Bioinformatics, 37(2), 2021, 243–249
doi: 10.1093/bioinformatics/btaa675
Advance Access Publication Date: 28 July 2020
Original Paper

Data and text mining

The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews

Elena Tutubalina [1,*], Ilseyar Alimova [1,*], Zulfat Miftahutdinov [1], Andrey Sakhovskiy [1], Valentin Malykh [1] and Sergey Nikolenko [1,2]

[1]Chemoinformatics and Molecular Modeling Laboratory, The Alexander Butlerov Institute of Chemistry, Kazan Federal University, Kazan 420008, Russian Federation and [2]Samsung-PDMI AI Center, Steklov Institute of Mathematics at St. Petersburg, St. Petersburg 191023, Russian Federation

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren
Received on April 8, 2020; revised on July 14, 2020; editorial decision on July 17, 2020; accepted on July 20, 2020

Abstract

Motivation: Drugs and diseases play a central role in many areas of biomedical research and healthcare. Aggregating knowledge about these entities across a broader range of domains and languages is critical for information extraction (IE) applications. To facilitate text mining methods for analysis and comparison of patient's health conditions and adverse drug reactions reported on the Internet with traditional sources such as drug labels, we present a new corpus of Russian language health reviews.

Results: The Russian Drug Reaction Corpus (RuDReC) is a new partially annotated corpus of consumer reviews in Russian about pharmaceutical products for the detection of health-related named entities and the effectiveness of pharmaceutical products. The corpus itself consists of two parts, the raw one and the labeled one. The raw part includes 1.4 million health-related user-generated texts collected from various Internet sources, including social media. The labeled part contains 500 consumer reviews about drug therapy with drug- and disease-related information. Labels for sentences include health-related issues or their absence. The sentences with one are additionally labeled at the expression level for identification of fine-grained subtypes such as drug classes and drug forms, drug indications and drug reactions. Further, we present a baseline model for named entity recognition (NER) and multilabel sentence classification tasks on this corpus. The macro F1 score of 74.85% in the NER task was achieved by our RuDR-BERT model. For the sentence classification task, our model achieves the macro F1 score of 68.82% gaining 7.47% over the score of BERT model trained on Russian data.

Availability and implementation: We make the RuDReC corpus and pretrained weights of domain-specific BERT models freely available at https://github.com/cimm-kzn/RuDReC.

Contact: elvtutubalina@kpfu.ru or AlimovaIlseyar@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In this work, we describe the design, composition and construction of a large dataset of user-generated texts (UGTs) about pharmaceutical products in Russian. Similar to the Food and Drug Administration in the USA and the Therapeutic Goods Administration in Australia, the Federal Service for Surveillance in Healthcare (*Roszdravnadzor*) in Russia accumulates data provided by volunteer reports on the risks of taking various medicines to ensure their safe use. Since some particular medications may interact with others in a non-obvious way, creating and using such resources leads to significant difficulties. Information from online sources is considered to be a valuable source for *Roszdravnadzor* or pharmaceutical companies to correct the use of a drug when necessary. Thus, our corpus has been designed with the explicit purpose to facilitate the methods for learning complex knowledge of primary interactions between different drugs, diseases and adverse reactions.

Figure 1 shows a brief overview of our study. The corpus, which we call the *Russian Drug Reaction Corpus* (RuDReC), contains an aggregation of texts of the patients' feedback on the use of drugs in various therapeutic groups or their experience with the healthcare system in general; we have taken care to ensure that we have collected representative samples intended for training advanced

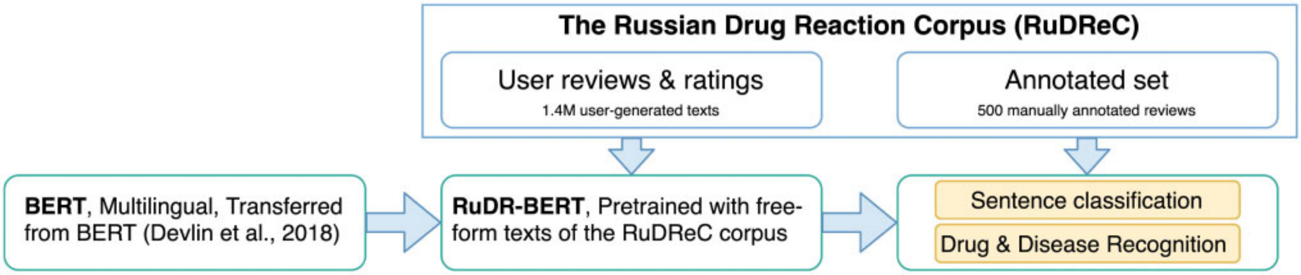

The Russian Drug Reaction Corpus (RuDReC)
User reviews & ratings
1.4M user-generated texts
Annotated set
500 manually annotated reviews
BERT, Multilingual, Transferred from BERT (Devlin et al., 2018)
RuDR-BERT, Pretrained with free-form texts of the RuDReC corpus
Sentence classification
Drug & Disease Recognition

