An analysis of full-size Russian complexly NER labelled corpus of Internet user reviews on the drugs based on deep learning and language neuron nets

AG Sboev[a,b,*], SG Sboeva[c], IA Moloshnikov[a], AV Gryaznov[a], RB Rybka[a], AV Naumov[a], AA Selivanov[a], GV Rylkov[a] and VA Ilyin[a]

[a]NRC "Kurchatov institute", Moscow, Russia

[b]MEPhI National Research Nuclear University, Kashirskoye sh., 31, Moscow, 115409, Russia

[c]I.M. Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia

ARTICLE INFO

Keywords:
Pharmacovigilance
Annotated corpus
Adverse drug events
Social media
UMLS
MESHRUS
Information extraction
Semantic mapping
Machine learning
Neural Networks
Deep Learning

Abstract

We present the full-size Russian compound NER-labeled corpus of Internet user reviews, along with an evaluation of accuracy levels reached with this corpus by a set of advanced deep learning neural networks used for the extraction of pharmacologically meaningful entities from Russian texts. The corpus annotation includes mentions of the following entities: Medication (33005 mentions), Adverse Drug Reaction (1778), Disease (17403), and Note (4490). Two of them – Medication and Disease – comprise a set of attributes. In order to select the most effective neuron models for further adaptation to Russian language texts, numerical analysis has been performed on CADEC and N2C2 corpora. Selected neuronet models were adapted to Russian-language texts. This justifies the usage of our corpus to estimate the current accuracy baseline of the problem for Russian texts. Special multilabel model basing on a language model and the set of features is developed, appropriated for presented corpus labeling. The influence of the choice of different modifications of the models: word vector representations, types of language models pre-trained for Russian, text normalization styles, and other preliminary processing are analyzed. The sufficient size of our corpus allows to study the effects of particularities of corpus labeling and balancing entities in the corpus. As a result, the state of the art for the pharmacological entity extraction problem for Russian is established on a full-size labeled corpus, and is shown to be on par with the accuracy level for solving a similar task for other languages, which is 63.1% for ADR recognition by the F1-exact metric.

[*]Corresponding author

sag111@mail.ru (A. Sboev); sboevasanna@mail.ru (S. Sboeva); ivan-rus@yandex.ru (I. Moloshnikov); artem.official@mail.ru (A. Gryaznov); rybkarb@gmail.com (R. Rybka); sanya.naumov@gmail.com (A. Naumov); sanya.naumov@gmail.com (A. Selivanov); sanya.naumov@gmail.com (G. Rylkov); ilyin0048@gmail.com (V. Ilyin)

orcid(s): 0000-0002-6921-4133 (A. Sboev)

1. Introduction

Nowadays, a great amount of texts collected in the open Internet sources contains a vast variety of socially significant information. In particular, such information relates to healthcare in general, consumption sphere and evaluation of medicines by the population. Due to time limitations, clinical researches may not reveal the potential adverse effects of a medicine before entering the pharmaceutical market. This is a very serious problem in healthcare. Therefore, after a pharmaceutical product comes to the market, pharmacovigilance (PV) is of great importance. Patient opinions 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 efficacy 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 uses of various drugs to treat different diseases, including “off-label” applications. This information would be very useful for a PV database where risks and advantages of drugs would be registered for the purpose of safety monitoring, as well as the possibility to form hypotheses of using existing drugs for treating other diseases. This leads to an increasing need for the analysis of information from electronic sources to assess the quality of medical care and drug provision. An active control on the base of social networks is implemented by a number of countries.

To automatically analyze such amount of information, special methods have to be developed. However, the quality of these methods directly depends on tagged corpora to train them. In particular, the United States Food and Drug Administration is creating a base, using medical forums, where consumers discuss the experience of using drugs. That base would be a valuable resource for the analysis and development of the texts in machine learning. However, expert assessment of such texts is too laborious, while extracting medical information using conventional processing methods is difficult due to the use of the informal vocabulary and the presence of reasoning. In this regard, one of the main tasks is the development of machine learning methods for extracting useful information from social media. It is an area of increasing interest, and even is becoming mandatory on the territory of the Eurasian Union: according to the decision of the Council of the Eurasian Economic Commission No. 87 of November 3, 2016 “On Approving the Rules of Good Practice for PV of the Eurasian Economic Union”, registration certificate holders are obliged to monitor the Internet and digital health records regularly for potential reports on suspected undesirable reactions. Unfortunately, there still have been no annotated corpora for PV in Russia. In this paper, we present the first Russian corpus of such type with complex annotation, for which we propose the name Russian Drug Reviews by SagTeam initiative project (RDRS) [1]. In addition, we present a deep learning neural network complex to extract pharmacologically meaningful entities from a Russian text.

The materials used to collect the corpus are outlined in Section 3.1, the technique of its annotation is described in Section 3.2. The developed machine learning complex is presented in Section 3.4. The conducted numerical experiments are discussed in Section 3.6.

2. Related works

In the world science, research concerning the abovementioned problems is conducted intensively, resulting in a great diversity of annotated corpora. These corpora can be divided into two groups: firstly, the ones of texts written by medics (clinical reports with annotations), and secondly, those of texts written by nonspecialists, namely, by the Internet customers who used the drugs. The distinctive features of any corpus are the number of entities, the number of annotation types, and approaches to entity normalization. The diversity of these features makes it difficult to compare the accuracy of entity recognition on the base of different corpora.

Clinical corpora and corpora of Internet user texts, on the one hand, have a certain similarity, but on the other hand, essential differences. Both are partially based on the similar annotation schemes, normalization procedures, and can be analyzed by similar algorithms. However, the former − clinical corpora − have more types of annotated entities, mostly related to disease and treatment indication rather than pharmacology entities. Types of included entities are usually stricter in clinical texts than in Internet user texts.

The variability of the natural language constructions in the speech of Internet users complicates the analysis of corpora based on Internet texts. In this section, we strive to clarify how these reasons influence the accuracy of entity recognition, taking into account the differences of the task formulation in papers about analysis of clinical corpora or corpora of Internet user texts. We consider the corpora closest to the one presented in this article. Table 1 provides a summary of the relevant corpora.

