applied sciences

MDPI

Article

Analysis of the Full-Size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models

Alexander Sboev [1,2,*], Sanna Sboeva [1], Ivan Moloshnikov [1], Artem Gryaznov [1], Roman Rybka [1], Alexander Naumov [1], Anton Selivanov [1], Gleb Rylkov [1] and Vyacheslav Ilyin [1,3,4]

1 Complex of NBICS Technologies, National Research Centre "Kurchatov Institute", 123182 Moscow, Russia; Sboeva_SG@rrcki.ru (S.S.); Moloshnikov_IA@nrcki.ru (I.M.); Gryaznov_AV@nrcki.ru (A.G.); Rybka_RB@nrcki.ru (R.R.); Naumov-AV@nrcki.ru (A.N.); Selivanov_AA@nrcki.ru (A.S.); Rylkov_GV@rrcki.ru (G.R.); ilyin_va@nrcki.ru (V.I.)
2 Moscow Engineering Physics Institute, National Research Nuclear University, 115409 Moscow, Russia
3 National Center for Cognitive Research, ITMO University, 197101 Saint Petersburg, Russia
4 Department of NBIC-Technologies, Moscow Institute of Physics and Technology, 141701 Dolgoprudny, Russia
* Correspondence: Sboev_AG@nrcki.ru

Abstract: The paper presents the full-size Russian corpus of Internet users' reviews on medicines with complex named entity recognition (NER) labeling of pharmaceutically relevant entities. We evaluate the accuracy levels reached on this corpus by a set of advanced deep learning neural networks for extracting mentions of these entities. The corpus markup includes mentions of the following entities: medication (33,005 mentions), adverse drug reaction (1778), disease (17,403), and note (4490). Two of them—medication and disease—include a set of attributes. A part of the corpus has a coreference annotation with 1560 coreference chains in 300 documents. A multi-label model based on a language model and a set of features has been developed for recognizing entities of the presented corpus. We analyze how the choice of different model components affects the entity recognition accuracy. Those components include methods for vector representation of words, types of language models pre-trained for the Russian language, ways of text normalization, and other pre-processing methods. The sufficient size of our corpus allows us to study the effects of particularities of annotation and entity balancing. We compare our corpus to existing ones by the occurrences of entities of different types and show that balancing the corpus by the number of texts with and without adverse drug event (ADR) mentions improves the ADR recognition accuracy with no notable decline in the accuracy of detecting entities of other types. As a result, the state of the art for the pharmacological entity extraction task for the Russian language is established on a full-size labeled corpus. For the ADR entity type, the accuracy achieved is 61.1% by the F1-exact metric, which is on par with the accuracy level for other language corpora with similar characteristics and ADR representativeness. The accuracy of the coreference relation extraction evaluated on our corpus is 71%, which is higher than the results achieved on the other Russian-language corpora.

Keywords: pharmacovigilance; annotated corpus; adverse drug events; social media; UMLS; MESHRUS; information extraction; machine learning; neural networks; deep learning; named entity recognition; coreference relation extraction; language models

Citation: 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. https://doi.org/10.3390/app12010491

Academic Editor: Jianbo Gao

Received: 25 November 2021
Accepted: 30 December 2021
Published: 4 January 2022

1. Introduction

Nowadays, Internet sources contain a vast variety of information subject to automated analysis by means of machine learning methods, the usage of which allows one to solve various socially significant tasks [1,2]. In particular, such information is related to healthcare in general, consumption sphere and evaluation of medicines by the population. Clinical trials may not reveal all potential adverse effects of a medicine due to time limitations.

This is a very serious problem in healthcare (See, e.g., the decision of the Council of the Eurasian Economic Commission No. 87 of 3 November 2016 "On Approving the Rules of Good Practice for PV of the Eurasian Economic Union"). Therefore, after a pharmaceutical product enters the market, pharmacovigilance (PV) is of great importance.

Patients' opinions expressed on the Internet, in particular in social networks, discussion groups and forums, may contain a considerable amount of information that would supplement clinical investigations in evaluating the efficiency of a medicine. Internet posts often describe adverse reactions in real time ahead of official reporting, or reveal unique characteristics of undesirable reactions that differ from the data of health professionals. Moreover, patients openly discuss a variety of ways they use medicines to treat different diseases, including "off-label" applications. Such information would be very useful for a PV database where the risks and advantages of drugs would be registered for the purpose of safety monitoring, as well as for forming hypotheses of using existing drugs for treating other diseases.

This leads to an increasing need for the analysis of Internet information to assess the quality of medical care and drug provision. In this regard, one of the main tasks is the development of machine learning methods for extracting useful information from social media. These methods have to be developed taking into account the presence of informal vocabulary and reasoning in such texts.

The quality of these methods directly depends on annotated corpora for their training. In this paper, we present the full-size Russian-language corpus of Internet user reviews, named Russian Drug Reviews corpus of the SagTeam project (RDRS) (the description of the corpus is presented at https://sagteam.ru/en/med-corpus/, accessed on 12 December 2021). The corpus comprises a complex annotation for named entity recognition (NER) and a part with coreference relation annotation. Moreover, we present a deep learning neural network model for automated extraction of mentions of the various types of entities present in our corpus. The model is based on the XLM-RoBERTa-large language model with a set of additional input features, and is capable of multi-tag labeling.

In Section 2, we analyze existing works in the domain of this research. In particular, we describe and compare the compositions of the existing corpora containing adverse drug reaction (ADR) annotation on the different text types, entity types, text sizes and styles (see Section 2.1). Moreover, we consider these corpora for the purpose of establishing correspondences between marked entities and concepts in the thesauri accepted for this area (see Section 2.2). In Section 2.3, we compare these corpora based on the complexity of annotated entities, so as to analyze the influence of such characteristics on the ADR extraction accuracy. In Section 2.4, we review methods of named entity recognition that can be built on these corpora, and in Section 2.5, we describe the status of the task of coreference relations extraction. Section 3 presents our corpus: the materials used to collect the corpus are outlined in Section 3.1, and the technique of its annotation is described in Section 3.2. Then, we provide the statistics of the annotated corpus to understand its diversity. The developed machine learning system is presented in Section 4. We describe the recurrent model that uses the chosen set of features as an input, the process of training a large language model to adapt it for texts of the target domain and the combination of models that we are using as the final pipeline for named entity recognition. The conducted numerical experiments are presented in Section 5 and discussed in Section 6.

2. Related Works

Research concerning the above-mentioned problems is conducted intensively worldwide, resulting in a great diversity of annotated corpora. From the linguistic point of view, these corpora could be distinguished into two groups: corpora of texts written by medicine specialists (clinical reports with annotations), and those of texts written by non-specialists, namely, by the Internet customers who used the medications. The variability of natural language constructions in the speech of Internet users complicates the analysis of corpora based on Internet texts. There are also other distinctive features of any corpus that influence

the accuracy of entity recognition on its base: the types of entities annotated, the numbers of their joint use in phrases, the numbers of phrases mentioning entities of certain types and approaches to entity normalization. Moreover, the metrics used for evaluating the results may vary. Not for every corpus is such information available. Below, we briefly describe six corpora: CADEC, n2c2-2018, Twitter annotated corpus, PsyTAR, TwiMed corpus and RuDReC.

2.1. Existing Drug Review Corpora

2.1.1. Corpus of Adverse Drug Event Annotations (CADEC)

Ref. [3] is a corpus of medical posts taken from the AskAPatient (Ask a Patient: Medicine Ratings and Health Care Opinions, http://www.askapatient.com/, accessed on 12 December 2021) forum and annotated by four medical students and two computer scientists. It comprises 1253 posts with 7398 sentences, containing consumers' ratings and reviews on 13 different drugs. The following entities were annotated: drug, ADR, symptom, disease and findings. In order to coordinate the annotation, all annotators did the markup together for several texts, and after that, the remaining texts were distributed among them. All annotated texts were checked by the three corpus authors so as to correct obvious mistakes, e.g., missing letters, misprints, etc.

2.1.2. Twitter and PubMed Comparable Corpus of Drugs, Diseases, Symptoms and Their Relations (TwiMed)

Ref. [4] contains 1000 tweets (TwiMed Twitter) and 1000 sentences from Pubmed (National Center for Biotechnology Information webcite—http://www.ncbi.nlm.nih.gov/pubmed/, accessed on 12 December 2021) (TwiMed Pubmed) for 30 drugs. The corpus is composed of annotations approved by two pharmaceutical experts. Its markup contains 3144 entities, 2749 relations and 5003 attributes labeled. The entity types are drug, symptom and disease.

2.1.3. Twitter Annotated Corpus

Ref. [5] consists of randomly selected tweets containing drug name mentions: generic and brand names of the drugs. The annotators group includes pharmaceutical and computer experts. Two types of markup are currently available: binary and span, the former having texts labeled just by the presence or absence of ADRs, and the latter including mention boundaries. The binary-annotated part [6] consists of 10,822 tweets, of which 1239 (11.4%) contain ADR mentions and 9583 (88.6%) do not. The span-annotated part [5] consists of 2131 tweets (which include 1239 tweets containing ADR mentions from the binary-annotated part). The semantic types annotated are: ADR, beneficial effect, indication, other (medical signs or symptoms).

2.1.4. PsyTAR Dataset

Ref. [7] contains 891 reviews on 4 drugs, collected randomly from the AskAPatient forum. Before annotation, the texts were cleared (by means of regular expressions) of any personal information, such as emails, phone numbers and URLs. The annotators group included pharmaceutical students and experts. They marked the following set of entities: ADR, withdrawal symptoms (WD), sign/symptom/illness (SSI), drug indications (DI) and other. Unfortunately, the original corpus does not contain mention boundaries in its markup. This complicates the NER task. A paper, ref. [8] presented a version of the PsyTAR corpus in the CoNLL format, where every word has a corresponding named entity tag.

2.1.5. n2c2-2018

Ref. [9] is a dataset from the National NLP Clinical Challenge of the Department of Biomedical Informatics (DBMI) at Harvard Medical School. The dataset contains clinical narratives and is based on past medication extraction tasks but examines a broader set of patients, diseases and relations as compared with earlier challenges. It was annotated

by four paramedic students and three nurses. The label set includes medications and associated attributes, such as dosage, strength of the medication, administration mode, administration frequency, administration duration, reason for administration and drug-related adverse effects. The number of texts was 505,274 in training, 29 in development and 202 in the testing subset.

2.1.6. Russian Drug Reaction Corpus (RuDReC)

Ref. [10] is a Russian-language corpus, the labeled part of which contains 500 reviews on drugs from a consumer forum OTZOVIK. A two-step procedure was performed for its annotation: First, 400 texts were used that had been labeled in accordance with the format of the Sagteam project (https://sagteam.ru/en/med-corpus/annotation/, accessed on 12 December 2021) by 4 experts of Sechenov First Moscow State Medical University who are now participants of our projects. In the second step, the corpus authors reformed the labeling by deleting/uniting tags, and after that, annotated 100 more reviews. Overall, RuDReC and our proposed corpus RDRS have an intersection of 467 texts. The influence of differences in their labeling on the ADR extraction accuracy is presented in Section 6.

2.2. Target Vocabularies in the Corpora Normalization

The analysis of internet user texts is more difficult because of informal text style and more natural vocabulary. Consequently, when creating corpora, the labeled entities are assigned to concepts of a unified international dictionaries and thesauri. In particular, annotated entities in CADEC were mapped to controlled vocabularies: SNOMED CT, The Australian Medicines Terminology (AMT) [11] and MedDRA. Any span of text annotated with any tag was mapped to the corresponding vocabularies. If a concept did not exist in the vocabularies, it was assigned the "concept_less" tag. In the TwiMed corpus, for drug entities, the SIDER database [12] was used, which contains information on market medicines extracted from public documents, while for symptom and disease entities the MedDRA ontology was used. In addition, the terminology of SNOMED CT concepts was used for entities belonging to the Ddisorder semantic group. In the Twitter dataset [5], ADR mentions were set in accordance with their unified medical language system (UMLS) concept ID. Finally, in the PsyTAR corpus, ADR, WD, SSI and DI entities were matched to UMLS Metathesaurus concepts and SNOMED CT concepts. Concerning the n2c2-2018 corpus, no normalization was applied to it.

2.3. Number of Entities and Their Proportions in the Corpora

In Table 1, we review the complexity characteristics of the existing corpora described above and evaluate the influence of these characteristics on the ADR extraction accuracy. For the TwiMed Twitter and TwiMed PubMed corpora, by ADRs we meant, following the article [13], symptoms related to drugs.

Only a few of the considered corpora contain overlapping entities, but their proportions are relatively small, except for CADEC, where there are parts of overlapping ADR entities, both continuous (5%), and discontinuous (9%). In this sense, CADEC appears to be the most complicated corpus from the considered set; this fact impedes ADR extraction. On the other hand, it has the largest absolute number of ADR mentions and the largest ratio of ADRs to symptoms, which positively affects the accuracy of their extraction.

We were unable to find the information about the ADR identification precision by the F1-exact metric for all corpora. However, on the basis of Table 1, we suggest a parameter that could be convenient for comparing the corpora. It is the fraction of the ADR mentions number to the total number of words in the corpus, and we use it further named as *saturation*.

Table 1. Comparison of structural characteristics of existing corpora with respect to ADR mentions, and the accuracy of ADR detection in these corpora. Abbreviations of corpora names: TA—Twitter Annotated Corpus, TT—TwiMED Twitter, TP—TwiMED PubMed, N2C2—n2c2-2018. Abbreviations of accuracy metrics: f1-e—f1-exact, f1-am—f1-approximate match, f1-r—f1-relaxed, f1-cs—sentence classification on ADR entity presence; NA—data not available for download and analysis.