Fig. 1. Overview of our study: (i) creating the raw and annotated parts of the *RuDReC* corpus, (ii) training a domain-specific version of BERT (RuDR-BERT) on collected texts and (iii) developing baselines and presenting evaluation results

machine learning methods. Recent advances in deep contextualized representations via language models such as BERT (Devlin *et al.*, 2019) or domain-specific biomedical models such as BioBERT (Lee *et al.*, 2020) offer new opportunities to improve the models for classification and entity recognition. Our primary goal has been, therefore, to construct a large (partially) annotated corpus to stimulate the development of automated text-mining methods for finding meaningful information in the patients' narratives in the Russian language.

The RuDReC corpus is meaningfully divided into two parts that are very different in size. The larger part is a raw corpus of 1.4M health comments that can be used to train modern distributed semantics models whose training is based on self-supervised objectives such as the next token prediction (as in, e.g. *word2vec*) or predicting masked tokens (as in, e.g. BERT). The second, smaller part, contains 500 richly annotated reviews to allow the training of downstream task-specific models. The primary downstream tasks in our case are named entity recognition (NER) and multilabel classification. The labeling in the second part consists of two main components: sentence labels and entity labels. We have split the review posts into sentences and labeled them for the presence of drug indications (DI) and symptoms of a disease, adverse drug reactions (ADR), drug effectiveness (DE), drug ineffectiveness (DIE). In the entity identification phase, we identified and extracted 6 entity types: drug names, drug classes, drug forms, ADR, DI and Findings. In total, we have labeled 2202 sentences and 4566 entities.

The resulting dataset and pretrained weights of domain-specific BERT have been made freely available for researchers at https://github.com/cimm-kzn/RuDReC. We hope that this new resource will intensify research on multilingual IE on adverse drug events and DE based on the data from patient narratives. The paper is organized as follows: Section 2 discusses related work; Section 3 introduces the RuDReC corpus, describes it qualitatively and quantitatively and shows the details of model training; Section 4 presents the results of our evaluation across two downstream tasks (sentence classification and NER), Section 5 shows some limitations of our approach, and Section 6 concludes the article.

2 Related work

Many systems for disease and chemical entity recognition from scientific texts have been developed over the past 15 years. This task is traditionally formulated as a sequence labeling problem and solved with Conditional Random Fields (CRF) that use a wide variety of features: individual words or lemmas, part-of-speech tags, suffixes and prefixes, dictionaries of medical terms, cluster-based and distributed representations and others (Gu *et al.*, 2016; Lee *et al.*, 2016; Miftahutdinov *et al.*, 2017).

In contrast to biomedical literature, research into the processing of UGTs about drug therapy has not reached the same level of maturity. Starting from 2014, some studies began to use the powers of social media and deep learning (especially suitable for training on large available datasets that are the main advantage of using UGTs) for pharmacovigilance purposes; in particular, researchers have considered the problems of text (post) classification and extraction of ADRs (Alvaro *et al.*, 2017; Gonzalez-Hernandez *et al.*, 2018; Karimi *et al.*, 2015; Zolnoori *et al.*, 2019). Recent studies primarily use neural architectures; in particular, Tutubalina and Nikolenko (2017), Dang *et al.* (2018) and Giorgi and Bader (2019) exploited LSTM-CRF models with domain-specific word embeddings, while Miftahutdinov *et al.* (2020) and Lee *et al.* (2020) used BERT-based architectures for NER.

The CSIRO Adverse Drug Event Corpus (CADEC) dataset collected by Karimi *et al.* (2015) became a *de facto* standard for the extraction of health-related entities such as ADRs from user reviews. It contains 1253 medical forum posts taken from the *AskaPatient* web portal about 12 drugs divided into two categories: *Diclofenac* and *Lipitor*. All posts were annotated manually by medical students and computer scientists who labeled five types of entities, including ADRs and names of medicines or drugs. Average inter-annotator agreement rates computed over a subset of 55 user posts with related span matching and tag settings showed that agreement across four annotators in a subset of *Diclofenac* posts was approximately 78%, while the agreement between two annotators in a subset of *Lipitor* posts was approximately 95%.

The Psychiatric Treatment Adverse Reactions (PsyTAR) corpus (Zolnoori *et al.*, 2019) is also an open source corpus of UGTs taken from *AskaPatient*. This dataset includes 887 posts about four psychiatric medications from two classes: (i) *Zoloft* and *Lexapro* from the Selective Serotonin Reuptake Inhibitor (SSRI) class and (ii) *Effexor* and *Cymbalta* from the Serotonin Norepinephrine Reuptake Inhibitor (SNRI) class. In contrast with the CADEC dataset, first, the authors labeled sentences in the posts for the presence of ADRs, withdrawal symptoms (WD), sign/symptoms/illness (SSI), DI, DE and DIE. Second, sentences were annotated with four types of entities: ADR, WD, DI, SSI. Two of the annotators were pharmacy students, and two annotators had a background in health sciences. The resulting pairwise agreement for a strict match was 0.86 for the entire dataset, ranging from 0.81 for the WD class to 0.91 for DI.