2.1. Corpora of clinical reports

CLEF corpus (CLinical E-Science Framework) [42]. The basis of this corpus is a dataset of 565 000 semantically annotated documents on 20 234 deceased patients from Royal Marsden Hospital. The documents are of three types: clinical narratives, histopathology reports, and imaging reports. The annotators marked text fragments (spans) with a type:

[1] Russian Drug Reviews by SagTeam initiative project (RDRS) - https://sagteam.ru/en/

Table 1
Summary of a few relevant corpora of annotated clinical texts.

Corpus name	Origin	Size	Entities
CLEF corpus	English reports from 565 000 records of 20 234 deceased patients from the Royal Marsden Hospital	~200 documents, 50 of each of the following types: clinical narratives, histopathology reports, and imaging reports	Condition; Drug or device; Intervention; Investigation; Laterality; Locus; Negation; Result; Sub-location
ShARe Clef eHealth 2013	English clinical reports from US intensive care (version 2.5 of the MIMIC II database). The corpus consists of discharge summaries and electrocardiogram, echocardiogram, and radiology reports	298 documents	Negation Indicator; Subject Class; Uncertainty Indicator; Course Class; Severity Class; Conditional Class; Generic Class; Body Location; DocTime Class; Temporal Expression
ShARe Clef eHealth 2014	English clinical reports (same as CLEF 2013)	433 documents	Negation Indicator; Subject Class; Uncertainty Indicator; Course Class; Severity Class; Conditional Class; Generic Class; Body Location; DocTime Class; Temporal Expression
SCCH	Russian clinical records for 60 patients from Federal State Autonomous institution "National Medical Research Center of Children Health" (NMRCCH)	112 documents	Disease; Symptom; Drug; Treatment; Body location; Severity; Course
ADE corpus	English medical case reports of MEDLINE database, which is a part of PubMed. The documents with medication adverse effects mentioned are selected.	2 972 documents	Drug; Adverse effect; Dosage
MADE1.0	English medical records of 21 randomly selected cancer patients monitored in the University of Massachusetts Memorial Hospital	1 089 notes	Adverse effect; Indication; Other Sign, Symptom, or Disease; Severity; Drugname; Dosage; Duration; Frequency; Route
I2b2 – NLP Data Set #3	English discharge summaries of patients from the non-profit hospital and physicians network "Partners Healthcare"	1 243 documents	Medications; Dosages; Modes; Frequencies; Durations; Reasons; List/narrative
IxaMed corpus	Randomly selected reports of Spanish medical consultations in Galdakao-Usansolo Hospital	75 verified reports	Drug; Procedure; Disease

drug, locus, and so on. The Clinical Narratives [43] part of the corpus contains 77 documents with mentions of the following entity types marked: Condition (739 mentions), Intervention (298), Investigation (325), drug or device (272), locus (490). In addition, the annotators marked words that modify spans (such as negation), and marked relations between spans. Two or more spans may refer to the same entity, in which case they are coreferent. Every document is marked up by two independent annotators, and the third one makes the final consensus annotation.

ShARe CLEF eHealth 2013 [40] and 2014 [32] are the corpora collected for the competitions on the medical texts information extraction task: CLEF eHealth 2013 Task 1, and CLEF eHealth 2014 Task 2 respectively. The 2013 corpus contains about 300 documents of 4 clinical report types: discharge summaries, radiology, electrocardiograms, and echocardiograms. The corpus of year 2014 is an extension, larger by 133 documents. The main dataset for both corpora is the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) [18]. The MIMIC comprises impersonal health data associated with about 40 000 critical care patients and includes demographics, vital signs, laboratory tests, medications, etc.

SCCH (Corpus of Scientific Center of Children Health) [49] is an annotated corpus of clinical texts in Russian [2] . The corpus includes 112 medical records of more than 60 patients from the Scientific Center of Children Health with allergic and pulmonary disorders and diseases. It comprises discharge summaries, radiology, echocardiography, ultrasound diagnostics reports, and recommendations. All documents are depersonalized: names have been deleted and dates distorted. The markup scheme is partially similar to the scheme presented in ShARe CLEF eHealth 2014 Task 2 and in-

[2] SCCH corpus website: http://nlp.isa.ru/datasets/clinical

cludes the following entities: Disease; Symptom; Drug; Treatment; Body location; Severity; Course. Currently, the documents contain the total of 45 000 tokens (words, punctuation symbols, and so on), among which more than 7 600 are annotated with entities and more than 4 000 are annotated with attributes and relations.

ADE (adverse drug effect) corpus [13] consists of 2 972 documents which have been randomly selected from nearly 30 000 PubMed documents and annotated manually by three annotators. The corpus contains three types of entities: Drug, Adverse effect, and Dosage. Annotators labeled relations in sentences between drugs and side effects; drugs and dosages; and all texts in corpus that do not contain any drug-related adverse effects. The goal was to extract adverse effects of drugs mentioned within the context of sentences. Mentions of drugs, disorders or dosages that did not fit into a relation were not annotated. With the purpose of preparing a larger dataset for a supervised classifier, the ADE corpus was expanded with machine-annotated drugs, conditions, and conditions that did not fall into adverse effect relations but were still within the same sentence. This corpus is called ADE-EXT [14]. It includes 2 269 new drugs, 3 437 new conditions and 5 968 false relations (co-occurring drug-condition pairs that were not previously annotated by humans were considered false).

MADE1.0 (Medication and Adverse Drug Events from Electronic Health Records) corpus [17] contains a cohort of medications and ADE information annotated by experts. It includes 1 089 depersonalized electronic health notes from 21 randomly selected cancer patients at the University of Massachusetts Memorial Hospital. The corpus provides a set of common evaluation tasks to assess the state of the art for natural language processing (NLP) systems applied to electronic health records supporting drug safety surveillance and PV. The MADE 1.0 was used in three shared NLP tasks: The named entity recognition (NER) task for medications and their attributes (dosage, route, duration, and frequency of taking), indications, ADEs, and severity. The relation identification (RI) task is the identification of relations between the named entities (medication-indication, medication-ADE, and attribute relations). The third shared task (NER-RI) evaluates NLP models that perform the NER and RI tasks jointly. A particularity of this corpus is that the annotators mark an ADE mention only in a direct linguistic cue that links an adverse effect to a drug name. The high quality of the corpus markup is provided thanks to selecting entities by pattern matching, which allows not to use additional normalization.