Corpus	CADEC	TA	TT	TP	N2C2	PSYTAR	RuDRec
Total number of mentions	6318	1122	899	475	1579	3543	720
Multi-word (%)	72.4	0.47	40	46.7	42	78	54
Single-word (%)	27.6	0.53	60	53.3	58	22	46
Discontinuous, non-overlapping (%)	1.3	0	0	0	0	0	0
Continuous, non-overlapping (%)	84	100	98	96.8	95	100	100
Discontinuous, overlapping (%)	9.3	0	0	0	0	0	0
Continuous, overlapping (%)	5.3	0	2	3.2	5	0	0
$\text{Saturation} = \frac{\text{total ADR}}{\text{total words in corpus}} (\cdot 10^3)$	53.38	NA	NA	16.5	1.35	39.17	10.61
$\frac{\text{total ADR}}{\text{total entities number}}$	0.69	0.72	0.67	0.47	0.02	0.70	0.41
$\frac{\text{total ADR}}{\text{number of indication, reason, etc.}}$	22.97	7.1	1.91	0.49	0.25	0.70	0.01
Accuracy	70.6 [14]	61.1 [15]	64.8 [16]	73.6 [17]	55.8 [18]	71.1 (see Appendix A)	60.4 [10]
Accuracy metric	f1-e	f1-am	f1-am	f1-cs	f1-r	f1-e	f1-e

2.4. Named Entity Recognition and Classification Methods

There are two main approaches for named entity recognition. The first is based on feature engineering and using recurrent neural networks [19]. The second uses deep learning language models to encode input text, and simple output layers for token classification. A few state-of-the-art methods for named entity recognition in social media texts were tested in the recent shared task #SMM4H [20–25]. Most of them utilize deep learning language models like ELMo [26] or transformer-based BERT [27]. Such models can extract high-level features for tokens of input text and encode words with real valued vectors that can be utilized by neural networks to detect tokens of the entities of interest. However, achieving the best performance requires adapting the language model to the domain-specific texts by pre-training. Additional manually constructed features such as external vocabularies are usually useful in tasks with a specific lexicon. In our previous paper [28], we compared the usage of up-to-date language models for the NER task on several English corpora and demonstrated that the XLM-RoBERTa model achieved the best accuracy. We therefore use that model in this work.

2.5. Coreference Resolution

There is a problem that some reviews present users' opinions about more than one real-world entity, for example, reports about the use of multiple medications that may have different effects. Therefore, in order to distinguish mentions of different drugs, diseases, etc., it would be useful to detect which mentions, on the contrary, refer to the same entities, and which coreference resolution does.

For the English language, there are several corpora for coreference resolution, such as CoNLL-2012 [29] or GAP [30], and even a corpus of pharmacovigilance records with ADR annotations that includes coreference annotation (PHAEDRA) [31]. For Russian texts, the coreference problem is underrepresented in the literature. Currently, there are only two corpora with coreference annotations for the Russian language: RuCor [32] and corpus from the shared task AnCor-2019 [33]. The latter is a continuation and extension of the first.

As for the methods for coreference resolution, the state-of-the-art approach is based on a neural network trained end-to-end to solve two tasks at the same time: mention extraction

and relation extraction. This approach was firstly introduced in [34] and has been used in several papers [35–39], with some modifications to get higher scores on the coreference corpus CoNLL-2012 [29].

3. Collection of the Corpus

3.1. Corpus Material

In this section, we present the design of our corpus. It is based on 2800 reviews from a medical section of the forum called Otzovik (OTZOVIK, Internet forum of user reviews: http://otzovik.com, accessed on 12 December 2021), which is dedicated to consumer reviews on medications. On that website, there is a section where users submit posts by filling special survey forms. The site offers two forms: simplified and extended, the latter being optional. In this form, a user selects a drug name and fills out the information about the drug, such as: adverse effects experienced, comments, positive and negative sides, satisfaction rate and whether they would recommend the medicine to friends. In addition, the extended form contains prices, frequency, scores on a five-point scale for such parameters as quality, packing, safety and availability. We used information only from the simplified form since the users had rarely filled the extended forms in their reviews. We considered only the fields Heading, General impression and Comment.

A sample post for "Глицин" (Glycine) is shown in Table 2. The reviews are written in colloquial language, and do not necessarily follow formal grammar and punctuation rules. Moreover, sometimes the consumers describe not only their personal experience, but opinions of their family members, friends or others.

Table 2. A sample post for "Глицин" (Glycine) from otzovik.com. Original text is quoted, and followed by English translation in parentheses.

Overall Impression	"Помог чересчур!" (Helped too much!)
Advantages	"Цена" (Price)
Disadvantages	"отрицательно действует на работоспособность" (It has a negative effect on productivity)
Would you Recommend It to Friends?	"Нет" (No)
Comments	"Начала пить недавно. Прочитала отзывы вроде все хорошо отзывались. Стала спокойной даже чересчур, на работе стала тупить, коллеги сказали что я какая то заторможенная, все время клонит в сон. Буду бросать пить эти таблетки." (I started taking recently. I read the reviews, and they all seemed positive. I became calm, even too calm, I started to blunt at work, colleagues said that I somewhat slowed down, feel sleepy all the time. I will stop taking these pills.)

3.2. Corpus Annotation

This section describes the corpus annotation methodology, including the markup structure, the annotation procedure with guidelines for complex cases and software infrastructure for the annotation.

3.2.1. Annotation Procedure

Mention labeling for the review texts has been performed by a group of four annotators using a guide developed jointly by machine learning experts and pharmacists. Two annotators were qualified pharmacists, and the two others were students with pharmaceutical education. Reliability was achieved through joint work of annotators on the same set of documents subsequently controlled with logging. After the initial annotation round, the annotations were corrected three times with cross-checking by different annotators, after which the final decision was made by an expert pharmacist. The corpus annotation comprised the following steps:

1. First, a guide was compiled for the annotators. It included the description of the entities and corresponding examples.
2. Upon testing on a set of 300 reviews, the guide was corrected, addressing complex cases. During that, iterative annotation was performed, from one to five iterations for a text, while tracking for each text and each iteration of the annotator's questions, controller's comments and correction status.
3. The resulting guide was used during the annotation of the remaining reviews. Two annotators marked up each review, and then a pharmacist checked the result. Complex cases found during the process were analyzed separately by the whole group of experts.
4. The obtained markup was automatically checked for any possible inaccuracies, such as incomplete fragments of words selected as mentions, terms marked differently in different reviews, etc. Texts with such inaccuracies were rechecked.

Inter-annotator agreement has been estimated using the metric described by Karimi et al. [3]. According to this metric, we calculated the agreement score of a pair of annotators i and j for every document as the ratio of the number of matching mentions to the maximum number of mentions labeled by one of the annotators in the current document:

$$\text{agreement}(i,j) = 100\frac{\text{match}(A_i, A_j, \alpha, \beta)}{\max(|A_i|, |A_j|)}.$$

Here, A_i and A_j denote the lists of mentions labeled by annotators i and j. $|A_i|$ and $|A_j|$ stand for the numbers of elements in these lists. Counting the matching mentions was performed in four ways, depending on two parameters: span strictness α and tag strictness β. Span strictness can be *strict* or *intersection*. In the strict spans comparison, only mentions with equal borders will be counted as matching, otherwise we count mentions as matching if they at least intersect each other (but a mention cannot match more than one mention of another annotator). Tag strictness can be *strict* if we count matching mentions only when both annotators label them with the same tag, or *ignored* otherwise. Then, the total pair-wise agreement score for each pair of annotators was averaged over all documents, and finally, averaged over all pairs of annotators. The average inter-annotator agreement is presented in Table 3.

Table 3. Average pair-wise agreement between annotators.

Span Strictness, α	Tag Strictness, β	Agreement
strict	strict	61%
strict	ignored	63%
intersection	strict	69%
intersection	ignored	71%

The annotation was carried out with the help of the WebAnno-based toolkit, which is an open source project under the Apache License v2.0. It has a web interface and offers a set of annotation layers for different levels of analysis. The annotators acted by the guidelines below.

3.2.2. Guidelines Applied in the Course of Annotation

The objects of annotation are attributes of drugs, diseases (including their symptoms) and undesirable reactions to those drugs. The annotators were to label mentions of these three entity types with their attributes defined below.

Medication

This entity type includes everything related to the mentions of drugs and drug manufacturers. Selecting a mention of such entity, an annotator had to specify an attribute out

of those listed in Table 4, thereby annotating it, for instance, as a mention of the attribute DrugName of the entity type medication. In addition, the attributes DrugName and MedMaker had sub-attributes based on the origin of the distributor and the manufacturer of the drug, respectively, domestic and foreign, that were labeled with the help of lookup in the State Registry of Medicinal Products [40].

Table 4. Attributes of the medication entity type.

DrugName	Marks a mention of a drug. For example, in the sentence «Препарат Aventis "Трентал" для улучшения мозгового кровообращения» (The Aventis "Trental" drug to improve cerebral circulation), the word "Trental" (without quotation marks) is marked as a DrugName. This attribute has two sub-attributes, DrugName/MedFromDomestic and DrugName/MedFromForeign, which are based on the origin of the drug distributor looked up in external sources.
DrugBrand	A drug name is also marked as DrugBrand if it is a registered trademark. For example, the word "Протефлазид" (Proteflazid) in the sentence «Противовирусный и иммунотропный препарат Экофарм "Протефлазид"» (The Ecopharm "Proteflazid" antiviral and immunotropic drug).
Drugform	Dosage form of the drug (ointment, tablets, drops, etc.). For example, the word "таблетки" (pills) in the sentence «Эти таблетки не плохие, если начать принимать с первых признаков застуды» (These pills are not bad if you start taking them since the first signs of a cold).
Drugclass	Type of drug (sedative, antiviral agent, sleeping pill, etc.). For example, in the sentence «Противовирусный и иммунотропный препарат Экофарм "Протефлазид"» (The Ecopharm "Proteflazid" antiviral and immunotropic drug), two mentions marked as Drugclass: "Противовирусный" (Antiviral) and "иммунотропный" (immunotropic).
MedMaker	The drug manufacturer. For example, the words "Материа медика" (Materia Medica) in the sentence «Седативный препарат Материа медика "Тенотен"» (The Materia Medica "Tenoten" sedative). This attribute has two sub-attributes: MedMaker/Domestic and MedMaker/Foreign.
Frequency	The drug usage frequency. For example, the phrase "2 раза в день" (two times a day) in the sentence «Неудобство было в том, что его приходилось наносить 2 раза в день» (Its inconvenience was that it had to be applied two times a day).
Dosage	The drug dosage (including units of measurement, if specified). For example, in the sentence «Ректальные суппозитории "Виферон" 15000 МЕ—эффекта ноль» (Rectal suppositories "Viferon" 150000 IU have zero effect), the mention "15000 ME" (150000 IU) is marked as Dosage.
Duration	This entity specifies the duration of use. For example, "6 лет" (6 years) in the sentence «Время использования: 6 лет».
Route	Administration method (how to use the drug). For example, the words "можно готовить раствор небольшими порциями" (can prepare a solution in small portions) in the sentence «удобно то, что можно готовить раствор небольшими порциями» (it is convenient that one can prepare the solution in small portions).
SourceInfodrug	The source of information about the drug. For example, the words "посоветовали в аптеке" (recommended to me at a pharmacy) in the sentence «Этот спрей мне посоветовали в аптеке в его состав входят такие составляющие вещества как мята» (This spray was recommended to me at a pharmacy, it includes such ingredient as mint).

Disease

This entity type is associated with diseases or symptoms. It indicates the reason for taking a medicine, the name of the disease and improvement or worsening of the patient's state after taking the drug. Attributes of this entity are specified in Table 5.

Table 5. Attributes of the disease entity type.

Diseasename	The name of a disease. If a report author mentions the name of the disease for which they take a medicine, it is annotated as a mention of the attribute Diseasename. For example, in the sentence «у меня вчера была диарея» (I had diarrhea yesterday) the word "диарея" (diarrhea) will be marked as Diseasename. If there are two or more mentions of diseases in one sentence, they are annotated separately. In the sentence «Обычно весной у меня сезон аллергии на пыльцу и депрессия» (In spring I usually have season allergy to pollen, and depression), both "аллергия" (allergy) and "депрессия" (depression) are independently marked as Diseasename.
Indication	Indications for use (symptoms). In the sentence «У меня постоянный стресс на работе» (I have a permanent stress at work), the word "стресс" (stress) is annotated as Indication. Moreover, in the sentence «Я принимаю витамин С для профилактики гриппа и простуды» (I take vitamin C to prevent flu and cold), the entity "для профилактики" (to prevent) is annotated as Indication too. For another example, in the sentence «У меня температура 39.5» (I have a temperature of 39.5) the words "температура 39.5" (temperature of 39.5) are marked as Indication.
BNE-Pos	This entity specifies positive dynamics after or during taking the drug. In the sentence «препарат Тонзилгон Н действительно помогает при ангине» (the Tonsilgon N drug really helps a sore throat), the word "помогает" (helps) is the one marked as BNE-Pos.
ADE-Neg	Negative dynamics after the start or some period of using the drug. For example, in the sentence «Я очень нервничаю, купила пачку "персен", в капсулах, он не помог, а по моему наоборот всё усугубил, начала сильнее плакать и расстраиваться» (I am very nervous, I bought a pack of "persen", in capsules, it did not help, but in my opinion, on the contrary, everything aggravated, I started crying and getting upset more), the words "по моему наоборот всё усугубил, начала сильнее плакать и расстраиваться" (in my opinion, on the contrary, everything aggravated, I started crying and getting upset more) are marked as ADE-Neg.
NegatedADE	This entity specifies that the drug does not work after taking the course. For example, in the sentence «...боль в горле притупляют, но не лечат, временный эффект, хотя цена великовата для 18-ти таблеток» (...dulls the sore throat, but does not cure, a temporary effect, although the price is too big for 18 pills) the words "не лечат, временный эффект" (does not cure, the effect is temporary) are marked as NegatedADE.
Worse	Deterioration after taking a course of the drug. For example, in the sentence «Распыляла его в нос течении четырех дней, результата на меня не какого не оказал, слизистая еще больше раздражалось» (I sprayed my nose for four days, it didn't have any results on me, the mucosa got even more irritated), the words "слизистая еще больше раздражалось" (the mucosa got even more irritated) are marked as Worse.

ADR

This entity type is associated with adverse drug reactions: undesirable effects that a consumer relates to the usage of a medicine. For example, the word "судороги" ("cramp") in the sentence «После недели приема Кортексина у ребенка начались судороги» (After a week of taking Cortexin, the child began to cramp).