The *Twitter and PubMed Comparable* corpus (TwiMed) (Alvaro et al., 2017) is the only open source corpus that contains two sources of information annotated at the entity level by the same experts (pharmacists) using the same set of guidelines. This dataset includes 1000 tweets and 1000 PubMed sentences retrieved using a set of 30 different drugs. This corpus contains annotations for 3144 entities (drugs, symptoms, and diseases), and 5003 attributes of entities (polarity, person, modality, exemplification, duration, severity, status, sentiment). In this case, there was a lower agreement in the annotation of tweets than in the annotation of PubMed sentences, most likely due to the noisy nature of tweets. The annotators did not perform terminology association. We note that the total number of sentences and tweets in the TwiMed corpus is three times smaller than in the CADEC and PsyTAR corpora.

To sum up, most existing research on information retrieval for drug-related events deals with PubMed abstracts, user reviews, tweets and clinical records in English (Alvaro *et al.*, 2017; Karimi *et al.*, 2015; Zolnoori *et al.*, 2019). Supplementary Table S1 presents basic statistics of existing relevant corpora.

There exist very few Russian corpora with annotations of the presence of drug reactions at the level of sentences. Alimova *et al.*

(2017) proposed a Russian corpus of user reviews from *Otzovik.com* with four types of sentence annotations: indication, beneficial effect, ADR, other. Shelmanov et al. (2015) created a corpus of clinical notes in the Russian language. The corpus contains 112 fully annotated texts from a multi-disciplinary pediatric center. Recently, the SMM4H 2020 Task (https://healthlanguageprocessing.org/smm4h-sharedtask-2020/) presented a multilingual corpus of tweets (including Russian-language tweets) annotated with the presence of ADRs. To our knowledge, the RuDReC corpus is the first large (partially) annotated corpus of health-related UGTs in Russian.

3 The RuDReC corpus

Our goal in this work is threefold:

1. create an open access corpus, which we call RuDReC that would conform to annotation guidelines based on the annotators' insights and existing English corpora such as CADEC and PsyTAR;
2. collect a large dataset of free-form health-related UGTs to ensure diversity of drug classes that are defined by their therapeutic use;
3. develop a domain-specific language representation model, pre-trained on the raw texts from the collected corpus and baselines for sentence classification and entity recognition tasks.

Our manually annotated corpus contains five sentence labels and six different entity types, as shown in Tables 1 and 2, respectively.

Figure 2 shows sample annotations produced using INCEpTION as the annotation platform (Klie *et al.*, 2018). It is important to note that we have obtained all reviews without accessing password-protected information; all data from our corpus are publicly available on the Internet.

3.1 Annotation

3.1.1 Data source

For the annotation process, we have used user posts in Russian from a popular and publicly accessible source *Otzovik.com*, which collects the patients' self-reported experiences for a wide range of medications. Each user fills out a form containing the drug description (including the reason for taking it), drug class, year of purchase, its route of administration, perceived efficiency and side effects and information about the disease. Users are also asked to rate the overall drug satisfaction from one (low) to five (high). The reviews are written in Russian; as is usually the case with UGTs, they do not necessarily have perfect grammar and may contain informal language patterns specific for different regions of Russia and other Russian-speaking countries.

3.1.2 Annotation guidelines

Our annotation process consisted of two stages. At the first stage, annotators with a background in pharmaceutical sciences were asked to read 400 reviews and highlight all spans of text, including drug names and patient's health conditions experienced before, during, or after the drug use. The objective of the first stage of the annotation process was to perform preliminary annotation across a set of reviews to choose the best annotation scheme. The authors informed the annotators with an analysis of existing annotation schemes for English language corpora (Alvaro *et al.*, 2017; Karimi *et al.*, 2015). At the second stage, annotators were asked to screen existing annotations and annotate new texts on an extended set of reviews.

At the first stage, the process of identification and extraction of entities' spans was conducted by four annotators with a background in pharmaceutical sciences from the I.M. Sechenov First Moscow State Medical University. Our analysis of existing corpora shows two main types of entities common to all schemes: Drug and Disease. After several discussions, annotators defined the following Disease subtypes: (i) disease name; (ii) indication (Indication); (iii) positive dynamics after or during taking the drug (BNE-Pos); (iv) negative dynamics after the start or some period of using the drug (ADE-Neg); (v) the drug does not work after taking the course (NegatedADE); (vi) deterioration after taking a course of the drug (Worse). As Drug subtypes, annotators have chosen: (i) drug names, (ii) drug classes and (iii) drug forms.