I2b2–NLP Data Set #3 [57] has been developing since 2006. The corpus contains data of 1 243 depersonalized summaries from the network of non-profit hospital and physicians "Partners Healthcare", which includes Brigham and Women's Hospital (BWH), and Massachusetts General Hospital (MGH). The annotation process was preceded with a golden standard markup, containing annotations of 17 texts. After that, an annotation guide was created, and different expert groups have annotated another 547 texts. Finally, the research group processed the annotation results and collected a dataset containing 251 annotated texts. The list of annotated entities includes: Medications, Dosages, Modes, Frequencies, Durations, Reasons. The corpus contains both structured and narrative components of clinical records. Thus, the first type of discharge summary is a formalised list structure, and the second is a narrative text. As shown in that work [57], entities are much better extracted from structural information than from a narrative text. As a result, it gives one a possibility to reach higher accuracy without normalization.

IxaMed corpus [36] is composed of electronic health records written in Spanish. The corpus consists of 142 154 discharge reports from the outpatient consultations in the Galdakao-Usansolo Hospital. The documents were created between 2008 and 2012 by about 400 doctors from different services. Experts made modifications if it is necessary and tagged the adverse drug reaction (ADR) events. All documents in the corpus were manually depersonalized by changing names and dates. All medical abbreviations were identified with the help of the dictionary by Yetano and Alberola [25], while drug brand names were looked up in the BOTPLUS5[3] database. Its gold standard is manually annotated by experts in pharmacology and pharmacovigilance. The annotation was performed in several stages. At first, the developers created an annotation guide and chose a preliminary marking system based on dictionaries and rules. Two independent annotators marked up 50 documents using the guide, matched the differences and refined the guide. Then, another 25 texts were annotated and joined with the first set to form a golden standard of 75 documents.

2.2. Normalization approaches applied to clinical corpora

For normalizing the entities in clinical corpora, different approaches for mapping to thesauruses were used. For instance, in ShARe CLEF eHealth 2013 [40] and 2014 [32], disease entities were mapped to a Concept Unique Identifier (CUI) according to the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) terminology [48], which belongs to one of the Unified Medical Language System (UMLS) [6] semantic types. In the CLEF corpus [42], the normalization of entities is based on the types from the

[3] Drug Database BOTPLUS5 - https://botplusweb.portalfarma.com/

Table 2
Summary of the existing clinical text corpora.

Corpus	Normalization					Num. Entities	F1(exact)	F1(partial)
	SNOMED CT	MedRA	UMLS	SRD	ATC			
ADE corpus [26]	−	+	−	−	+	10 666	0.846	−
CLEF corpus [43]	−	−	+	−	−	2 124	0.71	−
ShARe Clef eHealth 2013 [53]	+	−	+	−	−	11 151	0.75	0.873
ShARe Clef eHealth 2014 [19]	+	−	−	−	−	19 557	0.676	0.72
SCCH [49]	−	−	+	+	−	7 600	−	0.862
MADE1.0 [61]	−	−	−	−	−	79 003	0.841	−
I2b2 – NLP Data Set #3 [37]	−	−	−	−	−	22 000	0.856	0.849
IXAMed corpus [38]	+	−	−	−	−	6 158	0.703	−

Table 3
Summary of the existing Internet text corpora.

Corpus	Normalization					Num. Entities	F1(exact)	F1(partial)
	SNOMED CT	MedRA	UMLS	SIDER	AMT			
CADEC [8]	+	+	−	−	+	8118	0.806	−
PsyTar [63]	+	−	+	−	−	7414	0.49	−
TwiMed [12]	+	+	−	+	−	1200	−	0.648
Twitter annotated corpus [58]	−	−	+	−	−	1482 tweets (2130 tweets original dataset)	−	0.611

UMLS semantic network. The procedure of term normalization in SCCH [49] is based on two thesauruses: UMLS Metathesaurus for disease, symptom, and body site identification; a thesaurus based on the State Register of Drugs (SRD) [44] for drug identification. The only Russian thesaurus present in UMLS is MeSHRUS (Medical Subject Headings) [27], but it does not provide all nomenclature of drugs used in Russia. In the IxaMed corpus [36], the FreeLingMed system [35] is able to carry out medical named entity recognition, linking all the terms in SNOMED CT with their corresponding semantic tags (substances, disorders, procedures, findings). Another system, ProMiner [15], was used in the ADE corpus [13] to map names of drugs and adverse effects to standard ontologies. In this case, the drug names were mapped to the Anatomical Therapeutic Chemical (ATC) classification system [31] using the DrugBank [60] dictionary. The ATC hierarchically classifies several drugs according to their pharmacotherapeutic properties. The names of adverse effects were mapped to the Medical Dictionary for Regulatory Activities (MedDRA) [4] classification system. In the MADE 1.0 [17] and I2b2–NLP #3 [57], corpora there is no normalization.

Table 2 summarizes the information about the normalization types in the clinical corpora and presents their overall entity recognition accuracy (averaged over all entity types) measured by the F1 exact/partial metrics explained in Section 3.5.

2.3. Corpora of internet user texts

CADEC (corpus of adverse drug event annotations) [20] is a corpus of medical posts taken from the AskaPatient [4] forum and annotated by medical students and computer scientists. It collects ratings and reviews of medications from their consumers and contains consumer posts on 13 different drugs. There are 1253 posts with 7398 sentences. The following entities were annotated: Drug, ADR, Symptom, Disease, Findings. The annotation procedure involved 4 medical students and 2 computer scientists. In order to coordinate the markup, all annotators jointly marked up several texts, and after that the texts were distributed among them. All the annotated texts were checked by three corpus authors for obvious mistakes, e.g. missing letters, misprints, etc.