Note

We use this entity type for pharmaceutically meaningful entities that cannot be unequivocally assigned to any of the other entity types: when the author makes recommendations, tips and so on, but does not explicitly state whether the drug helps or not. These include phrases such as "I do not advise". For instance, the phrase «Нет поддержки для иммунной системы» (No support for the immune system) is annotated as a Note. Additionally labeled as Note are an author's subjective arguments instead of explicit reports on the outcomes. For example, "strange meds", "not impressed", "it is not clear whether it worked or not", "ambiguous effect" (example (d) in Figure 1). In borderline cases when the context of a phrase does not allow the annotator to decide unambiguously whether a phrase is an ADR mention, it is assigned both ADR and Note tags, and the influence of

including or excluding such ambiguous mentions into the resulting markup is analyzed later in Section 5.4.

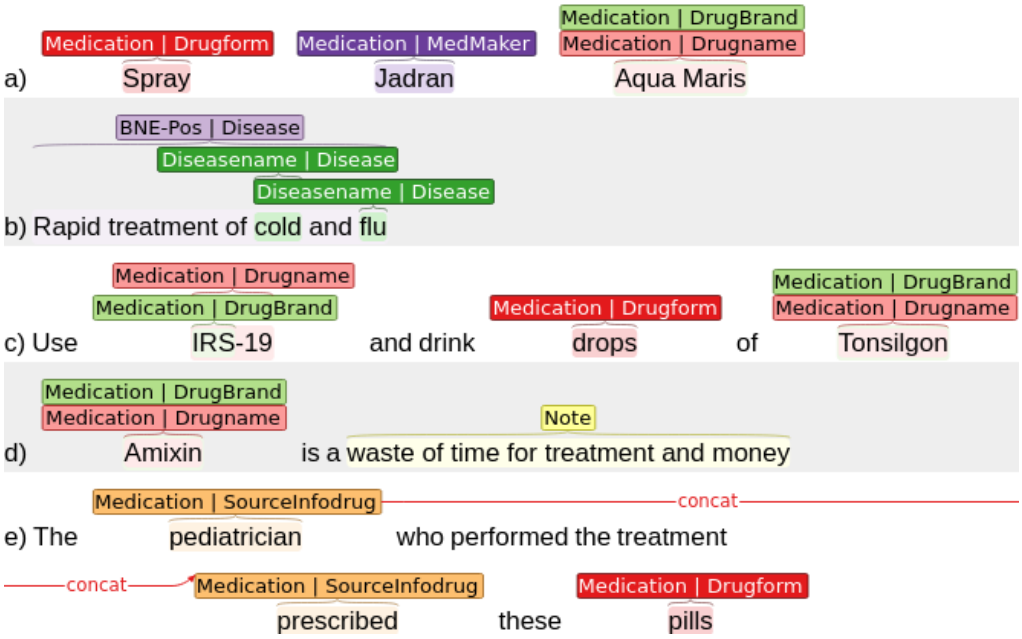

Figure 1. Examples of the text annotation from the corpus. Examples (**a**–**d**) depict intersecting annotations; example (**d**) depicts a mention of Note; example (**e**) depicts a discontinuous mention with concatenation relation.

By the complexity of their annotation, mentions can be divided into the following groups:

1. A simple markup: when a mention consists of one or more words and is related to a single attribute of entity. The annotators then just have to select a minimal but meaningful text fragment, excluding conjunctions, introductory words and punctuation marks.
2. Discontinuous annotation: when a mention is separated by words that do not belong to it. It is then necessary to annotate mention parts and connect them. In such cases, we use the "concatenation" relation. In the example (e) on Figure 1 "The pediatrician who performed the treatment prescribed these pills", the words "prescribed" and "pediatrician" are annotated as a concatenated parts of mention of the attribute SourceInfoDrug.
3. Intersecting annotations: words in a text can belong to mentions of different entities or attributes simultaneously. For example, in the sentence "Rapid treatment of cold and flu" (see Figure 1, example (b)), words "cold" and "flu" are mentions of attribute DiseaseName, but at the same time the whole phrase is a mention of attribute BNE-Pos. If a word or a phrase belongs to mentions of different attributes or entities at the same time (for example, DrugName and DrugBrand), it should be labeled with all of them: see, for instance, entity "Aqua Maris" in sentence "Spray Jadran Aqua Maris" (Figure 1, example (a)).

The percentages of such mentions for different entity types are presented in Table 6. An analysis of this table shows that the annotated entities differ greatly in word length and complex cases, in particular, the corpus contains a significant part of overlapping entities. This requires the development of an appropriate model for their effective recognition.

Table 6. Percentages of different types of mentions in the annotation of our corpus. A discontinuous mention consists of several labeled phrases separated by words not related to it. A mention is overlapping if some of its words are also labeled as another mention.

Entity Type	Total Mentions Count	Multi-Word (%)	Single-word (%)	Discontinuous, Non-Overlapping (%)	Continuous, Non-Overlapping (%)	Discontinuous, Overlapping (%)	Continuous, Overlapping (%)
ADR	1784	63.85	36.15	2.97	80.66	0.62	15.75
Drugname	8236	17.13	82.87	0	38.37	0.01	61.62
DrugBrand	4653	11.95	88.05	0	0	0.02	99.98
Drugform	5994	1.90	98.10	0	83.53	0.02	16.45
Drugclass	3120	4.42	95.58	0	94.33	0	5.67
Dosage	965	92.75	7.25	0.10	54.92	0.21	44.77
MedMaker	1715	32.19	67.81	0	99.71	0	0.29
Route	3617	34.95	65.05	0.53	88.80	0.06	10.62
SourceInfodrug	2566	48.99	51.01	6.16	91.00	0	2.84
Duration	1514	86.53	13.47	0.20	95.44	0	4.36
Frequency	614	98.96	1.14	0.33	88.93	0	10.75
Diseasename	4006	11.48	88.52	0.35	85.97	0.02	13.65
Indication	4606	43.88	56.12	1.13	77.49	0.30	21.08
BNE-Pos	5613	66.06	33.94	1.02	82.91	0.68	15.39
NegatedADE	2798	92.67	7.33	1.36	87.38	0.18	11.08
Worse	224	97.32	2.68	0.89	61.16	1.34	36.61
ADE-Neg	85	89.41	10.59	3.53	54.12	3.53	38.82
Note	4517	90.21	9.79	0.13	77.77	0.15	21.94

3.3. Classification Based on Categories of the Anatomical Therapeutic Chemical (ATC), ICD-10 Classifiers and MedDRA Terminology

After annotation, in order to resolve possible ambiguity in terms, we performed normalization and classification by matching the labeled mentions to information from external official classifiers and registers. The external sources for Russian are described below.

- The 10-th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) [41] is an international classification system for diseases which includes 22 classes of diagnoses, each consisting of up to 100 categories. ICD-10 allows us to reduce verbal diagnoses of diseases and health problems to unified codes.
- The Anatomical Therapeutic Chemical classification system (ATC) [42] is an international medication classification containing 14 anatomical main groups and 4 levels of subgroups. ICD-10 and ATC have a hierarchical structure, where "leaves" (terminal elements) are specified diseases or medications, and "nodes" are groups or categories. Every node has a code, which includes the code of its parent node.
- The State Registry of Medicinal Products ("Государственный реестр лекарственных средств, ГРЛС" in Russian) [40] is a registry of detailed information about the medications certified in the Russian Federation. It includes possible manufacturers, dosages, dosage forms, ATC codes, indications and so on.
- The Medical Dictionary for Regulatory Activities terminology (MedDRA®) is the international medical terminology developed under the auspices of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).

Among the international systems of standardization of concepts, the most complete and large metathesaurus is UMLS, which combines most of the databases of medical concepts and observations, including MeSH (and MESHRUS), ATC, ICD-10, SNOMED CT, LOINC and others. Every unique concept in UMLS has an identification code CUI, by which one can retrieve information about the concept from all the databases. However, within UMLS, it is only the MESHRUS database that contains the Russian language and can be used to associate words from our texts with CUI codes.

The classification was carried out by the annotators manually. For this purpose, we applied the procedure consisting of the following steps: automatic grouping of mentions, manual verification of mention groups (standardization) and matching the mention groups to the groups from ATC and ICD-10 or terms from MedDRA.

Automatic mention grouping was based on calculating the similarity between two mentions by the Ratcliff/Obershelp algorithm [43], which is based on searching two strings for matching substrings. In the course of the analysis, every new mention is added to one of the existing groups if the mean similarity between the mention and all the group items is more than 0.8 (value deduced empirically), otherwise a new group is created. The set of groups is empty at the start, and the first mention creates a new group with size 1. Each group is named by its most frequent mention. Next, the annotators manually check and refine the resulting set, creating larger groups or renaming them. Mentions of drug names are standardized according to the State Registry of Medicinal Products. This has given us 550 unique drug names mentioned in the corpus.

After that, the group names for the attributes DiseaseName, DrugName and DrugClass were manually matched with the term codes from the ICD-10 and ATC classifiers. As a result, 247 unique ICD-10 codes have been matched against the 765 unique phrases labeled as attribute DiseaseName; 226 unique ATC codes matched the 550 unique drug names and 70 unique ATC codes corresponded to 414 unique DrugClass mentions. Some drug classes mentioned in the corpus (such as homeopathy) do not have a corresponding ATC code, and are aggregated according to their anatomical and therapeutic classification in the State Registry of Medicinal Products.

Standardized terms for ADR and indications were manually matched with low-level terms (LLT) or preferred terms (PT) from MedDRA. In Table 7, we show the numbers of unique PT terms that match our mentions.

Table 7. Size characteristics of the collected corpus.

Entity Type	Mentions			
	Annotated	**Classification and Normalization**	**Num. Words in the Mentions**	**Reviews Coverage**
ADR	1784	316 (MedDRA)	4211	628
Medication	32,994		47,306	2799
Drugname	8236	550 (SRD), 226 (ATC)	9914	2793
DrugBrand	4653		5296	1804
Drugform	5994		6131	2193
Drugclass	3120	70 (ATC)	3277	1687
MedMaker	1715		2423	1448
Frequency	614		2478	516
Dosage	965		2389	708
Duration	1514		3137	1194
Route	3617		7869	1737
SourceInfodrug	2566		4392	1579
Disease	17,332		37,863	2712
Diseasename	4006	247 (ICD-10)	4713	1621
Indication	4606	343 (MedDRA)	7858	1784
BNE-Pos	5613		14,883	1764
ADE-Neg	85		347	54
NegatedADE	2798		9028	1104
Worse	224		1034	134
Note	4517		21,200	1876

3.4. Statistics of the Collected Corpus

We used the UDPipe [44] package to parse the reviews, in order to get sentence segmentation, tokenization and lemmatization. Given this, we calculated that the average number of sentences for the reviews is 10, the average number of tokens is 152 (with a standard deviation of 44) and the average number of lemmas is 95 (standard deviation equals to 23). The type/token ratio (TTR), calculated as the ratio of the unique lemmas in a review to the amount of tokens in it, is 0.64 for all reviews on average.

Detailed information about the annotated corpus is presented in Table 7, including:

1. The number of mentions for every attribute ("Mentions—Annotated" column in the table).
2. The number of unique normalized terms that match the mentions, and the number of unique classes from classifiers as described in Section 3.3 that the mentions belong to ("Mentions—Classification and Normalization").
3. The number of words belonging to mentions of the attribute ("Mentions—Number of words in the mentions").
4. The number of reviews containing any mentions of the corresponding attribute ("Mentions—Reviews Coverage").

The corpus contains 8236 mentions of drugs corresponding to 226 ATC codes. The 20% most popular ATC codes (by the number of reviews with the corresponding DrugName mentions) include 45 different codes which appear in 2614 reviews (93% of all reviews).

Of them, the 20 most popular ATC codes, which were reviewed in more than 50 posts (2511 posts in total), are listed in Figure 2.

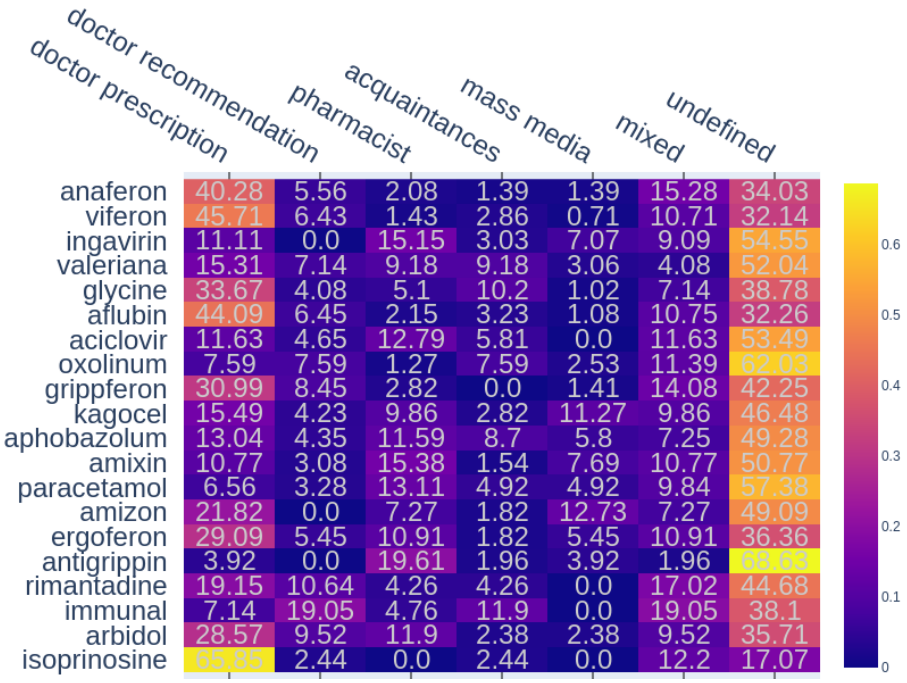

Figure 2. The percentages of different sources of information for the 20 most popular (in the collected corpus) drugs. The number in a cell means the ratio of reviews with co-occurring mentions of a drug and a particular source to the total number of reviews with this drug. If several different sources are mentioned in a review, it is counted as the "mixed" source.

The most popular second-level ATC codes are: L03 "Immunostimulants"—662 reviews (which is 23.6% of corpus); J05 "Antivirals for systemic use"—508 (18.5%) reviews; N05 "Psycholeptics"—449 (16.0%); N02 "Analgesics"—310 (11.1%); N06 "Psychoanaleptics"—294 (10.5%). The most popular drugs among immunostimulants by the reviews count are: Anaferon (144 reviews), Viferon (140) and Grippferon (71). The most popular antivirals for systemic use are the following: Ingavirin (99), Kagocel (71) and Amixin (58).