The posts were divided between the annotators, and 100 documents and annotation guidelines were given to another annotator from the Department of Pharmacology of the Kazan Federal University for the purpose of calculating the interannotator agreement. We note that this annotator did not interact with other annotators in discussions about the annotation scheme. Two metrics were used in our calculation of relaxed agreement for *Disease* and *Drug* entities, as described by Karimi *et al.* (2015). When annotation and span settings were both relaxed, the average agreement was approximately 70%.

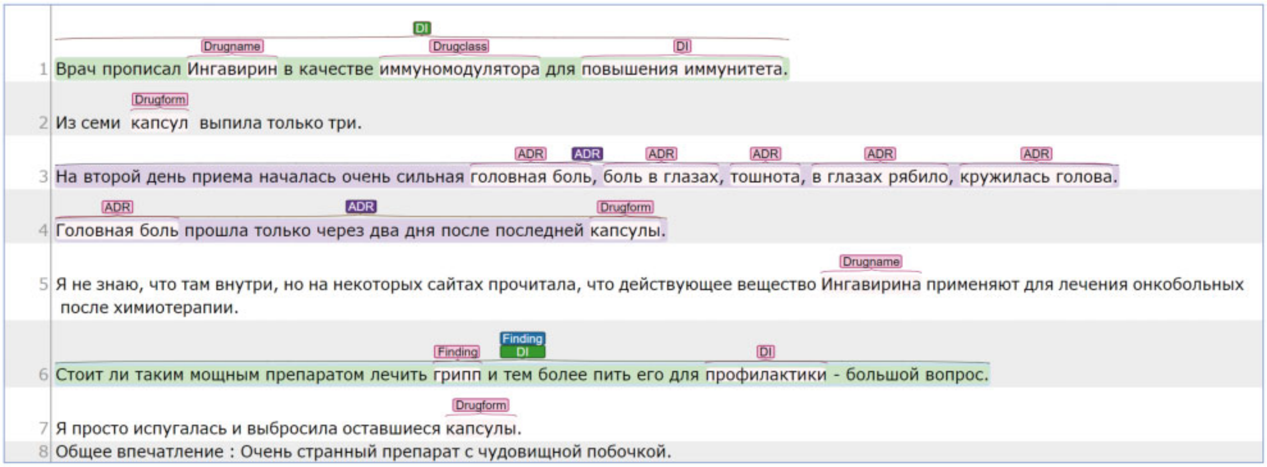
After completing the annotation process at the first stage, three of the authors screened the annotations. We came to several conclusions based on the results. First, there were relatively few examples of Worse and ADE-Neg types (198 examples in total). Second, entities of ineffective type were longer in comparison with other entity types: the average length of ineffective type entities was 15 words,

Table 1. Definitions for sentence labels annotated in the patients' comments

Sentence label	Definition
DE	A sentence is labeled as DE if it contains an explicit report about treated symptoms or that the patient's condition has improved after drug use.
DIE	A sentence is labeled as DIE if it contains a direct report that the patient's health status became worse or did not change after the drug usage.
DI	A sentence is labeled as DI if it contains any indication/symptom that specifies the reason for taking/prescribing the drug.
ADR	A sentence is labeled as ADR if it contains mentions of undesirable, untoward medical events that occur as a consequence of drug intake.
FINDING	A sentence is labeled as Finding if it describes disease-related events that are not experienced or denied by the reporting patient or his/her family members. These sentences often describe a patient's medical history, drug label, or absence of expected drug reactions.

Table 2. Definitions for entity types identified in patient comments

Entity type	Definition
DRUG NAME	Mentions of the brand name of a drug or product ingredients/active compounds.
DRUG CLASS	Mentions of drug classes such as *anti-inflammatory* or *cardiovascular*.
DRUG FORM	Mentions of routes of administration such as *tablet* or *liquid* that describe the physical form in which medication will be delivered into patient's organism.
DI	Any indication/symptom that specifies the reason for taking/prescribing the drug.
ADR	Mentions of untoward medical events that occur as a consequence of drug intake and are not associated with treated symptoms.
FINDING	Any DI or ADR that was not directly experienced by the reporting patient or his/her family members, or related to medical history/drug label, or any disease entities if the annotator is not clear about type.



1 Врач прописал Ингавирин в качестве иммуномодулятора для повышения иммунитета.

2 Из семи капсул выпила только три.

3 На второй день приема началась очень сильная головная боль, боль в глазах, тошнота, в глазах рябило, кружилась голова.

4 Головная боль прошла только через два дня после последней капсулы.

5 Я не знаю, что там внутри, но на некоторых сайтах прочитала, что действующее вещество Ингавирина применяют для лечения онкобольных после химиотерапии.