TwiMed corpus (Twitter and PubMed comparative corpus of drugs, diseases, symptoms, and their relations) [1] contains 1000 tweets and 1000 sentences from Pubmed [5] for 30 drugs. It was annotated for 3 144 entities, 2 749 relations, and 5 003 attributes. The re-

[4] Ask a Patient: Medicine Ratings and Health Care Opinions - http://www.askapatient.com/

[5] National Center for Biotechnology Information webcite - http://www.ncbi.nlm.nih.gov/pubmed/

Table 4
The F1 accuracy score of recognizing ADR entities in the existing corpora

Paper	Text type	Corpus	Number of ADR entities	F1 (exact)
Xiang Dai et. all [8]	Social media	Cadec	6318	0.687
Gupta et. all [12]		TwiMed (Twitter part)	~1200	0.648
W. Wang [58]		Twitter annotated corpus	NA	0.611
Ramamoorthy et. all [26]	Clinical documents	ADE corpus	5776	0.868
Wunnava et. all [61]		MADE1.0	1940	0.609

sulting corpus was composed of agreed annotations approved by two pharmaceutical experts. The entities marked were Drug, Symptom, and Disease.

Twitter annotated corpus [46] consists of randomly selected tweets containing drug name mentions: generic and brand names of the drugs. The annotator group comprised pharmaceutical and computer experts. Two types of annotations are currently available: Binary and Span. The binary annotated part [45] consists of 10 822 tweets annotated by the presence or absence of ADRs. Out of these, 1 239 (11.4%) tweets contain ADR mentions and 9583 (88.6%) do not. The span annotated part [46] consists of 2 131 tweets (which include 1 239 tweets containing ADR mention from the binary annotated part). The semantic types marked are: ADR, beneficial effect, indication, other (medical signs or symptoms).

PsyTAR dataset [62] contains 891 reviews on four drugs, collected randomly from an online healthcare forum [6] . They were split into 6 009 sentences. To prepare the data for annotation, regular expression rules were formulated to remove any personal information such as emails, phone numbers, and URLs from the reviews. The annotator group included pharmaceutical students and experts. They marked the following set of entities: ADR, Withdrawal Symptoms (WD), Sign, Symptom, Illness (SSI), Drug Indications (DL) and other.

2.4. Normalization approaches applied to Social corpora

The normalization task of internet user texts is more difficult because of informal text style and more natural vocabulary. Still, as in the case of clinical texts, thesauruses are used. In particular, annotated entities in CADEC were mapped to controlled vocabularies: SNOMED CT, The Australian Medicines Terminology (AMT) [33], 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 [23] was used, which contains information on marketed medicines extracted from public documents, while for Symptom and Disease entities the MedDRA ontology was used. In addition, the SNOMED CT concept terminology was used, which belongs to the Disorder semantic group. In the Twitter dataset [46], when annotating ADR mentions, they were normalized to their UMLS identifiers. Finally, in PsyTAR corpus, ADRs, WDs, SSIs and DIs entities were matched to UMLS and SNOMED CT concepts.

As in the part concerning clinical corpora, in Table 3 we summarize a set of papers addressing named entity recognition and type classification tasks on the presented social corpora. The average entity recognition scores F_1^{exact} and F_1^{partial} (described in Section 3.5) are presented for different normalization types.

2.5. Comparison of entity identification accuracy in clinical and internet texts

The overall entity identification accuracy (see Tables 2 and 3) is higher in clinical texts by about 12% (by the F_1^{exact} metric) and 19% by F_1^{partial}. This may be explained by the fact that clinical texts are more formal and strictly-structured than Internet texts. Table 4 confirms this conclusion for the case of identifying ADR mentions solely. There, the accuracy of ADR identification in clinical texts is about 9% higher by F_1^{exact}.

The results mentioned above are obtained using different approaches and features. Several works ($F_1^{\text{exact}} = 0.676$ [19] on the ShARe CLEF eHealth 2014 corpus, $F_1^{\text{partial}} = 0.862$ [49] on SCCH, $F_1^{\text{exact}} = 0.49$ [63] on PsyTar) are based on rule-based approaches which use the UMLS dictionary, morphological, syntactical, and negation dependency features. Other works use a combination of rule-based approach with machine learning models like CRF or SVM ($F_1^{\text{exact}} = 0.856$, $F_1^{\text{partial}} = 0.849$ [37] on the I2b2 corpus) along with dictionary (SNOMED CT) and traditionally morphology (prefix, suffix) feature set. Indeed, in the highly formalised clinical texts, even a simple classification model of SVM shows good results ($F_1^{\text{exact}} = 0.71$ [53] on CLEF, $F_1^{\text{exact}} = 0.75$, $F_1^{\text{partial}} = 0.873$ [43] on ShARe CLEF eHealth 2013) on base of such features as bag of words, part-of-speech (PoS) and semantic categories of

[6] Ask a Patient: Medicine Ratings and Health Care Opinions - http://www.askapatient.com/

Table 5
Specifications of the corpus.

Type	Origin	Size	Language	Named Entities
Social	Reviews from users of the Otzovik medical forum	1660 posts	Russian	ADR; Medication (Drugname; DrugBrand; Drugform; Drugclass; Domestic; Foreign; Frequency; Dosage; Duration; Route; SourceInfodrug); Disease (Diseasename; Indication; ADE-Neg; BNE-Pos; NegatedADE; Notes; Worse); Note

Table 6
A sample post for "Глицин" (Glycine) from otzovik.com. Original text is quoted, and followed by English translation in parentheses.

Overall impression	"Помог чересчур!" (Too much help!)
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.)

words based on UMLS.

However, in unformalised Internet texts, the conventional machine learning algorithms show inferior accuracy. For instance, CRF achieves $F_1^{\text{exact}} = 0.703$ [38] on the IXAMed corpus, but just $F_1^{\text{partial}} = 0.611$ [58] on the Twitter corpus. Both works use a wide range of features including word forms, linguistic features, PoS, semantic tags, word embedding, and SNOMED CT or COSTART, SIDER dictionaries respectively. An up-to-date level of accuracy on Internet texts is usually reached by a modern deep learning approach of bi-directional LSTM (bi-LSTM) ($F_1^{\text{partial}} = 0.648$ [12] on TwiMed) based on PoS features along with pretrained word embedding models. In some works, the bi-LSTM architecture additionally comprises a character-level representation CNN layer [26] ($F_1^{\text{exact}} = 0.846$ on the ADE corpus), a CRF layer [8] ($F_1^{\text{exact}} = 0.806$ on CADEC) or a combination of CNN with recurrent neural network layers [61] ($F_1^{\text{exact}} = 0.841$ on MADE1.0). Therefore, in this work, in order to estimate the quality of our corpus we implement the deep learning bi-LSTM architecture along with an advanced neuronet model with use of up-to-date pretrained Russian language models.