The proportions of reviews about domestic and foreign drugs to the total number of reviews are 44.9% and 39.7%, respectively. The remaining documents (15.4%) contain mentions of multiple drugs, both domestic and foreign, or mentions of drugs for which the annotators were unable to determine the origin. Among the domestic drugs are the following: Anaferon (144 reviews), Viferon (140), Ingavirin (99) and Glycine (98). Examples of mentioned foreign drugs include: Aflubin (93), Amison (55), Antigrippin (51) and Immunal (42).

Regarding diseases, the most frequent ICD-10 top level categories are "X—Diseases of the respiratory system" (1122 reviews); "I—Certain infectious and parasitic diseases" (300 reviews); "V—Mental and behavioural disorders" (170 reviews); and "XIX—Injury, poisoning and certain other consequences of external causes" (82 reviews). The top five low-level codes from ICD-10 by the number of reviews are presented in Figure 3.

Analyzing the consumers' motivation to acquire and use drugs ("sourceInfoDrug" attribute) showed that review authors mainly mention using drugs based on professional recommendations. Of the reviews, 989 mention doctors' prescriptions, 262 refer to pharmaceutical specialists' recommendations and 252 refer to doctors' recommendations. Some reviews report using drugs recommended by relatives (207 reviews), or chosen on the basis of advertisement (97) or the Internet (15).

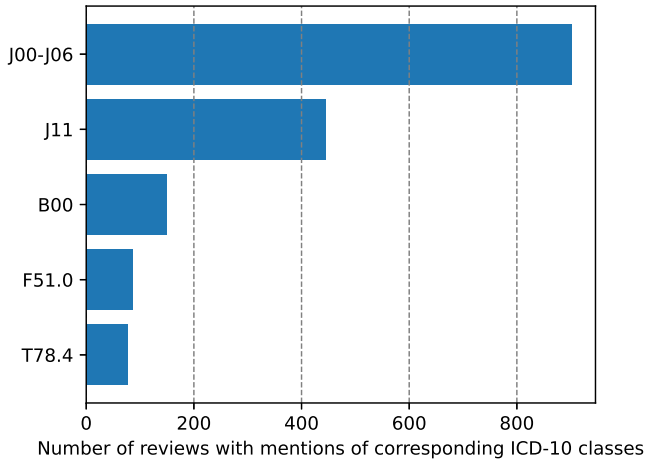

Figure 3. Top five low-level disease categories from ICD-10 by the number of reviews in our corpus. J00-J06—Acute upper respiratory infections, J11—Influenza with other respiratory manifestations, virus not identified, B00—Herpesviral [herpes simplex] infections, F51.0— Nonorganic insomnia and T78.4—Allergy, unspecified.

The heatmap presented on Figure 2 shows the percentages of different sources of recommendation for a few popular drugs. The sources were manually merged into five groups by the annotators.

It could be seen that most recommendations are coming from professionals. For example, Isoprinosine (used in 65.85% cases by medical prescription), Aflubin (44.09%), Anaferon (47.30%) and others. However, for such drugs as Immunal (11.9%) or Valeriana (9.18%), the rate of usage on the advice of patients' acquaintances is close to doctors' recommendations or higher. Of all the drugs, Amizon and Kagocel are most frequently (12.73% and 11.27%, respectively) mentioned by the users as chosen on the basis of information from mass media (advertisement, internet and others).

The distribution of the tonality (positive or negative) for the sources of information is presented in Figure 4. A source is marked as "positive" in a review if a positive dynamic is reported after the use of the drug (i.e., the review includes a BNE-pos attribute). "Negative" tonality is marked if a negative dynamic or deterioration in health has taken place or the drug has had no effect (i.e., Worse, ADE-Neg or NegatedADE mentions appear). Reviews that report both positive and negative dynamics are considered neutral and do not count towards the distribution.

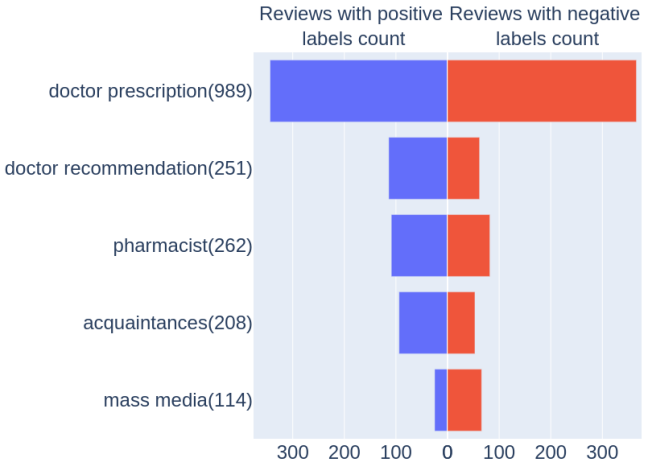

Figure 4. Distribution of tonality for the different sources. A number in brackets shows the number of reviews that mention a certain source of information, including reviews without reported effects or neutral reviews (with both good and bad effects).

The diagram in Figure 4 shows that drugs recommended by doctors or pharmacists are mentioned more often as having a positive effect, while using drugs based on an advertisement often leads to deterioration in health.

Diagrams in Figure 5 show the percentages of reviews where popular drugs were mentioned along with labeled effects.

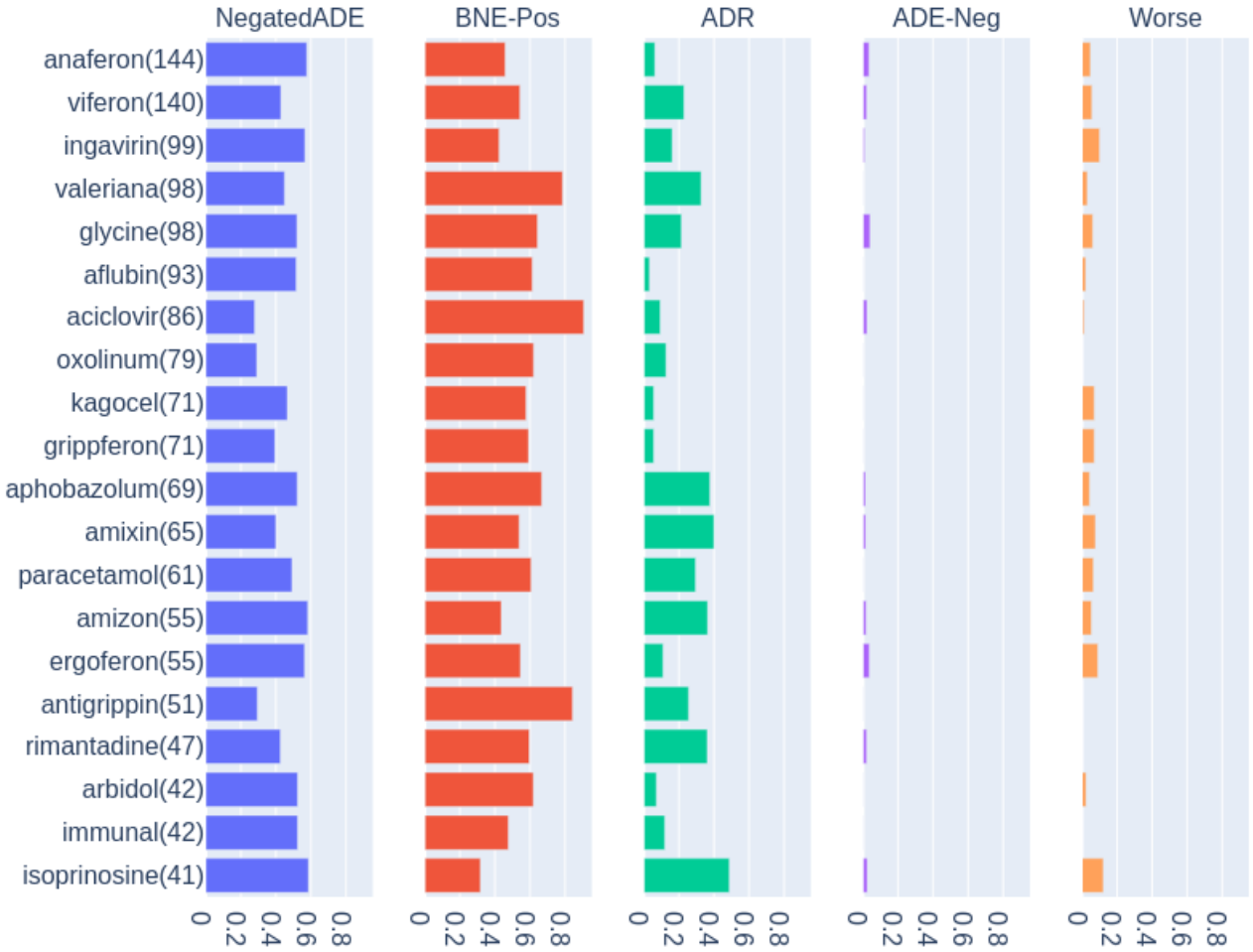

Figure 5. Distributions of labels of effects reported by reviewers after using drugs. The top 20 drugs by the reviews count are presented. The number in brackets is the number of reviews with mentions of a drug. The diagrams show the proportion of reviews mentioning a specific type of effect to the total amount of reviews on the drug.

The following drugs have the highest occurrences of ADR in reviews: immunomodulator "Isoprinosine" (48.8% of reviews with this drug contain ADR mentions), antiviral "Amixin" (40.0%), tranquilizer "Aphobazolum" (37.7%), antiviral "Amizon" (36.4%) and antiviral "Rimantadine" (36.3%). For some drugs, users mention negative dynamics of the disease after the start or some period of their usage (ADE-Neg). Examples of such drugs are "Anaferon" (3.5% of reviews with this drug mention ADE-Neg effects), "Viferon" (2.1%), "Glycine" (4.1%) and "Ergoferon" (3.6%). Some of the drugs cause a deterioration in health after taking the course (Worse label): immunomodulator "Isoprinosine" (12.2%), antiviral "Ingavirin" (10.1%), "Ergoferon" (9.1%) and others.

3.5. Coreference Annotation

Coreference annotation has been performed in two steps. Firstly, we used a state-of-the-art neural network model for coreference resolution [36], and adapted it to the Russian language by training on the corpus AnCor-2019. Using this model, we predicted coreference

for reviews in our corpus. We chose 91 reviews which had more than 2 different drug names and disease names (after the manual grouping described in Section 3.3) and more than 4 coreference clusters, and 209 reviews which had more than 2 different drug names and more than 2 coreference clusters. These 300 reviews were given to our annotators for manual checking of the coreference clusters predicted by the model.

The annotators had guidelines for coreference and a set of examples. According to the guidelines, they were supposed to pay attention to mentions of pharmacological types, pronouns and words typical for references (e.g., "such", "former" and "latter"). They did not annotate as coreference the following cases:

- Mentions of the reader ("you" in "I wouldn't recommend you to buy it if you don't want to waste money");
- Split antecedents, where two or more mentioned entities are then referred to by a common phrase ("I tried Coldrex, and after a while I decided to buy Antigrippin. Both drugs usually help me.");
- Generic mentions: phrases that describe some objects or events but not particular entities (e.g., "doctors" in "Many doctors recommend this medication. Since I respect the doctors' opinion, I decided to buy it.");
- Phrases that establish a relationship between different entities; for example, when one is a more general notion to which the other belongs ("Valeriana" and "sedative drug" in "Valeriana is a good sedative drug that usually helps me").

Table 8 shows the number of coreference clusters and coreferent mentions in 300 drug reviews from our corpus compared to the corpus AnCor-2019.

Table 8. Number of coreference chains and mentions compared to the other Russian coreference corpus.

Corpus	Texts Count	Mentions Count	Chains Count
AnCor-2019	522	25,159	5678
Our corpus	300	6276	1560

It should be noted that not all coreferent mentions correspond to mentions of our main entity annotation: sometimes a single coreferent mention can unite multiple medical mentions or connect pronouns that are not involved in the medical annotation. Table 9 represents the number of medical mentions of various types that intersect with coreferent mentions.

Table 9. Mention types involved in coreference chains.

Entity Type	Attribute Type	Number of Mentions Involved in Coreference Chains
Medication	Drugname	529
	Drugform	286
	Drugclass	204
	MedMaker	170
	Route	98
	SourceInfodrug	75
	Dosage	50
	Frequency	1
Disease	Diseasename	163
	Indication	125
	BNE-Pos	107
	NegatedADE	36
	Worse	5
	ADE-Neg	2
ADR		34

4. Machine Learning Methods

4.1. NER Task Formulation

We consider the NER problem of detecting pharmaceutically relevant entities as a multi-label classification of tokens—words and punctuation marks—in sentences. For each token, the output is a set of tags that comprises a tag in the BIO format for each attribute of each entity type (DrugName, DrugBrand and so on): the "B" tag indicates the first word of the mention of the particular attribute, the "I" tag is used for subsequent words within the mention and the "O" tag means that the word is outside of an entity mention. A token can be inside multiple mentions, allowing for intersecting mentions of different attributes.

We evaluate two methods for entity recognition on our corpus. The first (Model A) is based on a bidirectional long short term memory (BiLSTM) neural network topology with different features representing an input text: dictionaries, part of speech tags and several methods of word-level representations, including FastText [45], ELMo [26], BERT, words character long short term memory (LSTM) coding, etc. At its output, Model A produces one of the three tags—B, I or O—indicating the input token's belonging to a particular entity attribute. For each attribute of each entity type, an independent instance of the model is trained. The second (Model B) is a multi-label model which predicts all tags of a token using a single instance of the neural network. It combines the pre-trained multilingual language model XLM-RoBERTa [46] and the LSTM neural network with several of the most efficient features. Details of the implementation of both methods and the features they use for input encoding are presented below.

4.2. Features Used for Text Representation

4.2.1. Tokenization and Part-of-Speech Tagging

To pre-process the text, we use the UDPipe [44] tool. After parsing, each word is assigned 1 out of 17 parts of speech. They are represented as a one-hot vector, and then processed with an embedding layer, the output of which is then used within the input for the neural networks Model A and Model B. For Model B, the text is split into phrases using UDPipe version 2.5. Long phrases are split up into 45-word chunks.