6 Стоит ли таким мощным препаратом лечить грипп и тем более пить его для профилактики - большой вопрос.

7 Я просто испугалась и выбросила оставшиеся капсулы.

8 Общее впечатление : Очень странный препарат с чудовищной побочкой.

Fig. 2. Example of sentence and entity annotation

while, e.g. ADRs had an average of 5 words. Finally, the BNE-Pos entity types contained a lot of overly broad entities that were not related to medical concepts, such as ‘helped’, ‘effective’ and so on.

To mitigate these problems, we made several changes to the annotation scheme. First, we combined *Worse* and *ADE-Neg* with *NegatedADE* entity types into a single class DIE and spanned DIE annotation on the sentence level, similar to the PsyTAR corpus. Second, we spanned *BNE-Pos* entities on the sentence level and renamed them to DE, also in agreement with the PsyTAR corpus. Finally, following the CADEC corpus, we combined the *Indication* and *Disease* entity types into a single DI type.

At the second stage, two annotators from the Kazan Federal University were asked to continue the annotation process according to sentence classes and entity types presented in Tables 1 and 2. After completing the annotation process, two of the authors screened the annotations to correct span mistakes.

3.2 Analysis of the annotated set

Our dataset includes reviews about four groups of drugs:

1. sedatives (brain and nervous system);
2. nootropics (brain and nervous system);
3. immunomodulators (immune disease);
4. antivirals (infections).

Sedatives and nootropics both belong to the neurotropic group of drugs, i.e. drugs that have an effect on the central and peripheral nervous systems. This group includes antidepressants, mood stabilizers, nootropics and sedatives. Immunomodulators, in particular, immunostimulants and immunosuppressants, are substances that modify the immune response and affect immunocompetent cells. Antiviral drugs are intended for the treatment of various viral diseases (influenza, herpes, HIV infection, etc.); they are also used for preventive purposes.

The annotated corpus consists of 500 reviews about drugs from these four groups. Reviews were selected randomly for annotation. The examples of annotated entities for each group are presented in Supplementary Table S2.

Supplementary Figure S1 presents statistics on therapeutic groups. Every user fills out this information as well as 5-star ratings when writing a review. The majority of the reviews (60%) are describing the antiviral drugs, which are of the most common ones used in everyday life. The second by number group is sedatives and antidepressants (27%), which are on the raise in recent years. Supplementary Figure S2 presents statistics on ratings in our corpus. Another interesting feature is that the prevalence of the highest rating (5) is not overpowering the other ratings, which are more or less uniformly distributed. This is a common feature that the intermediate rating is mostly skipped in many domains, but the collected data is showing unusual uniformity.

Table 3 lists the statistics for the annotated corpus part as a whole, as well as one for each group of drugs. There are several interesting features one could note here. First of all, immunomodulatory drugs have longer reviews in terms of both the sentences and tokens. The average length is 30% larger than for any other group, and the maximal length is up to twice larger, although the minimal length is the same as for other groups. Second, the average number of sentences in Russian reviews is higher than in the English CADEC and PsyTAR corpora (9.71 versus 6).

Table 4 presents the frequency of annotated sentences in the entire corpus as well as in each drug group. There are several features that should be mentioned regarding these annotations. There are interesting disproportionalities in the frequencies (normalized columns) of different types of labels. The immunomodulators group has the lowest representation of ADRs, while the antidepressants (sleeping) have the highest one.

Table 5 presents the statistics of annotated entities in the entire corpus as well as in each drug group. The drug class and drug form labels are surprisingly scarce in the nootropic group. The most common among others DI class is in the antidepressant group. The analysis of part of speech (PoS) tags of each word in entities showed that users in social media use more verbs to express symptoms and ADRs in comparison to formal medical concepts. In the annotated part of the RuDREC corpus, 18.26% of disease-related entities’ words are verbs, while only 2.53% words, included in the MedDRA dictionary from UMLS v. 2020AA, are verbs.

3.3 A large collection of health reviews

Text collections used for training domain-specific BERT were obtained by web page crawling. User reviews were collected from the following popular medical web portals. These online resources mostly contain drug reviews about pharmaceutical products, health facilities and pharmacies. Duplicate comments were removed. The statistics on this part of the RuDReC corpus are given in Table 6. The collection contains 1.4 million of patient narrative texts, 1 104 054 unique tokens and 193 529 197 tokens in total.

3.4 Pretraining and fine-tuning domain-specific BERT

We used the multilingual version of BERT-base (Multi-BERT) as initialization for training domain-specific BERT further called **RuDR-BERT**.