3. Materials and methods

3.1. Corpus material

In this section, we report on the design of our corpus. Its basis were 1 660 reviews from a medical forum called OTZOVIK[7] , which is dedicated to consumer reviews on medications. On that website there is a partition 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 5-point scale for such parameters as quality, packing, safety, availability. A sample post for "Глицин" (Glycine) is shown in Table 6.

We used information only from the simplified form, since the users had rarely filled extended forms in their reviews. We considered only the fields Heading, General impression and Comment. Furthermore, some of the reviews are written in common language and do not follow formal grammar and punctuation rules. The consumers described not only their personal experience, but sometimes opinions of their family members, friends or others. The main specifications of our corpus are shown in Table 5.

3.2. Corpus Annotation

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

[7] OTZOVIK - Internet forum from which user reviews were taken - http://otzovik.com

3.2.1. Annotation process

Creating a reliably annotated corpus depends on experts a lot. The group of 4 annotators made annotation using a manual developed jointly by machine learning experts and pharmacists. Two annotators were certified 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 by means of journaling. 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 entities description and examples.
2. Upon testing on a set of 300 reviews, the guide was corrected, addressing complex cases. During that, iterative annotation was performed, from 1 to 5 iterations for a text, while tracking for each text and each iteration the annotator questions, controller comments, and correction status.
3. The resulting guide was used for annotating the remaining reviews. Two annotators marked up each review, and then a pharmacist checked the result. When complex cases were found, they were analyzed separately by the whole group of experts.
4. The obtained markup was automatically checked for any inaccuracies, such as incomplete fragments of words selected as mentions, terms marked differently in different reviews, etc. Texts with such inaccuracies were rechecked.

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 under the guidelines below.

3.2.2. Guidelines applied in the course of annotation

The annotation goal was to get a corpus of reviews in which named entities reflecting pharmacotherapeutic treatment are labelled, and annotate medication characteristic semantically. With this in mind, the objects of annotation were attributes of drugs, diseases (including their symptoms), and undesirable reactions to those drugs. The annotators were to label mentions of these three entities with their attributes defined below.

Medication. This entity includes everything related to the mentions of drugs and drugs manufacturers. Selecting a mention of such entity, an annotator had to specify an attribute out of those specified in Table 7, thereby annotating it, for instance, as a mention of the attribute “DrugName” of the entity “Medication” . In addition, the attributes “DrugBrand” and “MedFrom” were annotated with the help of lookup in an external source [44].

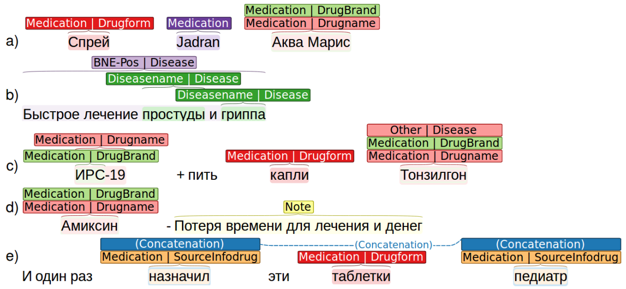

Figure 1: Examples of markup. a) “Spray Jadran Aqua Maris”, b) “Rapid treatment of cold and flu”, c) “IRS-19 + drink drops of Tonsilgon” d) “Amixin – waste of time and money for treatment”, e) “And once were these pills prescribed by my pediatrician”

Disease. This entity 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 state after taking the drug. Attributes of this entity are specified in Table 8.

ADR. This entity is associated with adverse drug reactions in the text. For example, one post said: «После недели приема Кортексина у ребенка начались судороги» (After a week of taking Cortexin, the child began to cramp). In this sentence, the word “судороги” (“cramp”) is labeled as an ADR entity.

Note. We use this entity when the author makes recommendations, tips, and so on, but does not explicitly state whether the drug helps or not. These include phrases like “I do not advise”. For instance, the phrase «Нет поддержки для иммунной системы» (No support for the immune system) is annotated as a Note.

The typical situations that had to be handled during the annotation are the following:

1. A simple markup, when a mention consists of 1 or more words and it 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 mentions separated by words that are not part of 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 Fig. 1 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

Table 7
Attributes belonging to the Medication entity

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.
DrugBrand	A drug name is also marked as DrugBrand if it is a registered trademark. For example, in the sentence «Противовирусный и иммунотропный препарат Экофарм “Протефлазид”» (The Ecopharm “Proteflazid” antiviral and immunotropic drug), the word “Протефлазид” (Proteflazid) is marked as DrugBrand.
Drugform	Dosage form of the drug (ointment, tablets, drops, etc.). For example, 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), the word “таблетки” (pills) is marked as DrugForm.
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. This attribute has two values: Domestic and Foreign. For example, in the sentence «Седативный препарат Материа медика “Тенотен”» (The Materia Medica “Tenoten” sedative) the word combination “Материа медика” (Materia Medica) is marked as MedMaker/Domestic.
MedFrom	This is an attribute of a Medication entity that takes one of the two values – Domestic and Foreign, characterizing the manufacturer of the drug. For example, in the sentence «Седативные таблетки Фармстандарт “Афобазол”» (The Pharmstandard “Afobazol” sedative pills) the drug name “Афобазол” (Afobazol) has its MedFrom attribute equal to Domestic.
Frequency	The drug usage frequency. For example, 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), the phrase “2 раза в день” (two times a day) is marked as Frequency.
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 МЕ” (150000 IU) is marked as Dosage.
Duration	This entity specifies the duration of use. For example, in the sentence «Время использования: 6 лет» (Time of use: 6 years), “6 лет” (6 years) is marked as Duration.
Route	Application method (how to use the drug). For example, in the sentence «удобно то, что можно готовить раствор небольшими порциями» (it is convenient that one can prepare the solution in small portions), the mention “можно готовить раствор небольшими порциями” (can prepare a solution in small portions) is marked as a Route.
SourceInfodrug	The source of information about the drug. For example, 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), the word combination “посоветовали в аптеке” (recommended to me at a pharmacy) is marked as SourceInfoDrug.

simultaneously. For example, in the sentence “Rapid treatment of cold and flu” (see Fig. 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 a mentions of different attributes or entities at the same time (for example, “drugname” and “drugbrand”), it should be annotated with all of them: see, for instance, entity “Aqua Maris” in sentence “Spray Jadran Aqua Maris” (Fig. 1, example (a)).