Such vector representation of a part of speech, later referred to as PoS, also contains a binary vector of answers to the following questions (1 if yes, 0 otherwise):

- Are all letters capital?
- Are all letters in lowercase?
- Is the first letter capital?
- Are there any numbers in the word?
- Do more than a half of the word consist of numbers?
- Does the entire word consist of numbers?
- Are all letters Latin?

4.2.2. Emotion Markers

Adding the frequencies of emotional words as extra features is motivated by the positive influence of these features on determining the author's gender [47]. Emotional words are taken from the dictionary (Information Retrieval System "Emotions and feelings in lexicographical parameters: Dictionary of the emotive vocabulary of the Russian language"—http://lexrus.ru/default.aspx?p=2876, accessed on 12 December 2021) which contains 37 emotion categories, such as anxiety, inspiration, faith, attraction, etc. On the basis of the n emotion categories available in the dictionary, an n-dimensional binary vector is formed for each word, where each vector component reflects the presence of the word in a certain emotion category.

In addition, this word feature vector is concatenated with emotional features of the whole text. These features are English Linguistic Inquiry and Word Count (LIWC) and psycholinguistic markers.

The former is a set of specialized English Linguistic Inquiry and Word Count (LIWC) dictionaries [48], adapted for the Russian language by linguists [49]. The LIWC values are

calculated for each document based on the occurrence of its words in the corresponding psychosocial dictionaries.

Psycholinguistic text markers [50] reflect the emotional intensity of the text. They are calculated as ratios of certain frequencies of parts of speech in the text. We use the following markers: the ratio of the number of verbs to the number of adjectives per unit of text; the ratio of the number of verbs to the number of nouns per unit of text; the ratio of the number of verbs and verb forms (participles and adverbs) to the total number of all words; and the number of question marks, exclamation points and average sentence length.

The combination of these features are further referred to as "ton".

4.2.3. Dictionaries

The following dictionaries from open databases and registers are used as additional features for the neural network model.

1. CUI codes obtained from the MESHRUS thesaurus as described in Appendix B. The two approaches described there are referred to as MESHRUS and MESHRUS-2.
2. Categories from the Vidal medication handbook [51]: adverse effects, drug names in English and Russian, diseases. The dataset words are mapped to the words or phrases from the Vidal handbook. To establish the categories, the same approach as for MESHRUS is used. The difference is that, instead of setting indices for every word (as CUI in the UMLS), we assign a single index to all words of the same category. That way, words from the dataset are not mapped to special terms, but checked for category relations.
3. Categories from MedDRA are obtained as described in Section 3.3.

The resulting binary vector (one-hot representation in the case of CUI codes and vectors reflecting belonging to categories in the case of Vidal and MedDRA) is then processed with an embedding layer.

4.2.4. Language Models

Language models, pre-trained on large bodies of unlabeled texts, represent words by vectors in a space where words with similar meanings are close to each other. We use the following models: FastText [45], Embeddings from Language Model (ELMo) [26], Bidirectional Encoder Representations from Transformer (BERT) [27] and XLM-RoBERTa [46].

The approach of FastText is based on the Word2Vec model principles, where word distributions are predicted by their context, but FastText uses character trigrams as its basis vector representation. Each word is represented as a sum of its trigram vectors, which are then used as the base for continuous bag of words or skip-grams algorithms [52]. Such a model is simpler to train due to decreased dictionary size: the number of character n-grams is less than the number of unique words. Another advantage of this approach is that morphology is accounted automatically, which is important for the Russian language.

Instead of using fixed vectors for every word similar to how FastText does, ELMo word vectors are sentence-dependent. ELMo is based on the bidirectional language model (BiLM), which learns to predict the next word in a word sequence. Vectors obtained with ELMo are contextualized by means of grouping the hidden states (and initial embedding) in a certain way (concatenation followed by weighed summation). However, predicting the next word in a sequence is a directional approach and therefore is limited in taking the context into account. This is a common problem in training NLP models, and is addressed in BERT.

BERT is based on the transformer mechanism, which analyzes contextual relations between words in a text. The BERT model consists of an encoder extracting information from a text and a decoder which gives output predictions. In order to address the context accounting problem, BERT uses two learning strategies: word masking and logic check of the next sentence. In the first strategy, 15% of the words are replaced with a token "MASK", the original words later being the target for the neural network to predict. In the second learning strategy, the neural network is used to determine whether two input sentences

are a logical sequence or just a random set of unrelated phrases. In BERT training, both strategies are used simultaneously by minimizing their combined loss function.

XLM-RoBERTa is a model similar to a masked BERT language model based on Transformers [53]. The main differences between XLM-RoBERTa and BERT are the following. Firstly, XLM-RoBERTa was trained on a larger multilingual corpus from the CommonCrawl project which contains 2.5TB of texts. Russian is the second language by texts count in this corpus after English. Minibatches during model training included texts in different languages. Secondly, XLM-RoBERTa was trained only for the masked token prediction task; its loss function did not involve the next sentence prediction learning strategy. Thirdly, it used a different tokenization algorithm: while BERT used WordPiece [54], XLM-RoBERTa used SentencePiece [55]. The vocabulary size in XLM-RoBERTa is 250,000 unique tokens for all languages.

There are two versions of the model: XLM-RoBERTa-base (with 270M parameters) and XLM-RoBERTa-large (with 550M), of which we use the latter.

4.3. Neural Network Architecture

4.3.1. Model A: BiLSTM Neural Network

The topology of Model A is depicted in Figure 6.

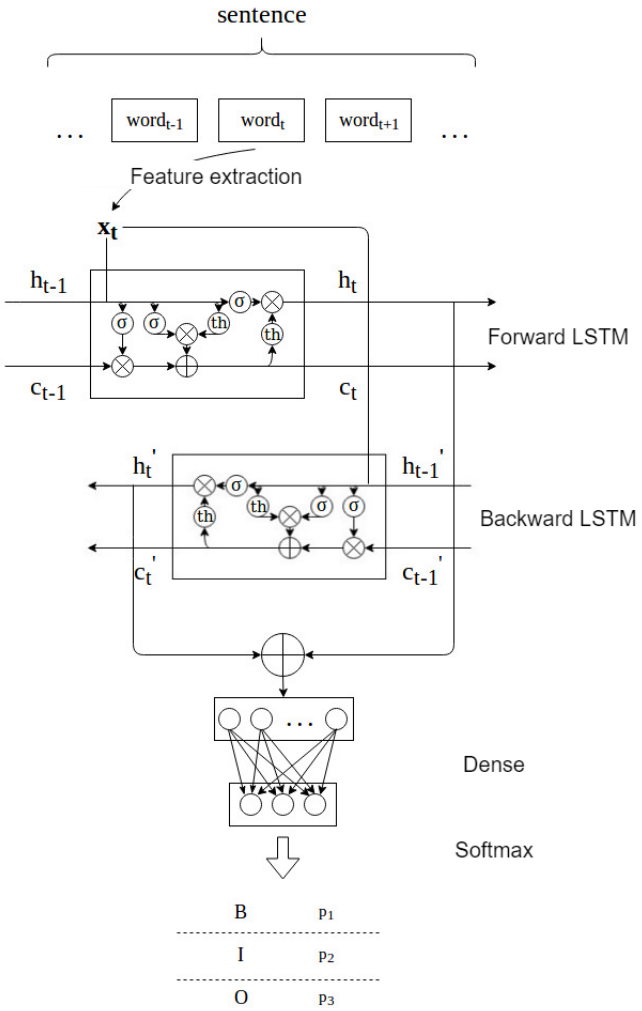

Figure 6. The network architecture of Model A. Input data goes to a bidirectional LSTM, where the hidden states of forward LSTM and backward LSTM get concatenated, and the resulting vector goes to a fully connected ("dense") layer with size 3 and SoftMax activation function. The output p_1, p_2 and p_3 are the probabilities for the word to belong to the classes B, I and O, i.e., to have a B, I or O tag.

Its inputs are various combinations of features described in Section 4.2, and additionally, word encoding obtained with a characters-convolution-based neural network CharCNN [56] (see Figure 7).

At the input of CharCNN, each word is represented as a fixed-length character sequence. The number of characters is a hyperparameter, which in this study has been chosen empirically with the value of 52. If the word has fewer characters than this number, the remaining characters are filled with the ≪PADDING≫ symbol. Character vocabulary is formed from the training dataset, and also includes special characters ≪PADDING≫ and ≪UNKNOWN≫, the latter allowing for a possible future occurrence of characters not present in the training set. For coding each character of the word, an embedding layer [57] is used, which replaces every character from the vocabulary with a real vector of size 30. The values of these vectors are initialized randomly from the uniform distribution in the range of $[-0.5; 0.5]$, and then trained. After encoding by the embedding layer, the matrix of encoded characters representing a word is processed by a convolution layer [58] (with 30 filters and a kernel size of 3) and global maxpooling function that provides a maximization function of all the values for each filter [59].

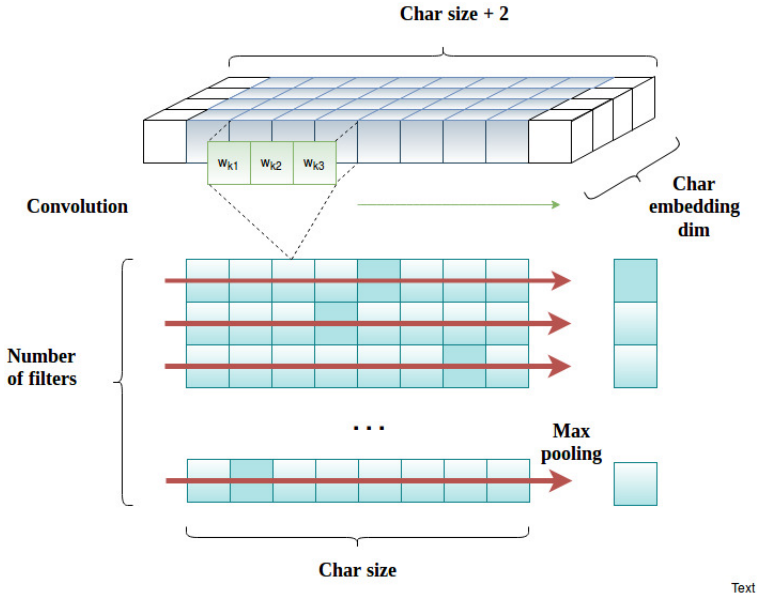

Figure 7. The scheme of character feature extraction on the basis of a char convolution neural network. Each input vector, after being processed by the embedding layer, is expanded with two extra padding objects (white boxes). $w_{(k1)}, w_{(k2)}, w_{(k3)}$ are the weights of the convolution filter k.

At the output of the model, we put either a fully connected layer [19] or conditional random fields (CRF) [60], which output the probabilities for a token to have a B, I or O tag for the corresponding entity (for instance, B-ADR, I-ADR or O-ADR).

4.3.2. XLM-RoBERTa

To tune the language model to texts of a medical nature, we performed an additional training of XLM-RoBERTa-large on a dataset (https://huggingface.co/sagteam/xlm-roberta-large-sag, accessed on 12 December 2021), containing two sets of texts: the first one, consisting of 250,000 reviews on medicines (an average with 1000-token-long), has been collected from the website irecommend.ru (accessed on 12 December 2021), and the second one has been borrowed from the unannotated part of RuDReC [10]. The calculations of XLM-RoBERTa-large for one epoch were performed using a computer with one Nvidia Tesla v100 and the Huggingface Transformers library, and took five days.

Then, we fine-tuned the language model for solving the NER task as depicted in Figure 8.

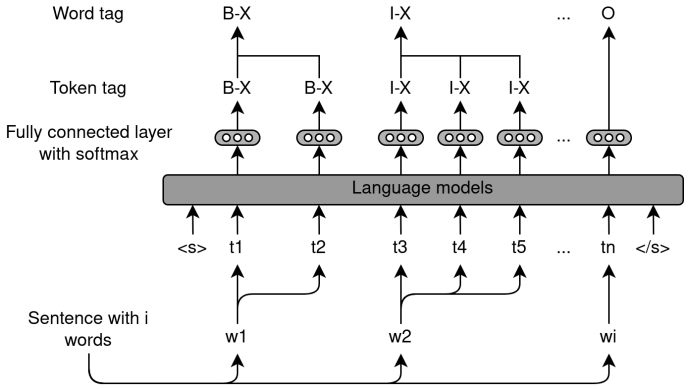

Figure 8. Fine-tuning of a language model for the word classification task. X stands for the attribute name.

It is the commonly used fine-tuning algorithm of the Simple Transformers project [61]. As the output layer for classifying words, a fully connected layer with the softmax activation function is added to the model. The output classes are "B-DrugBrand", "I-DrugBrand", "B-DrugClass", "I-DrugClass" and so on for all the attributes of all the entity types, and finally, "O" ("outside of any mention").

4.3.3. Model B: XLM-RoBERTa-Based Multi-Model

Model B is a multi-tag model that combines the fine-tuned XLM-RoBERTa language model described in Section 4.3.2 with a simplified variant of Model A, with CRF excluded and ELMo word representation substituted by the output of the fine-tuned language model.

The output vector of class activations from the fine-tuned language model is concatenated (see Figure 9) with a vector of features out of those described in Section 4.2 (MESHRUS, MESHRUs-2, PoS and ton), and also concatenated with the output of Char-CNN described in Section 4.3.1. The resulting vector is then processed by the LSTM neural network model depicted in Figure 10 so as to obtain multi-tagged labeling.

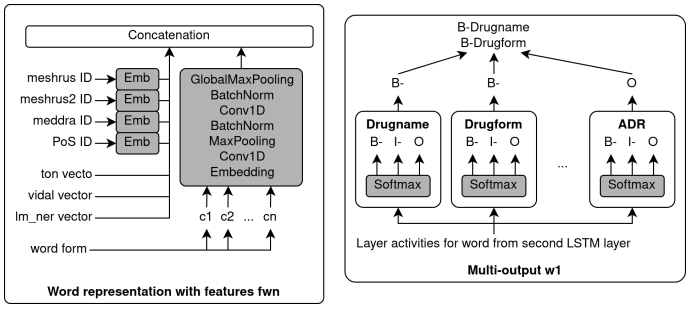

Figure 9. On the left: word vector representation within Model B. **On the right**: the multi-output scheme for word classification within Model B.