Similar to the study by Lee *et al.* (2020), we observed that 800K and 840K pretraining steps were sufficient. This roughly corresponds to a single epoch on each corpus. The batch size was set to 32 examples. Other hyperparameters such as learning rate scheduling for pretraining RuDR-BERT are the same as those for Multi-

Table 3. Basic statistics on reviews, sentences and tokens

	Entire Corpus	Sedatives	Immunomodulators	Nootropics	Antivirals
No. of reviews	500	90	67	46	297
Total no. of sentences	4855	829	813	410	2803
Avg no. of sentences in each review	9.71	9.21	12.13	8.91	9.44
No. of sentences in each review (range)	1–35	2–22	2–35	3–17	1–25
Total no. of tokens	68036	11536	12217	5930	38353
Avg no. of tokens (words) in each review	136.07	128.17	182.34	128.91	129.13

Table 4. Number of sentences annotated in the entire corpus and each therapeutic group

	Entire Corpus		Sedatives		Immunomodulators		Nootropics		Antivirals	
	Raw	Norm.	Raw	Norm.	Raw	Norm.	Raw	Norm.	Raw	Norm.
DI	949	1.90	182	2.02	132	1.97	83	1.80	552	1.86
ADR	379	0.78	100	1.11	27	0.40	42	0.91	210	0.71
FINDING	172	0.34	36	0.40	25	0.37	20	0.43	91	0.31
DE	424	0.85	86	0.96	69	1.03	53	1.15	216	0.73
DIE	278	0.56	45	0.50	35	0.52	26	0.57	172	0.58
All	2202	4.40	449	4.99	288	4.30	224	4.87	1241	4.18

Table 5. Number of entities annotated in the entire corpus and each therapeutic group

	Entire Corpus		Sedatives		Immunomodulators		Nootropics		Antivirals	
	Raw	Norm.	Raw	Norm.	Raw	Norm.	Raw	Norm.	Raw	Norm.
DRUG NAME	1043	2.07	200	2.22	151	2.25	95	2.07	597	2.01
DRUG CLASS	330	0.66	79	0.88	64	0.96	8	0.17	179	0.60
DRUG FORM	836	1.67	155	1.72	163	2.43	35	0.76	483	1.63
DI	1401	2.80	293	3.26	191	2.85	116	2.52	801	2.70
ADR	720	1.44	202	2.24	43	0.64	93	2.02	382	1.29
FINDING	236	0.47	50	0.56	33	0.49	26	0.54	127	0.43
All	4566	9.13	979	10.88	645	9.62	372	8.09	2570	8.65

Table 6. Text collection statistics for web-based comments

Category for reviewing	Written by	Number of texts
Pharmaceutical products	users	261983
Beauty products	users	466199
Drugs	doctors	7451
Drugs	users	31500
Health facilities and pharmacies	users	642178
Total		1409311

BERT unless stated otherwise. We decided to adopt the initial vocabulary of Multi-BERT for preprocessing in both pretraining corpora and fine-tuning sets. The language model was fine-tuned using a BERT implementation from https://github.com/google-research/bert. We trained **RuDR-BERT** on a single machine with 8 NVIDIA P40 GPUs. The training of all models took approximately 8 days.

We fine-tuned several BERT models, including RuDR-BERT, on two tasks:

i. NER (with entity types as shown in Table 5);
ii. sentence classification (the classes are presented in Table 4).

Following our previous work on NER (Miftahutdinov *et al.*, 2020), we use different BERT models with a softmax layer over all possible tags as the output for NER. Word labels are encoded with the BIO tag scheme. We note that the model was trained on the sentence level. All NER models were trained without an explicit selection of parameters on the RuDReC corpus. The loss function became stable (without significant decreases) after 35–40 epochs. We use Adam optimizer with polynomial decay to update the learning rate on each epoch with warm-up steps in the beginning. For sentence classification, we use the Tensorflow implementation of BERT with sigmoid activation over dense output layer and cross-entropy loss function. For each label, we used the sigmoid value of 0.5 as a classification threshold. We fine-tuned each model for 10 epochs with a batch size of 16. We defined the first 10% of the training steps as warm-up steps.

4 Experiments and evaluation

For our experiments, we used three versions of BERT:

1. BERT$_{base}$, the Multilingual Cased (Multi-BERT) pretrained on 104 languages; it has 12 heads, 12 layers, 768 hidden units per layer and a total of 110M parameters;
2. RuBERT, the Russian Cased BERT pretrained on the Russian part of Wikipedia and news data; it has 12 heads, 12 layers, 768 hidden units per layer and a total of 180M parameters; Multi-BERT was used for initialization, while the vocabulary of Russian subtokens was built on the training dataset (Kuratov and Arkhipov, 2019);
3. RuDR-BERT, Multilingual Cased BERT pretrained on the raw part of the RuDReC corpus (1.4M reviews); Multi-BERT was used for initialization, and the vocabulary of Russian subtokens and parameters are the same as in Multi-BERT.

