concept normalization for online user-generated texts. In Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, USA, 23–26 August 2017 ; IEEE: Piscataway, NJ, USA, 2017; pp. 462–469.
26. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
27. Mondal, I.; Purkayastha, S.; Sarkar, S.; Goyal, P.; Pillai, J.; Bhattacharyya, A.; Gattu, M. Medical entity linking using triplet network. *arXiv* **2020**, arXiv:2012.11164.
28. Ji, Z.; Wei, Q.; Xu, H. Bert-based ranking for biomedical entity normalization. *AMIA Summits Transl. Sci. Proc.* **2020**, *2020*, 269. [PubMed]
29. Zolnoori, M.; Fung, K.W.; Patrick, T.B.; Fontelo, P.; Kharrazi, H.; Faiola, A.; Shah, N.D.; Wu, Y.S.S.; Eldredge, C.E.; Luo, J.; et al. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data Brief* **2019**, *24*, 103838. [CrossRef]
30. Weissenbacher, D.; Sarker, A.; Paul, M.; Gonzalez, G. Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In Proceedings of the 2018 EMNLP workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task, Brussels, Belgium, 31 October 2018; pp. 13–16.
31. Weissenbacher, D.; Sarker, A.; Magge, A.; Daughton, A.; O'Connor, K.; Paul, M.; Gonzalez, G. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task, Florence, Italy, 1–2 August 2019 ; pp. 21–30.
32. Klein, A.; Alimova, I.; Flores, I.; Magge, A.; Miftahutdinov, Z.; Minard, A.L.; O'connor, K.; Sarker, A.; Tutubalina, E.; Weissenbacher, D.; et al. Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020. In Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, Online, 12 December 2020 ; pp. 27–36.
33. Magge, A.; Klein, A.; Miranda-Escalada, A.; Ali Al-Garadi, M.; Alimova, I.; Miftahutdinov, Z.; Farre, E.; Lima López, S.; Flores, I.; O'Connor, K.; et al. Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task, Mexico City, Mexico, 10 June 2021 ; Association for Computational Linguistics: Mexico City, Mexico, 2021; pp. 21–32. [CrossRef]
34. Magge, A.; Tutubalina, E.; Miftahutdinov, Z.; Alimova, I.; Dirkson, A.; Verberne, S.; Weissenbacher, D.; Gonzalez-Hernandez, G. DeepADEMiner: A deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 2184–2192. [CrossRef] [PubMed]
35. Shelmanov, A.; Smirnov, I.; Vishneva, E. Information extraction from clinical texts in Russian. In Proceedings of the Computational Linguistics and Intellectual Technologies: Annual International Conference "Dialog", Moscow, Russia, 27–30 May 2015; Volume 17.
36. Tutubalina, E.; Alimova, I.; Miftahutdinov, Z.; Sakhovskiy, A.; Malykh, V.; Nikolenko, S. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics* **2020**, *37*, 246–249. [CrossRef]
37. Sboev, A.; Sboeva, S.; Gryaznov, A.; Evteeva, A.; Rybka, R.; Silin, M. A neural network algorithm for extracting pharmacological information from Russian-language internet reviews on drugs. *J. Physics: Conf. Ser.* **2020**, *1686*, 012037. [CrossRef]
38. Sboev, A.; Moloshnikov, I.; Selivanov, A.; Rylkov, G.; Rybka, R. The Two-Stage Algorithm for Extraction of the Significant Pharmaceutical Named Entities and Their Relations in the Russian-Language Reviews on Medications on Base of the XLM-RoBERTa Language Model. In Proceedings of the Biologically Inspired Cognitive Architectures Meeting, Guadalajara, Mexico, 22–25 September 2022 ; Springer: Berlin/Heidelberg, Germany, 2022; pp. 463–471.
39. Sboev, A.; Rylkov, G.; Rybka, R.; Gryaznov, A.; Sboeva, S. Data-driven model for identifying related pharmaceutically-significant entities in clinical texts. In Proceedings of International Conference on Numerical Analysis and Applied Mathematics, Rome, Italy, 15–16 December 2022; AIP Publishing LLC: Melville, NY, USA, 2022; Volume 2425, p. 340003.
40. Kuratov, Y.; Arkhipov, M. Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv* **2019**, arXiv:1905.07213.
41. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
42. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE International Conference on Computer Vision, Cambridge, MA, USA, 20–23 June 1995; pp. 19–27.
43. Hamborg, F.; Meuschke, N.; Breitinger, C.; Gipp, B. news-please: A Generic News Crawler and Extractor. In Proceedings of the 15th International Symposium of Information Science, Zadar, Croatia, 19-21 May 2017; pp. 218–223. [CrossRef]
44. Gokaslan, A.; Cohen, V. OpenWebText Corpus. 2019. Available online: http://Skylion007.github.io/OpenWebTextCorpus (accessed on 28 November 2022) .
45. Trinh, T.H.; Le, Q.V. A simple method for commonsense reasoning. *arXiv* **2018**, arXiv:1806.02847.
46. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. *arXiv* **2015**, arXiv:1508.05326.
47. Williams, A.; Nangia, N.; Bowman, S.R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv* **2017**, arXiv:1704.05426.