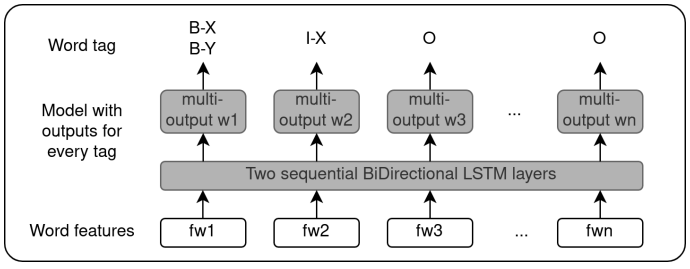

Figure 10. The architecture of Model B. Vector representations fw$n$ of each word n are obtained as depicted in Figure 9 **on the left**. Elements of the output layer denoted as "multi-output" are explained in Figure 9 **on the right**. X and Y stand for the attribute names.

Here, the output classes are "B-DrugBrand", "I-DrugBrand", "O-DrugBrand" and so on (where MedMaker/Domestic, MedMaker/Foreign, DrugName/MedFromDomestic and DrugName/MedFromForeign are considered separate attributes).

The hyperparameters of the multi-tag model have been adjusted automatically with the help of Weights&Biases Sweeps [62]. With six parallel processing agents, it took about 24 h on a computer with three Tesla K80.

4.3.4. Coreference Model

For coreference resolution, we chose a state-of-the-art neural network architecture from [36]. The key feature of this model is end-to-end learning: the task of mentions detection and the task of mentions linking and forming coreference clusters are learned at the same time rather than one after another. The model uses the BERT language model to retrieve vector representations for words of an input text.

In order to adapt the network architecture to the Russian language, we used RuBERT, the BERT language model trained on the Russian part of Wikipedia and news data. After tuning the neural network hyperparameters and training options, the optimal hyperparameters were chosen as follows: maximum span width = 30, maximum antecedents for every mention: 50, hidden fully connected layers size = 150, numbers of sequential hidden layers = 2, maximum epoch training: 200, language model learning rate = 10^{-5}, task model learning rate = 0.001 and embedding sizes = 20.

5. Experiments and Results

5.1. Methodology

In the experiments, we pursued the following objectives:

- To find the optimal model for mention detection (in Section 5.2). In Sections 5.2.1 and 5.2.2, respectively, we choose the optimal language model and combination of input features for Model A. In Section 5.2.3, we compare several variants of neural network topology for Model A. Then, we evaluate the XLM-RoBERTa model separately, and combine it with the optimal features found for Model A, resulting in the creation of Model B;
- To compare the ADR mention extraction accuracy on our corpus against the available data of a similar type for the Russian language (see Section 6);
- To show how the following characteristics of the corpus affect the ADR extraction accuracy: the proportion of phrases containing ADR, the proportion of ADR and Indication mentions, the corpus size, etc. (in Section 5.3);
- To evaluate the influence of the strictness of ADR labeling on the ADR identification precision (in Section 5.4).

The reason for the focus on ADR when calibrating models is that this entity type is practically important while at the same time the most difficult for automated identification because it is strongly dependent on its context.

The performance of the entity detection models is estimated with the help of the chunking metric that was introduced at the Conll-2000 shared task and has been used to compare named entity recognition systems since then. The script (https://www.clips.uantwerpen.be/conll2000/chunking/, accessed on 12 December 2021) receives a file as its input, where each line contains a token, its true tag and its predicted tag. Tags could be "O" if the token does not belong to any mentions, "B-X" if the token starts a mention of some type X or "I-X" if it continues a mention of type X. If a tag "I-X" appears after "O" or "I-Y" (mention of some other type), it is treated as "B-X" and starts a new mention. We use the F1-exact score that estimates the accuracy of full entity matching. The script calculates F1-exact as the F_1 score based on the percentage of detected mentions that are correct (precision) and the percentage of correct mentions that were detected (recall):

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

The coreference resolution performance is estimated with three commonly used metrics [63]: MUC, which is based on counting coreference relations added or missing in the generated markup compared to the ground truth; B^3, where recall and precision are calculated for every mention as the fractions of correct mentions in the coreference chain to which this mention belongs in the generated markup; CEAFe, which is calculated by finding an optimal mapping between coreference chains from the ground truth markup to coreference chains in the generated markup, and then using another similarity metric to compare mentions in the obtained pairs of chains.

5.2. Estimation of the Influence of Language Models, Input Features and Network Topology on the ADR Detection Accuracy

5.2.1. Choosing the Best Embedding for Model A

We consider the following pre-trained language models for word embedding: FastText, ELMo and BERT (see Section 4.2.4). FastText [64] has been taken from the open repository (https://fasttext.cc/docs/en/crawl-vectors.html, accessed on 12 December 2021), where it is available pre-trained on a large body of Russian texts from the CommonCrawl project (http://commoncrawl.org/, accessed on 12 December 2021), and then has been pre-trained on reviews from Otzovik.com (https://otzovik.com/health/, accessed on 12 December 2021) from the categories "medicines" (2,555,833 texts, 15.36 words in a text on average, 39,256,947 words total) and "hospitals" (3,290,912 texts, 15.04 words in a text on average, 49,500,274 words total).

The ELMo model was taken from the DeepPavlov [65] open-source library (https://deeppavlov.readthedocs.io/en/master/intro/pretrained_vectors.html, accessed on 12 December 2021), where it is available pre-liminarily trained on the Russian WMT News [66]. The multilingual BERT model, pre-trained on Wikipedia texts, was taken from the Google repository (https://github.com/google-research/bert/, accessed on 12 December 2021), and subsequently trained on the above-mentioned drug and hospital reviews.

Each of these pre-trained models is used as the input to our neural network Model A described in Section 4.3.1. The dataset RDRS-1660 (the first version of our corpus which contains 1660 reviews) is split into 5 folds for cross-validation. On each fold, the training set is split into training and validation sets in the ratio 9:1. Training is performed for a maximum of 70 epochs, with early stopping by the validation loss. Cross entropy is used as the loss function, with nAdam as the optimizer and cyclical learning rate mechanism [67].

The results of embedding comparison experiments are given in Table 10 and demonstrate the superiority of the ELMo model. BERT leads to lower F1 values with larger deviation ranges, and with the FastText model, the F1 score is the lowest. Combining ELMo with BERT by concatenating their output vectors worsens the accuracy. As a result, we use ELMo in the next section when comparing different input feature combinations.

Table 10. Accuracy (%) of recognizing ADR, medication and disease entities in the first version of our corpus (1660 reviews) by Model A with different language models.

Word Vector Representation	Vector Dimension	ADR	Medication	Disease
FastText	300	22.4 ± 1.6	70.4 ± 1.1	44.1 ± 1.7
ELMo	1024	24.3 ± 1.7	73.4 ± 1.5	46.4 ± 0.6
BERT	768	22.1 ± 2.4	71.4 ± 3.3	45.5 ± 3.2
ELMO||BERT	1024||768	18.7 ± 9.8	74.1 ± 1.1	47.9 ± 1.6

5.2.2. Influence of Different Input Features

In order to evaluate the contribution of any particular feature out of those described in Section 4.2, we evaluate Model A with ELMo in combination with emotion markers, PoS and MESHRUS, MESHRUS-2 and Vidal dictionaries. In these experiments, texts are passed to the language model split into independent sentences.

The results presented in Table 11 (compare to the results of ELMo in Table 10) show that adding these features improves the accuracy for the least-represented class ADR.

Table 11. The accuracy (by the F1-exact metric) of recognizing entities of different types in the first version of our corpus (RDRS-1600) using models with different features and topology.

Topology and Features	ADR	Medication	Disease
Model A—Influence of features			
ELMo + PoS	26.2 ± 3.0	72.9 ± 0.6	46.6 ± 0.9
ELMo + ton	26.6 ± 3.9	73.5 ± 0.5	47.3 ± 1.0
ELMo + Vidal	26.8 ± 1.0	73.2 ± 1.1	45.8 ± 1.2
ELMo + MESHRUS	27.4 ± 2.2	73.3 ± 1.5	46.5 ± 1.2
ELMo + MESHRUS-2	27.4 ± 0.9	73.1 ± 0.4	46.7 ± 1.4
Model A—Topology modifications			
ELMo with 3-layer LSTM	28.2 ± 5.1	74.7 ± 0.7	51.5 ± 1.8
ELMo with CRF	28.8 ± 2.7	73.2 ± 1.1	46.9 ± 0.4
Model A—Best combination			
ELMo with 3-layer LSTM and CRF + ton, PoS, MESHRUS, MESHRUS-2, Vidal	32.4 ± 4.7	74.6 ± 1.1	52.3 ± 1.4
XLM-RoBERTa			
XLM-RoBERTa-large	40.1 ± 2.9	79.6 ± 1.3	56.9 ± 0.8

Addition of any of the individual features separately leads to an increase in ADR recognition accuracy by 2% to 3%. In particular, part of speech and tonality tags give a 2% increase. These features are of a generic nature, which is the reason why these features give less increase in the accuracy compared to the features based on the MESHRUS vocabulary. The latter contains a lot of medical terminology, so words marked with features of MESHRUS are more important for the NER model. This is why MESHRUS and MESHRUS-2 give a 3% accuracy increase. Increasing the depth of the network with additional LSTM layers helps the model to extract more high-level features and gives a 4% increase compared to base Model A with ELMo, but it makes the process of convergence of the neural network harder. The CRF layer helps to predict more probable sequences of tags. It gives us 4% more accuracy without other additions. Combining all the features gives a significant increase in accuracy for ADR mentions (+8%).

5.2.3. Finding the Best Model Topology

We compare several variations of the topology of Model A: replacing the last fully connected layer with a CRF layer, or changing the number of biLSTM layers (see the part "Topology modifications" in Table 11). Eventually, a combination of dictionary features, emotion markers, 3-layer LSTM and CRF achieves the highest accuracy for ADR and disease entities. For medication, the combination of ELMo and 3-layer LSTM shows slightly better results. This is therefore the accuracy level of Model A (see "Model A—Best combination" in Table 11).

Then, in order to evaluate the effectiveness of XLM-RoBERTa-large, we run it without additional input features, as described in Section 4.3.2 (see the last row in Table 11). Overall, XLM-RoBERTa-large outperforms all the experiments with Model A, and so we use it as the basis for Model B (described in Section 4.3.3).

5.3. The Influence of Corpus Characteristics on the ADR Detection Accuracy

First of all, we conducted experiments on the latest version (RDRS-2800) of our corpus that contains 2800 texts, obtained by extension of the first version RDRS-1660 (containing 1660 texts) so as to assess the dependence of ADR detection accuracy on the number of

ADR mentions. Such direct expansion of the corpus (see RDRS-1600 and RDRS-2800 in Table 12) results in an increase in the ADR identification precision by 13% for ADR, 6% for disease, and 4% for medication.

Table 12. Subsets of the RDRS corpus with respect to the number and proportion of ADR mentions, and their ADR detection accuracy.

Corpus	RDRS-2800	RDRS-1600	RDRS-1250	RDRS-610	RDRS-1136	RDRS-500
Number of reviews	2800	1659	1250	610	1136	500
Number of reviews containing ADR	625	339	610	610	610	177
Percentage of reviews containing ADR	0.22	0.2	0.49	1	0.54	0.35
Number of ADR entities	1778	843	1752	1750	1750	709
Average number of ADR per review	0.64	0.51	1.4	2.87	1.54	1.42
Number of reviews containing Indication	1783	955	670	59	154	297
Total number of entities	52,186	27,987	21,807	3782	6126	9495
Number of Indication entities	4627	2310	1518	90	237	720
Ratio of ADR to Indication entities	0.38	0.36	1.15	19.44	7.38	0.98
F1-exact of ADR detection	52.8 ± 3.8	40.1 ± 2.9	61.1 ± 1.5	71.3 ± 3.4	68.6 ± 3.3	61.6 ± 2.9
Saturation $(\cdot 10^3)$	4.25	3.41	9.77	72.57	42.99	9.08

Figure 11 presents the results of training Model B on different fractions of the training set of RDRS-2800, and shows that the ADR detection accuracy stops growing when the training set reaches 80% of its size.

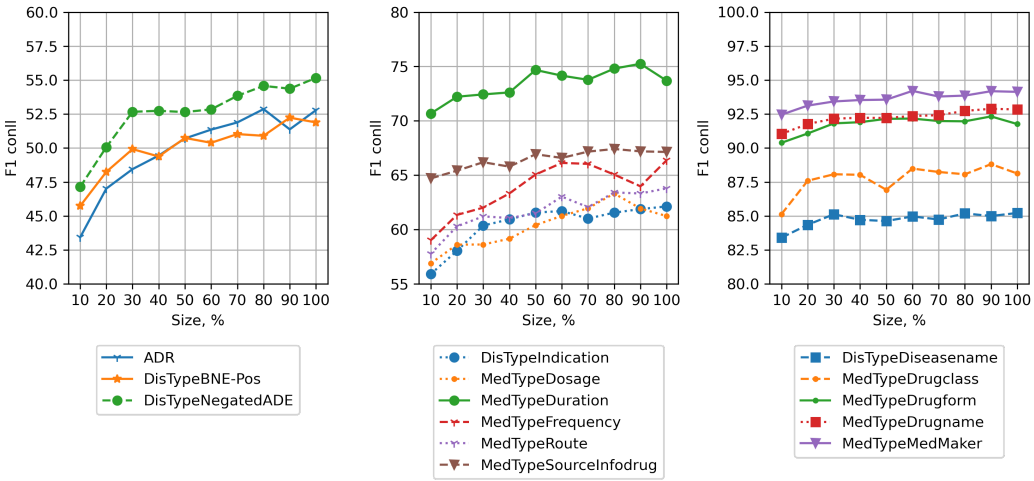

Figure 11. Dependence of the ADR recognition accuracy (by the F1-exact metric) on the size of the training set for different tags in RDRS-2800.

Similar behavior is observed for the accuracy of recognizing other entity types (see Table 13).

Table 13. F1-exact of detecting mentions with different tags for RDRS-1250 (the balanced version of our corpus) and RDRS-2800 (the full version). * Negative is the union of tags Worse, NegatedADE and ADE-Neg.