Table 7. Performance of fine-tuned RuDR-BERT on sentence classification with comparison to Multi-BERT and RuBERT, measured by F1-score

Model	DE	DIE	ADR	DI	Finding	Macro F1-score
RuBERT	67.7±2.82	62.27±3.47	66.65±2.96	81.63±2.38	28.51±4.8	61.35±3.28
Multi-BERT	63.61±4.22	60.19±3.52	63.45±2.61	79.58±4.1	24.32±2.85	58.23±3.46
RuDR-BERT	76.61±4.08	72.06±5.29	74.15±5.01	85.06±2.49	36.24±6.91	68.82±4.76

Table 8. Performance of fine-tuned RuDR-BERT on the NER task in comparison with Multi-BERT and RuBERT, measured by F1-score with exact matching criteria

Model	ADR	DI	Finding	Drug class	Drug form	Drug name	Macro F1-score
RuBERT	54.51±3.9	69.43±4.98	27.87±5.92	92.78±1.14	95.72±1.38	92.11±1.56	72.07±2.03
Multi-BERT	54.65±2.38	67.63±3.62	25.75±7.86	92.36±2.72	94.89±0.97	91.05±0.61	71.06±2.46
RuDR-BERT	60.36±2.13	72.33±2.12	33.31±7.55	94.12±2.31	95.89±1.82	93.08±1.08	74.85±2.09

4.1 Multilabel sentence classification

We compare all models on fivefold cross validation in terms of F1 score. The fine-tuning of each model took approximately 1 h on one NVIDIA GTX 1080 Ti GPU.

Table 7 performs the result of RuBERT, Multi-BERT and fine-tuned RuDR-BERT models in terms of F1 score. According to the results, the following conclusions can be drawn. First, the RuDR-BERT model achieved the best results among other comparable models. Second, the RuBERT model outperformed the Multi-BERT model on 3.12% in terms of the macro F1 score. The highest improvement was achieved for DE (+4.09%) and Finding entity types (+4.19%). Third, the performance of RuDR-BERT on Finding (36.24%) is significantly lower than on ADR (74.15%) and DI (85.06%). It could be explained by similar contexts and a much lower number of training examples.

4.2 Drug and disease recognition

We compare all models on fivefold cross -validation in terms of F1 scores computed by exactly matching criteria via a CoNLL script. We trained each model on a single machine with 8 NVIDIA P40 GPUs. The training of all models took approximately 10 h.

Table 8 shows the performance of RuBERT, Multi-BERT and fine-tuned RuDR-BERT in terms of the F1 score. There are several conclusions to be drawn based on the results in these tables. First, on all types of entities, the domain-specific RuDR-BERT achieves better scores than both RuBERT and Multi-BERT. Second, RuBERT, with a vocabulary of Russian subtokens generated on Wikipedia and news, outperforms Multi-BERT. Third, similar to sentence classification, the performance of RuDR-BERT on Finding is significantly lower than on ADR and DI. Finally, all models achieve much higher performance for the detection of drugs rather than diseases; it can be explained by boundary problems in multi-word expressions. In particular, RuDR-BERT achieves the F1 score of 81.34% on disease-related entities and F1 score of 94.65% of drug-related entities. To obtain metrics for disease-related entities, we replaced ADR, DI and Finding entity types with *Disease* entity type in the gold standard and predicted data. The same procedure was done for drug-related entities except that Drug name, Drug form and Drug class were replaced by *Drug*. The average number of tokens on drug-related entities is 1.06, while the average number of tokens on disease-related entities is 1.77.

5 Limitations

There are several issues that may potentially limit the applicability of RuDReC; they are mostly shared with other available datasets.

Validation of drugs by the state register of medicines. We believe that automatic systems for extracting meaningful information concerning pharmaceutical products should validate whether the pharmaceutical products have registered with the State Register of Medicines (https://grls.rosminzdrav.ru/). The State Register of Medicines is a list of domestic and foreign medicines, medical prophylactic and diagnostic products registered by the Ministry of Health of Russia. Our annotator from the Department of Pharmacology of Kazan Federal University conducted a manual study of 649 unique product names that review authors put as review titles in their free-form reviews, checking whether the drugs were present in the State Register of Medicines for each product name. The results of this labeling showed that 373 (57.5%) of the names do have a match in the system and belong to one of the groups from the Anatomical Therapeutic Chemical (ATC) Classification System (J0, D0, G0, A0). Note that, this is a preliminary result, and it has not been validated with multiple annotators; however, it indicates the need for an additional validation step for automatic systems.

Normalization challenge. There are three major international terminologies for the Russian language: Medical Dictionary for Regulatory Activities (MedDRA), Medical Subject Headings (MeSH) thesaurus and International Classification of Diseases (ICD). One challenge is that layperson expressions of disease-related words are fuzzier and broader than the corresponding MedDRA terms. Another challenge is that social media patients discuss different concepts of illness and a wide diversity of drug reactions. Moreover, social network data usually contains a lot of noise, such as misspelled words, incorrect grammar, hashtags, abbreviations and different variations of the same word. In our dataset, there is no mapping of entity mentions to formal medical terminology, which we leave as future work.