4. Another complex situation is when an analogue (or, in some cases, several analogues) of the drugs are mentioned in a text, for example, when a customer wrote about a drug and then described an alternative that helped them. In this case, the “Other” attribute is used (example (c)).

Moreover, there often were author subjective arguments instead of explicit reports on the outcomes. We labeled that as a mention of entity “Note”. For example, “strange meds”, “not impressed”, “it is not clear whether it worked or not”, “ambiguous effect” (example (d) in Fig. 1).

3.3. Normalization

After annotation, in order to resolve possible ambiguity in terms we performed normalization by matching the labeled mentions to the 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) [34] is an international classification system for diseases which includes 22 classes of diagnoses, each consisting of up to 100 categories. The ICD-10 makes it possible to re-

Table 8
Attributes belonging to the Disease entity

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. Also, 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.

duce verbal diagnoses of diseases and health problems to unified codes.

- The Anatomical Therapeutic Chemical (ATC) [31] is an international medication classification system containing 14 anatomical main groups and 4 levels of subgroups. The ICD-10 and the 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.

- State Register of Medicinal Products (SRD)("Государственный реестр лекарственных средств (ГРЛС)" [44] in Russian) is a register of detailed information about the medications certified in the Russian Federation. It includes possible manufacturers, dosages, dosage forms, ATC codes, indications, and so on.

- MeSH Russian (MESHRUS) [27] is a Russian version of the Medical Subject Headings (MESH) database [8]. 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 doesn't contain concept definitions.

Among the international systems of standardization of concepts, the most complete and large metathezaurus 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 the UMLS has an identification code CUI, using which one can get information about the concept from all the databases. However, within UMLS it is only the MESHRUS database that contains Russian language

[8] Home page of the MeSH database site: https://www.nlm.nih.gov/mesh/meshhome.html

and can be used to associate words from our texts with CUI codes.

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

Automatic mentions grouping is based on calculating the similarity between two mentions by the Ratcliff/Obershelp algorithm [41], 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 $\boldsymbol{G}$ 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 $\boldsymbol{G}$ set 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.

After that, the group names for attributes "Diseasename", "Drugname" and "Drugclass" are manually matched with ICD-10 and ATC terms to assign term codes from the classifiers. As a result, 141 unique ICD-10 codes were matched against the 1 333 mentions of attribute "Diseasename"; 171 unique ATC codes matched the 2 360 mentions of attribute "Drugname"; and 26 unique ATC codes corresponded to 1 092 mentions of "Drugclass". Some drug classes that were mentioned in corpus (such as homeopathy) did not have a corresponding ATC code, and were aggregated according to their anatomical and therapeutic classification in the SRD.

Normalization based on MESHRUS concepts. 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. We perform two approaches to automatically find and map concepts from MESHRUS to words from corpus.

The following preprocessing algorithm is used for mapping words from the corpus to concepts from the dictionary: words are lemmatized, put into a single register and filtered by length, frequency and parts of speech.

The first approach is to map the filtered words $\boldsymbol{W} = \{w_i\}_{i=0}^{N}$ from the corpus to MESHRUS concepts $\{k_j\}$. As a criterion for comparing words and concepts, we used the cosine similarity between their vector representations obtained using the FastText [2] model (see Section 3.4.1): a word w_i is assigned the CUI code v_j (see Fig. 7) whose corresponding concept k_j has the highest similarity measure $\cos\left(\mathrm{FastText}(w_i), \mathrm{FastText}(k_j)\right)$. If this similarity measure is lower than the empirical threshold $T = 0.55$, no CUI code is assigned to w_i.

The second approach is based on the mapping of syntactically and lexically related phrases extracted at the sentence level. Prepositions, particles and punctuation are not taken.

For each word $w_i \subset \boldsymbol{W}$, its adjacent words $[w_{i-1}, w_{i+1}]$ are selected. Together with the word 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 from the MESHRUS dictionary: $c_{j_l} = [c_{j-1}, c_j, c_{j+1}]$, and $c_{j_s} = [c_j, c_{j_{\text{parent}}}]$.

Further, for each word $w_i \subset \boldsymbol{W}$ and word $c_j \subset concept_k \subset$ MESHRUS, by analogy with the literature [49], the following metrics are calculated:

1. $\mathsf{lexical\_involvement}(w_i, c_j) = F_1\left(\frac{|w_{i_l} \cap c_{j_l}|}{|w_{i_l}|}, \frac{|w_{i_l} \cap c_{j_l}|}{|c_{j_l}|}\right)$
2. $\mathsf{cohesiveness}(w_i, c_j) = F_1\left(\frac{|w_{i_s} \cap c_{j_s}|}{|w_{i_s}|}, \frac{|w_{i_s} \cap c_{j_s}|}{|c_{j_s}|}\right)$
3. $\mathsf{centrality}$ 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 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.

The normalization results are shown in Table 9. The "MESHRUS – total" column contains the number of words from $\boldsymbol{W}$ that were annotated as parts of mentions of a particular attribute, the "MESHRUS – unique" column shows the number of unique codes related to the mentions.

3.3.1. Statistics of the collected corpus

Detailed information about the collected corpus is presented in Table 9 including:

1. The number of mentions for every attribute ("Mentions – Annotated" column in the table).
2. The number of unique mentions of the attribute after manual standardization procedure described in Section 3.3 ("Mentions – Standardized").
3. The number of words belonging to mentions of the attribute ("Mentions – Number words in the mentions").

Table 9
General information about the collected corpus.