Corpus	RDRS-1250	RDRS 2800
BNE-Pos	51.2	50.3
Diseasename	87.6	88.3
Indication	58.8	62.2
Dosage	59.6	63.2
DrugBrand	81.5	83.8
Drugclass	89.7	90.4
Drugform	91.5	92.4
Drugname	94.2	95.0
DrugName/MedFromDomestic	61.7	76.2
DrugName/MedFromForeign	63.5	74.4
Duration	75.5	74.7
Frequency	63.4	65.0
MedMaker	92.5	93.8
MedMaker/Domestic	65.1	87.1
MedMaker/Foreign	74.4	85.0
Route	58.4	61.2
SourceInfodrug	66.0	67.3
Negative *	52.2	52.0

Note also that direct expansion from 1600 to 2800 mentions gives only a small increase in the average number of ADR mentions per review (0.22 versus 0.2). So, its saturation by ADR stays lower than in most of the existing corpora surveyed in Table 1.

In order to study the effect of increasing saturation of the corpus by ADR mentions, we experiment with subsets of RDRS that have various sizes and various ADR mention shares per review (see Table 12).

Increasing the proportion of ADR by balancing the corpus by the amount of documents with ADR (in the RDRS-1250, 50% of reviews have ADR in it) leads to a more significant increase in ADR precision of 21%. At the same time, it does not cause a significant change in the disease and medication detection accuracy (see Table 14).

Table 14. Accuracy of recognizing three entity types for three subsets of the corpus, different by size and balancing.

	Number of Entities			F1-Exact		
RDRS Subset	**ADR**	**Medication**	**Disease**	**ADR**	**Medication**	**Disease**
RDRS-2800	1778	33,008	17,408	52.8 ± 3.4	84.1 ± 0.8	63.5 ± 0.5
RDRS-1250	1752	13,750	6307	61.1 ± 1.5	84.2 ± 0.6	62.9 ± 1.5
RDRS-1600	843	17,931	9840	40.1 ± 2.7	79.6 ± 1.3	56.9 ± 0.9

This may be explained by the higher saturation of the corpus by these entity types, which stays practically unchanged after balancing the corpus. Corpus RDRS-610 includes only sentences with ADR, and corpus RDRS-1136 includes sentences 50% of which has ADR

in it and 50% does not. The experiments on these corpora, which has an ADR saturation closer to that of CADEC, show a further increase of ADR detection accuracy up to 71.3.

5.4. Influence of Annotation Strictness on ADR Detection Accuracy

Here, we conduct two sets of experiments: with and without including mentions that are labeled as both ADR and Note. The results (compare red and blue lines in Figure 12) show that restricting the dataset to only unambiguous ADR mentions leads to a 3% accuracy decrease.

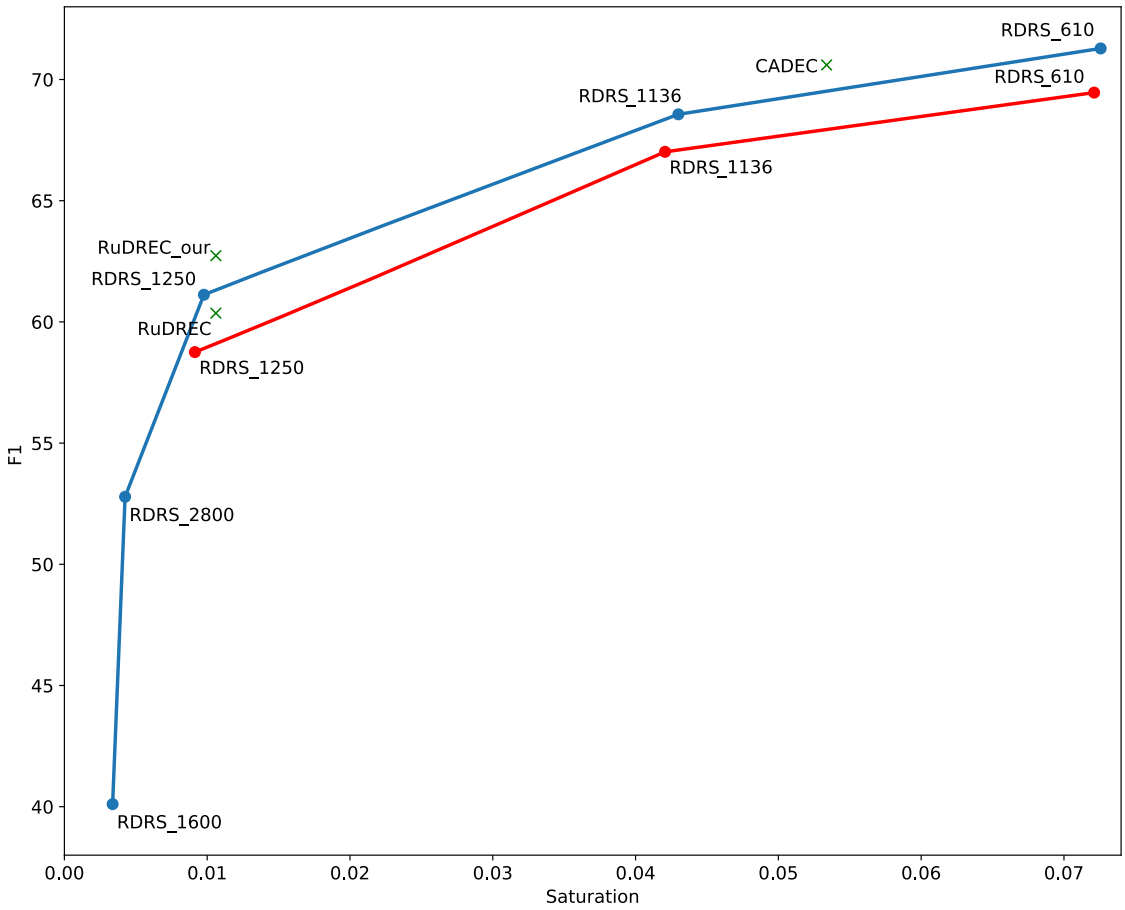

Figure 12. Dependency of ADR recognition precision on their saturation in the corpora. Red line—different subsets of our corpus (see Table 12) with ADR annotation (without Note tags). Blue line—different subsets of our corpus with overlapping of ADR and Note entities, RuDREC—published accuracy for RuDREC corpus [10], RuDREC_our—our accuracy for RuDREC corpus and CADEC—published accuracy for CADEC corpus [14].

5.5. Evaluation of the Accuracy of Coreference Relation Extraction on Our Corpus by Models Trained on Different Corpora

We evaluate the coreference resolution model described in Section 4.3.4 on our corpus with the coreference annotation described in Section 3.5. For this purpose, the corpus is split into train, validation and test subsets. Training the model is performed on the training subset of our corpus, or on the training subset of AnCor-2019, or on both.

Table 15 presents the coreference resolution accuracy dependent on what corpus the model is trained and tested on. The results show that the best accuracy on the testing subset of our corpus is achieved when training is performed on the training subset of our corpus, but not on AnCor-2019 nor on both.

Table 15. Results of the coreference resolution model trained and tested on different corpora.

Training Corpus	Testing Corpus	Avg F1	B^3 F1	MUC F1	CEAFe F1
AnCor-2019	Our corpus	58.7	56.4	61.3	58.3
AnCor-2019	AnCor-2019	58.9	55.6	65.1	55.9
Our corpus	Our corpus	71.0	69.6	74.2	69.3
Our corpus	AnCor-2019	28.7	26.5	33.3	26.4
AnCor-2019 + Our corpus	Our corpus	49.4	47.6	52.2	48.4
AnCor-2019 + Our corpus	AnCor-2019	31.8	31.4	40.7	23.3

6. Discussion

Currently, there is a significant diversity of full-sized labeled corpora in different languages for analyzing safety and effectiveness of drugs. We present the first full-size Russian corpus of Internet users' reviews with compound NER labeling and with the labeling of coreference relations in a part of the corpus.

Based on the results of the developed neural network models, we investigate the place of our corpus in this diversity depending on the corpora characteristics. Experiments performed on subsets with different saturation by a certain entity allow for giving a more realistic conclusion about the quality of the corpus concerning this entity.

The results of Model B developed on the base of XLM-RoBERTa-large outperform the existing results [10] by 2.3% for ADR detection accuracy on the corpus of a limited size. This justifies the quality of the developed Model B and the applicability of its results to establish state of the art for entity extraction precision on the created corpus.

In general, the results of experiments with sets of different sizes and different saturation show that the ADR identification accuracy strongly depends on the saturation of the corpus by these entities (see Figure 12). Therefore, a comparison of similar types of corpora, such as ours and CADEC, should be carried out on datasets that have similar values of ADR saturation.

In general, entities conform to three groups according to the ranges of their extraction accuracy: 42.5–55%, 55–75%, and 82.5–95% (see Figure 11). The first group, with the lower precision values, consists of entities that are more dependent on the informal language of writing a review context and are present only in a part of all reviews (e.g., ADR, BNE-Pos, etc.). The last group, with the largest precision values, consists of entities more dependent on domain-specific vocabulary, making extracting such entities easier.

The coreference relation extraction experiments show that the highest coreference resolution accuracy is achieved when the model is trained and tested on our corpus. All the other choices of the training set worsen the accuracy. This can be explained by the essential difference of the corpora from different domains.

7. Conclusions

The primary basic result of this work is the creation of the full-size Russian multi-tag NER-labeled corpus of Internet users' reviews on drugs, including the part of the corpus with annotated coreference relations. The corpus has a complex annotation scheme with 18 types of mentions, intersecting mentions, discontinuous mentions and coreference annotation. This allows us to build systems that can extract more detailed information demanded in the field of Russian pharmacovigilance. A multi-label neural network model for entity recognition, appropriated for labeling the presented corpus, is developed based on combining a language model XLM-RoBERTa with the selected set of input features. The model is capable of multi-tag labeling. It allows us to extract intersecting and discontinuous entities. The results obtained using this model show that the ADR detection accuracy on our corpus is comparable to that obtained on corpora of other languages with similar characteristics. Thus, this accuracy level may be considered the state of the art of this

task for Russian texts. The presence of a part with annotated coreference relations in our corpus allows us to evaluate the coreference resolution accuracy on texts of the profile under consideration.

Further work will be aimed at creating methods for recognizing entities with increased accuracy and solving the problem of normalization, i.e., establishing the correspondence of the selected entities with concepts from international dictionaries and thesauri (ICD, MedDRA, etc.).

Author Contributions: Conceptualization, A.S. (Alexander Sboev) and S.S.; methodology, I.M. and R.R.; software, I.M., A.G., A.N., A.S. (Anton Selivanov) and G.R.; validation, A.G., A.S. (Anton Selivanov) and I.M.; formal analysis, A.S. (Alexander Sboev) and V.I.; investigation, A.S. (Alexander Sboev), S.S. and I.M.; resources, A.S. (Alexander Sboev), R.R. and V.I.; data curation, S.S. and A.G.; writing—original draft preparation, A.S. (Alexander Sboev), R.R., A.G., I.M. and A.S. (Anton Selivanov); writing—review and editing, A.S. (Alexander Sboev) and R.R.; visualization, A.G., G.R. and I.M.; supervision, A.S. (Alexander Sboev); project administration, R.R. and I.M.; 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 20-11-20246.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be obtained by sending a request from the website of our project: https://sagteam.ru/en/med-corpus/, accessed on 12 December 2021; models will be presented on the page of our team on the Huggingface repository: https://huggingface.co/sagteam, accessed on 12 December 2021; code will be prepared and uploaded to the GitHub repository https://github.com/sag111, accessed on 12 December 2021.

Acknowledgments: 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 12 December 2021.

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

Abbreviations

The following abbreviations are used in this manuscript:

ADE	adverse drug event
ADR	adverse drug reaction
BiLSTM	bidirectional long short term memory
CNN	convolutional neural network
CRF	conditional random fields
CUI	controlled unique identifier from UMLS
DI	drug indications
ID	identifier
LSTM	long short term memory
NA	not available
NER	named entity recognition
NLP	natural language processing
PoS	part of speech
PV	pharmacovigilance
SSI	sign/Symptom/Illness
TA	Twitter annotated corpus
TP	TwiMED pubMed corpus
TT	TwiMED twitter corpus
TTR	type/token ratio
WD	withdrawal symptoms

Appendix A. ADR Recognition in the PsyTAR Corpus

For comparison purposes, we obtained the ADR recognition accuracy for the modification of the PsyTAR corpus [8] that contains sentences in the CoNLL format. It is publicly available (https://github.com/basaldella/psytarpreprocessor, accessed on 12 December 2021) and contains train, development and test parts of 3535, 431 and 1077 entities, respectively, and 3851, 551 and 1192 sentences, respectively. We used the XLM-RoBERTa-large model, for which we performed the fine-tuning only for the ADR tag, excluding the other tags WD, SSI and SD. The result on the test part was 71.1% according to the F1-exact metric described in Section 5.1.

Appendix B. Features Based on MESHRUS Concepts

MeSH Russian (MESHRUS) [68] is a Russian version of the Medical Subject Headings (MeSH) database (home page of the MeSH database website: https://www.nlm.nih.gov/mesh/meshhome.html, accessed on 12 December 2021). MeSH is a dictionary designed for indexing biomedical information that contains concepts from scientific journal articles and books, and is intended for their indexing and searching. The MeSH database is filled from articles in English; however, there exist translations of the database to different languages. We used the Russian version, MESHRUS. It is a less complete analogue of the English version: for example, it does not contain concept definitions. MESHRUS contains a set of tuples $(k; v)$ matching Russian concepts k with their relevant CUI codes v from the UMLS thesaurus. A concept k can consist of a word or a sequence of words.

The following pre-processing algorithm is used: words are lemmatized, put into a single register and filtered by length, frequency and parts of speech. In order to automatically find concepts from MESHRUS corresponding to words from our corpus, we perform two approaches.

The first approach is to map the filtered words $W = \{w_i\}_{i=0}^{N}$ from the corpus to MESHRUS concepts $\{C_j\}$. As a criterion for comparing words and concepts, we use the cosine similarity between their vector representations obtained using the FastText [45] model (see Section 4.2.4): a word w_i is assigned the CUI code v_j (see Figure A1) whose corresponding concept C_j has the highest similarity measure $\cos\big(\text{FastText}(w_i), \text{FastText}(C_j)\big)$. If this similarity measure is lower than the empirical threshold $T = 0.55$, no CUI code is assigned to w_i. Here, $\text{FastText}(C_j)$ is the vector representation of the output of concept C_j obtained by processing words of C_j to FastText, encoded as a sequence of n-grams.