The risk of fake reports on the Internet. A recent study by Smith *et al.* (2018) demonstrates that it is possible to harvest and compare ADRs found in social media with those from traditional sources. One major challenge for automatic methods is fact checking. A similar research question is currently being investigated in the *CLEF-2020 CheckThat! Shared Task 1* that deals with whether a given tweet is trustworthy, i.e. whether it is supported by factual information (the task uses a sample of tweets about COVID-19).

Robustness of trained models. Our annotated corpus for training NER and classification models includes reviews on several therapeutic groups, but it may not be representative of drugs from other classes, for example, antineoplastic agents. On the other hand, the RuDReC corpus includes a large collection of 1.4M user-generated health reviews about a large assortment of pharmaceutical products and patient experience with hospital care that could improve the robustness of language models.

6 Conclusion

In this work, we present a new open access corpus named RuDReC for researchers of biomedical natural language processing and pharmacovigilance. In this article, we have discussed the challenges of annotating health-related Russian comments and have presented several baselines for the classification and extraction of health entities. The RuDReC corpus provides opportunities for researchers in a number of areas to:

1. develop and evaluate text-mining models for gathering of meaningful information about DE and ADRs from layperson reports;
2. analyze and compare variations of reported patient health conditions and drug reactions of different therapeutic groups of medications with drug labels.

We foresee three directions for future work. First, transfer learning and multitask strategies on several tasks on English and Russian texts remain to be explored. Second, a promising research direction is to try pretraining domain-specific BERT-based models with a custom vocabulary. Third, future research will focus on the creation of mapping between entity mentions and existing multilingual terminologies such as MedDRA and MeSH.

Acknowledgements

The authors thank all annotators for their contribution. They also thank Timur Madzhidov and Valery Solovyev for valuable feedback.

Funding

This work was supported by the Russian Science Foundation [18-11-00284].

Conflict of Interest: none declared.

References

Alimova,I. *et al.* (2017) A machine learning approach to classification of drug reviews in Russian. In: *2017 Ivannikov ISPRAS Open Conference (ISPRAS)*. IEEE, Moscow, Russia, pp. 64–69.

Alvaro,N. *et al.* (2017) TwiMed: twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill.*, **3**, e24.

Dang,T.H. *et al.* (2018) D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*, **34**, 3539–3546.

Devlin,J. *et al.* (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Giorgi,J. and Bader,G. (2019) Towards reliable named entity recognition in the biomedical domain. *Bioinformatics.*, **36**, *280–286*.

Gonzalez-Hernandez,G. *et al.* (2018) Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *J. Am. Med. Inf. Assoc.*, **25**, 1274–1283.

Gu,J. *et al.* (2016) Chemical-induced disease relation extraction with various linguistic features. *Database*, **2016**, baw042.

Karimi,S. *et al.* (2015) CADEC: a corpus of adverse drug event annotations. *J. Biomed. Inf.*, **55**, 73–81.

Klie,J.-C. *et al.* (2018) The inception platform: machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Santa Fe, New Mexico, pp. 5–9.

Kuratov,Y. and Arkhipov,M. (2019) Adaptation of deep bidirectional multilingual transformers for Russian language. International Conference "Dialogue 2019" Moscow, Russia, May 29—June 1, 2019, pp. 1–7.

Lee,H.-C. *et al.* (2016) AuDis: an automatic CRF-enhanced disease normalization in biomedical text. *Database*, **2016**, baw091.

Lee,J. *et al.* (2020) BioBERT: pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.*, **36**, 1234–1240.

Miftahutdinov,Z. *et al.* (2017) Identifying disease-related expressions in reviews using conditional random fields. In *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog*, Vol. **1**, pp. 155–166.

Miftahutdinov,Z. *et al.* (2020). On biomedical named entity recognition: experiments in interlingual transfer for clinical and social media texts. In: Jose J.M. *et al.* (eds), *European Conference on Information Retrieval*, LNCS, Lisbon, Portugal, pp. 281–288.

Shelmanov, A.O. et al. (2015) *Information extraction from clinical texts in Russian. In Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, Conference Paper*, Moscow, Russia, vol. **1**, pp 560,

Smith,K. *et al.* (2018) Methods to compare adverse events in twitter to FAERS, drug information databases, and systematic reviews: proof of concept with adalimumab. *Drug Safety*, **41**, 1397–1410.

Tutubalina,E. and Nikolenko,S. (2017) Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. *J. Healthc. Eng.*, **2017**, 1–9.

Zolnoori,M. *et al.* (2019) A systematic approach for developing a corpus of patient reported adverse drug events: a case study for SSRI and SNRI medications. *J. Biomed. Inf.*, **90**, 103091.