Entity type	Mentions					MESHRUS		MESHRUS-2	
	Annotated	Standardized	Num. words in the mentions	Total entries	Reviews coverage	total	unique	total	unique
ADR	844	163	4 159	764	448	455	112	0	0
Medication	17 779	1 710	52 782	11 017	1 659	3 612	458	1 234	131
Drugname	4 730	384	7 497	2 360	1 654	852	166	47	14
DrugBrand	2 564	234	2 936	1 254	1 036	23	19	2	1
Drugform	3 236	30	6 496	1 501	1 243	651	40	368	48
Drugclass	1 735	36	4 069	1 204	974	682	47	624	35
MedMaker	998	210	2 838	905	851	553	60	178	25
Frequency	365	121	2 805	350	303	6	5	0	0
Dosage	506	93	2 632	407	387	20	9	1	1
Duration	897	85	4 012	772	703	9	9	0	0
Route	1 511	503	6 646	1 351	834	288	65	14	7
SourceInfodrug	1 225	8	4 733	913	861	528	38	0	0
Disease	9 285	2913	39 548	7 505	1 603	3 272	791	82	16
Diseasename	2 235	141	5 059	1 333	921	499	55	76	12
Indication	2 334	708	7 230	2 014	971	757	112	3	1
BNE-Pos	2 967	1 301	15 932	2 293	1 021	794	185	1	1
ADE-Neg	115	93	906	99	68	58	43	0	0
NegatedADE	1 535	603	9 617	1 183	641	260	97	2	2
Worse	83	68	696	76	51	29	21	0	0
Note	2 319	1 546	19 781	1 777	1 004	875	278	2	2

4. We introduce the concept of entries – the number of reviews that contain a standardized mention – so as to be able to compare drugs and diseases by their coverage in the reviews, not taking into account multiple repetitions of the same standardized mention in a review. The “Mentions – Total entries” column shows the total number of entries over all the standardized mentions of the corresponding attribute (this number can exceed the total number of reviews because some reviews contain several different standardized mentions).
5. The number of reviews containing any mentions of the corresponding attribute (“Mentions – Reviews coverage”).
6. The last four columns contain the results of automated normalization with the help of the MESHRUS dictionary using the two normalization approaches that were explained in the previous section, here called MESHRUS and MESHRUS-2. The numbers show how many times unique words with assigned CUI codes were labeled as parts of any mentions of the attribute under consideration and how many unique CUI indices were assigned to these words.

The corpus contains consumer posts on 384 drugs, mentioned 2360 times. These drugs relate to 36 drug classes according to the classification from the State Register of Drugs [44]. The top 20% of the drugs (by the entry counts of their corresponding Drugname mentions) include 77 different products with the total of 1 699 entries (which is 71.99% of all entries). Among them, 29 drugs were reviewed in more then 20 documents, the sum of their entries count is 1114.

The most popular drug classes mentioned in corpus are antiviral (74 drugs) and sedative (39 drugs). The sums of entries of these drugs have parts from all drug name attribute entries equal to 48.52% and 17.07% correspondingly. The proportions of entry counts of the most popular drugs to the total number of entries of antiviral drugs are: “Виферон” (Viferon) (6.9%), “Ингаверин” (Ingavirin) (5.41%) and “Ацикловир” (Acyclovir) (4.54%). For the sedative drugs, these are: “Глицин” (Glycine) (16.38%), “Валериана” (Valeriana) (14.39) “Афобазол” (Afobazol) (8.93%).

The proportions of domestic drugs and foreign drugs to the total number of drug entries are 38.8% and 61.2%, respectively. The foreign drugs with the highest entry percentages are: “Афлубин” (Aflubin) (2.37%), “Иммунал” (Immunal) (1.22%), “Амизон” (Amison) (1.18%), and “Антигриппин” (Antigrippin) (1.18%). The domestic ones are “Виферон” (Viferon) (3.34%), “Анаферон” (Anaferon) (3.17%), “Глицин” (Glycine) (2.79%), and “Ингавирин” (Ingavirin) (2.62%).

Regarding disease names, the most frequent ones are “острые респираторные инфекции верхних и нижних дыхательных путей” (acute respiratory infections of the upper and lower respiratory tract) (554 entries); “грипп и пневмония” (influenza and pneumonia) (262 entries); “вирусные инфекции, характеризующиеся поражениями кожи

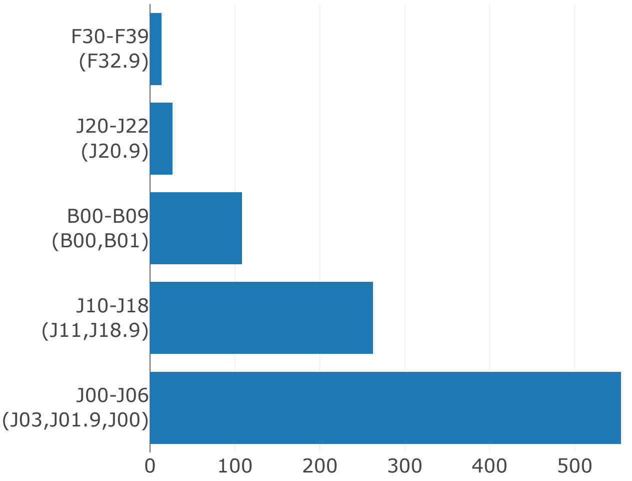

Figure 2: Top 5 disease categories from the ICD-10 by the number of entries in our corpus. F32.9: Unspecified depressive episode; J20.9: Unspecified acute bronchitis; B00-B09: Viral infections characterized by lesions of the skin and mucous membranes; J10-J18: Influenza and pneumonia; J00-J06: ARVI.

и слизистых оболочек" (viral infections characterized by lesions of the skin and mucous membranes) (108); "другие вирусные болезни" (other viral diseases) (38) and others. The top 5 disease categories from the ICD-10 by the entries count are presented in Fig. 2.

Analysing the consumers' motivation to acquire and use drugs ("sourceInfoDrug" attribute) showed that the major part of the drugs, 793 (86.68%) entries, were used on professional recommendations: medical or pharmaceutical specialists. 120 (13.31%) entries of drug usage referred to advice of non-professional sources: relatives, friends, advertisement and so on.