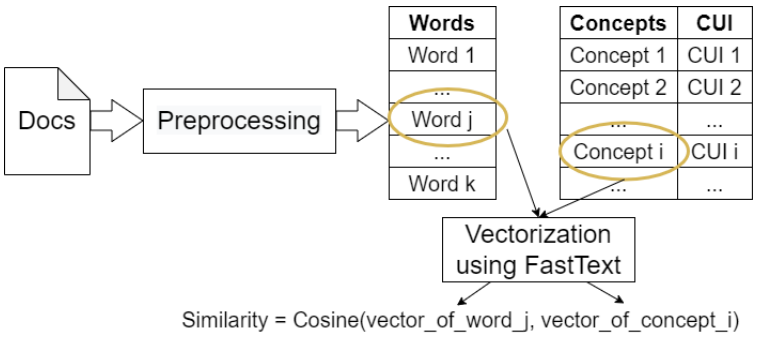

Figure A1. The matching scheme between words of corpus and concepts of UMLS.

The second approach is based on the mapping of syntactically and lexically related phrases extracted on the sentence level. Prepositions, particles and punctuation are not taken into account. Syntactic relations are obtained from dependency trees generated with UDpipe v2.5.

For each word $w_i \subset W$, its adjacent words $[w_{i-1}, w_{i+1}]$ are selected. Together with w_i itself, they form a lexical set w_{i_l}. Then, for the current word w_i, we find the word $w_{i_{\text{parent}}}$ that is its parent in the dependency tree (if there is no parent, then the syntactic set contains only w_i). These w_{i_l} and $w_{i_{\text{parent}}}$ in turn form a syntactic set w_{i_s}.

Similarly, such lexically and syntactically related sets c_{j_l} and c_{j_s} are formed for each filtered word c_j of the concept C_k from the MESHRUS dictionary: $c_{j_l} = [c_{j-1}, c_j, c_{j+1}]$ and $c_{j_s} = [c_j, c_{j_{\text{parent}}}]$.

Then, for each word $w_i \subset W$ and word $c_j \subset C_k$, by analogy with the literature [69], the following metrics are calculated:

1. $\text{lexical\_involvement}(w_i, c_j) = F_1\left(\frac{|w_{i_l} \cap c_{j_l}|}{|w_{i_l}|}, \frac{|w_{i_l} \cap c_{j_l}|}{|c_{j_l}|}\right)$;
2. $\text{cohesiveness}(w_i, c_j) = F_1\left(\frac{|w_{i_s} \cap c_{j_s}|}{|w_{i_s}|}, \frac{|w_{i_s} \cap c_{j_s}|}{|c_{j_s}|}\right)$;
3. centrality, which is 1 if the word $w_{i_{\text{parent}}}$ of the syntax set w_{i_s} is represented in the syntax set c_{j_s} of words from the dictionary; 0 otherwise.

Here, $F_1(x, y)$ is the harmonic mean of x and y, $|N|$ denotes the length of set N and $M \cap N$ is the intersection of the two sets. The final metric of similarity between the word w_i and the dictionary concept C_j is calculated as the mean of all three metric values.

For each word, its corresponding concept is selected by the highest similarity value provided that the similarity is greater than the specified threshold 0.6.

References

1. Helow, K.; Salem, A.B.M. Are Artificial Intelligence (AI) And Machine Learning (ML) Having An Effective Role In Helping Humanity Address The New Coronavirus Pandemic? *Wseas Trans. Biol. Biomed.* **2020**, *17*, 110–115. [CrossRef]
2. Madanan, M.; Sayed, B.; Akhmal, N.; Velayudhan, N. An Artificial Intelligence Approach Based on Hybrid CNN-XGB Model to Achieve High Prediction Accuracy through Feature Extraction, Classification and Regression for Enhancing Drug Discovery in Biomedicine. *Int. J. Biol. Biomed. Eng.* **2021**, *15*, 190–201. [CrossRef]
3. 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]
4. Alvaro, N.; Miyao, Y.; Collier, N. TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill.* **2017**, *3*, e6396. [CrossRef]
5. Sarker, A.; Nikfarjam, A.; Gonzalez, G. Social media mining shared task workshop. In *Biocomputing 2016: Proceedings of the Pacific Symposium*; World Scientific: Singapore , 2016; pp. 581–592.
6. Sarker, A.; Gonzalez, G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomed. Inform.* **2015**, *53*, 196–207. [CrossRef]
7. 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]
8. Basaldella, M.; Collier, N. BioReddit: Word embeddings for user-generated biomedical NLP. In Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), Hong Kong, China, 3 November 2019; pp. 34–38.
9. Henry, S.; Buchan, K.; Filannino, M.; Stubbs, A.; Uzuner, O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J. Am. Med Inform. Assoc.* **2019**, *27*, 3–12. [CrossRef]
10. 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*, 243–249. [CrossRef]
11. NEHTA. *Australian Medicines Terminology v3 Model–Common v1.4*; Tech. rep. EP-1825:2014; National E-Health Transition Authority: Rundle Mall, Australia, 2014.
12. Kuhn, M.; Letunic, I.; Jensen, L.J.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **2015**, *44*, D1075–D1079. [CrossRef]
13. Gupta, S.; Gupta, M.; Varma, V.; Pawar, S.; Ramrakhiyani, N.; Palshikar, G.K. Co-training for extraction of adverse drug reaction mentions from tweets. In *European Conference on Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 556–562.
14. Li, Z.; Yang, Z.; Wang, L.; Zhang, Y.; Lin, H.; Wang, J. Lexicon Knowledge Boosted Interaction Graph Network for Adverse Drug Reaction Recognition from Social Media. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 2777–2786. [CrossRef]
15. Wang, W. Mining adverse drug reaction mentions in twitter with word embeddings. In Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, Kohala Coast, HI, USA 4–8 January 2016.
16. Gupta, S.; Gupta, M.; Varma, V.; Pawar, S.; Ramrakhiyani, N.; Palshikar, G.K. Multi-Task Learning for Extraction of Adverse Drug Reaction Mentions from Tweets. In *European Conference on Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 59–71.
17. Li, Z.; Yang, Z.; Luo, L.; Xiang, Y.; Lin, H. Exploiting adversarial transfer learning for adverse drug reaction detection from texts. *J. Biomed. Inform.* **2020**, *106*, 103431. [CrossRef]

18. Sankaran, N.; Kaivalya, M.; Sreeranga, R.; Venkat, R. Evaluation of Transfer Learning for Adverse Drug Event (ADE) and Medication Entity Extraction. In Proceedings of the 3rd Clinical Natural Language Processing Workshop, Online, 19 November 2020.
19. Chiu, J.P.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
20. Magge, A.; Klein, A.; Miranda-Escalada, A.; Al-Garadi, M.A.; 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, Online, 10 June 2021 ; pp. 21–32.
21. Zhou, T.; Li, Z.; Gan, Z.; Zhang, B.; Chen, Y.; Niu, K.; Wan, J.; Liu, K.; Zhao, J.; Shi, Y.; et al. Classification, extraction, and normalization: Casia_unisound team at the social media mining for health 2021 shared tasks. In Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, Online, 10 June 2021; pp. 77–82.
22. Sakhovskiy, A.; Miftahutdinov, Z.; Tutubalina, E. KFU NLP Team at SMM4H 2021 Tasks: Cross-lingual and Cross-modal BERT-based Models for Adverse Drug Effects. In Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, Online, 10 June 2021; pp. 39–43.
23. El-karef, M.; Hassan, L. A Joint Training Approach to Tweet Classification and Adverse Effect Extraction and Normalization for SMM4H 2021. In Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, Online, 10 June 2021; pp. 91–94.
24. Dima, G.A.; Cercel, D.C.; Dascalu, M. Transformer-based Multi-Task Learning for Adverse Effect Mention Analysis in Tweets. In Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, Online, 10 June 2021; pp. 44–51.
25. Ji, Z.; Xia, T.; Han, M. PAII-NLP at SMM4H 2021: Joint Extraction and Normalization of Adverse Drug Effect Mentions in Tweets. In Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, Online, 10 June 2021; pp. 126–127.
26. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
28. Sboev, A.; Selivanov, A.; Rylkov, G.; Rybka, R. On the accuracy of different neural language model approaches to ADE extraction in natural language corpora. *Procedia Comput. Sci.* **2021**, *190*, 706–711. [CrossRef]
29. 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, 13 July 2012; pp. 1–40.
30. Webster, K.; Recasens, M.; Axelrod, V.; Baldridge, J. Mind the GAP: A Balanced Corpus of Gendered Ambiguou. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 605–617. [CrossRef]
31. Thompson, P.; Daikou, S.; Ueno, K.; Batista-Navarro, R.; Tsujii, J.; Ananiadou, S. Annotation and detection of drug effects in text for pharmacovigilance. *J. Cheminform.* **2018**, *10*, 37. [CrossRef]
32. Toldova, S.; Roytberg, A.; Ladygina, A.A.; Vasilyeva, M.D.; Azerkovich, I.L.; Kurzukov, M.; Sim, G.; Gorshkov, D.V.; Ivanova, A.; Nedoluzhko, A.; et al. RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian. In Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies "Dialogue", Bekasovo, Russia, 4–8 June 2014; pp. 681–695.
33. Ju, T.S. RU-EVAL-2019: Evaluating Anaphora And Coreference Resolution For Russian. In Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies "Dialogue", Moscow, Russia, 29 May–1 June 2019.
34. Lee, K.; He, L.; Lewis, M.; Zettlemoyer, L. End-to-end neural coreference resolution. *arXiv* **2017**, arXiv:1707.07045.
35. Lee, K.; He, L.; Zettlemoyer, L. Higher-order coreference resolution with coarse-to-fine inference. *arXiv* **2018**, arXiv:1804.05392.
36. Joshi, M.; Levy, O.; Weld, D.S.; Zettlemoyer, L. BERT for coreference resolution: Baselines and analysis. *arXiv* **2019**, arXiv:1908.09091.
37. Xu, L.; Choi, J.D. Revealing the myth of higher-order inference in coreference resolution. *arXiv* **2020**, arXiv:2009.12013.
38. Joshi, M.; Chen, D.; Liu, Y.; Weld, D.S.; Zettlemoyer, L.; Levy, O. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 64–77. [CrossRef]
39. Toshniwal, S.; Wiseman, S.; Ettinger, A.; Livescu, K.; Gimpel, K. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. *arXiv* **2020**, arXiv:2010.02807.
40. Rosminzdrav. State Register of Registered Drugs in Russia. Available online: http://grls.rosminzdrav.ru/ (accessed on 12 December 2021).
41. World Health Organization. International Statistical Classification of Diseases. Available online: https://icd.who.int/browse10/2019/en (accessed on 12 December 2021).
42. Miller, G.; Britt, H. A new drug classification for computer systems: The ATC extension code. *Int. J. Bio-Med. Comput.* **1995**, *40*, 121–124. [CrossRef]
43. Ratcliff, J.W.; Metzener, D.E. Pattern-matching-the gestalt approach. *Dobbs J.* **1988**, *13*, 46.

44. Straka, M.; Hajic, J.; Strakov'a, J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016.
45. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
46. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv* **2019**, arXiv:1911.02116.
47. Suero Montero, C.; Munezero, M.; Kakkonen, T. Investigating the role of emotion-based features in author gender classification of text. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Kathmandu, Nepal, 6–12 April 2014; pp. 98–114. [CrossRef]
48. Tausczik, Y.R.; Pennebaker, J.W. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **2010**, *29*, 24–54. [CrossRef]
49. Litvinova, O.; Seredin, P.; Litvinova, T.; Lyell, J. Deception detection in Russian texts. In Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; pp. 43–52.
50. Sboev, A.; Gudovskikh, D.; Rybka, R.; Moloshnikov, I. A quantitative method of text emotiveness evaluation on base of the psycholinguistic markers founded on morphological features. *Procedia Comput. Sci.* **2015**, *66*, 307–316. [CrossRef]
51. Tolmachova, E. *VIDAL: Directory of medicines in Russia*; Vidal Rus VIDAL: Directory of medicines in Russia. Available online: https://www.vidal.ru/ (accessed on 12 December 2021).
52. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
53. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
54. Schuster, M.; Nakajima, K. Japanese and Korean voice search. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 5149–5152.
55. Kudo, T.; Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv* **2018**, arXiv:1808.06226.
56. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
57. Gal, Y.; Ghahramani, Z. A theoretically grounded application of dropout in recurrent neural networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1019–1027.
58. Dumoulin, V.; Visin, F. A guide to convolution arithmetic for deep learning. *arXiv* **2016**, arXiv:1603.07285.
59. Boureau, Y.L.; Ponce, J.; LeCun, Y. A theoretical analysis of feature pooling in visual recognition. In Proceedings of the 27th International conference on machine learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 111–118.
60. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289. [CrossRef]
61. Rajapakse, T.C. Simple Transformers. 2019. Available online: https://github.com/ThilinaRajapakse/simpletransformers (accessed on 12 December 2021).
62. Biewald, L. Experiment Tracking with Weights and Biases, 2020. Available online: wandb.com (accessed on 12 December 2021).
63. Moosavi, N.S.; Strube, M. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 632–642.
64. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning Word Vectors for 157 Languages. In Proceedings of the International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018.
65. 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.
66. Koehn, P. Statmt—Internet Resource about Research in the Field of Statistical Machine Translation. Available online: https://www.statmt.org/ (accessed on 12 December 2021).
67. Smith, L.N. Cyclical learning rates for training neural networks. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 464–472.
68. System Center Service Manager Library. Russian Version of the Medical Subject Headings (MeSH) Database. Available online: https://www.nlm.nih.gov/mesh/meshhome.html (accessed on 12 December 2021).
69. Shelmanov, A.; Smirnov, I.; Vishneva, E. Information extraction from clinical texts in Russian. In Proceedings of the International Conference on Computer Linguistics and Intellectual Technologies "Dialogue", Moscow, Russia, 27–30 May 2015; Volume 17.