The distribution heatmap of entry percentages for different sources for the 20 most popular drugs is presented in Fig. 3. It could be seen that most recommendations are coming from professionals. For example "Изопринозин" (Isoprinosine) (used in 69.23% cases by medical recomendations), "Афлубин" (Aflubun) (48.21%), "Анаферон для детей" (Anaferon for children) (47.30%) and others. However, for such drugs as "Иммунал" (Immunal) (14.29%) or "Парацетамол" (Paracetamol) (7.41%) the rate of usage on the advice of patients' acquaintances is close to doctors' recommendations or higher. "Кагоцел" (Kagocel) has the highest percentage for advertisement as the source (9.3%) compared to other drugs.

The distribution of the tonality (positive or negative) for the sources of information is presented in Fig. 4. A source is marked as "positive" if positive dynamic is appeared after the use of drug (i.e. review includes "BNE-pos" attribute). "Negative" tonality is marked if negative dynamic or deterioration in health has taken place or drug has had no effect (i.e. "Worse", "ADE-Neg" or "NegatedADE" mentions appear). It follows from the diagram that drugs prescribed by the doctor are mentioned more often as having positive effect, while using drugs based on an advertisement often leads to deterioration in health.

Diagrams in Fig. 5 show parts of reviews where drugs were mentioned along with labeled effects from all reviews with this drug (only top 20 drugs by entries count presented on figure). The following drugs have largest parts for ADR in reviews: immunomodulator – "Изопринозин" (Isoprinosine) (57.7%), sleeping pills – "Донормил" (Donormil) (45.5%); antiviral – "Амизон" (Amizon) (35.7%), "Тенферон лайт" (Genferon Light) (34.8%), "Амиксин" (Amiksin) (30%), etc.

Users mention that some drugs causing negative dynamics after start or some period of using it (ADE-Neg). Examples of such drugs are "Донормил" (Donormil) (13%), "Кортексин" (Cortexin) (9%), "Тенферон лайт" (Genferon Light) (8%), "Амиксин" (Amiksin) (6%), "Глицин" (Glycine) (6.6%). Also homeopathic drugs were marked as the ones with no effect: "Анаферон детский" (Anaferon for children) (64%), "Анаферон" (Anaferon) (54.6%), "Тенотен" (Tenoten) (52%).

According to reviews some of the drugs causes deterioration in health after taking the course ("Worse" label): immunomodulator – "Изопринозин" (Isoprinozine) (15%), "ИРС19" (IRS19) (13%), "Амиксин" (Amiksin) (10%), "Парацетамол" (Paracetamol) (7%) and other.

This corpus is used further to get a baseline accuracy estimate for the named entity recognition task.

3.4. Model

We consider the problem of mention detection and classification as a multi-label classification of tokens – words and punctuation marks – in sentences. For each of the three entities – ADR, Medication and Disease – its own neural network is trained. That way, mentions of different entities can intersect, so that one word can have several tags.

The output for each token is a tag in the BIO format: the "B" tag indicates the first word of a mention of the considered entity, 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.

The input to the model is a sequence of features extracted from tokens. We use part-of-speech tags, word vector representations, common features and coded word characters as an input to the model in all experiments and consider it the basic set of features. Further, in the tables with experiments, we will indicate only additional features and type of word vector representation model, implying that the basic features are present.

3.4.1. Features

Tokenization and Part-of-Speech tagging. For these tasks we used Udpipe [51] tool. After parsing

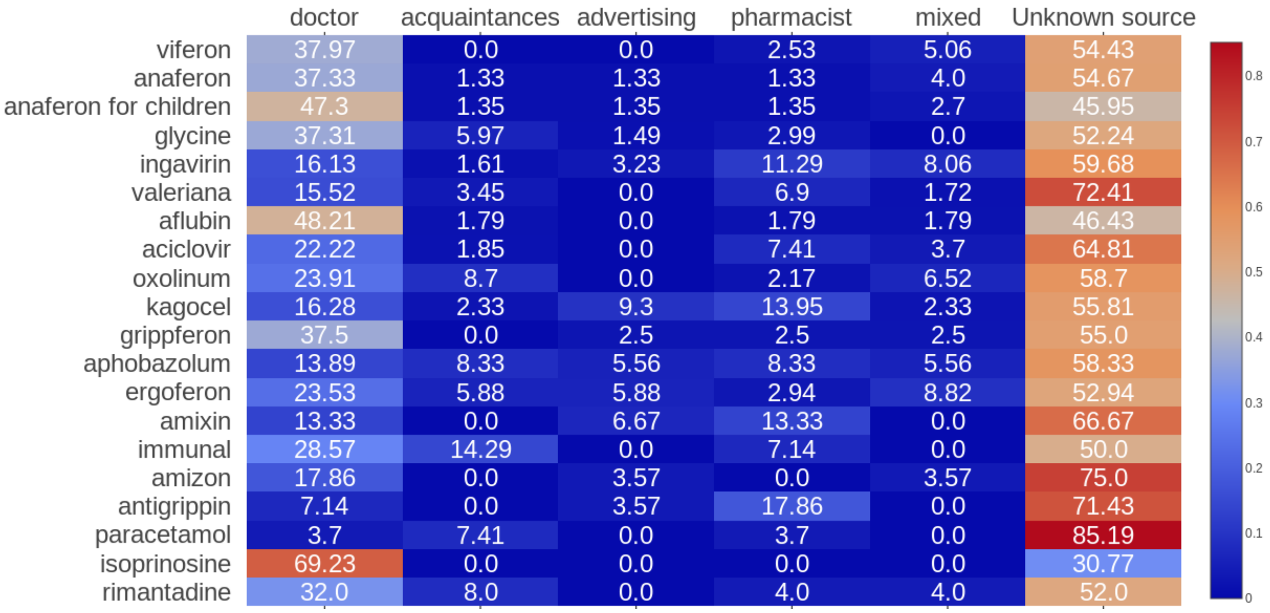

Figure 3: The distribution heatmap of entity percentages for different sources for the 20 most popular drugs. The number in a cell means the percentage of entries of a certain drug name used by recommendation from the corresponding source of information. If there were several different sources mentioned, it counted as "mixed" source

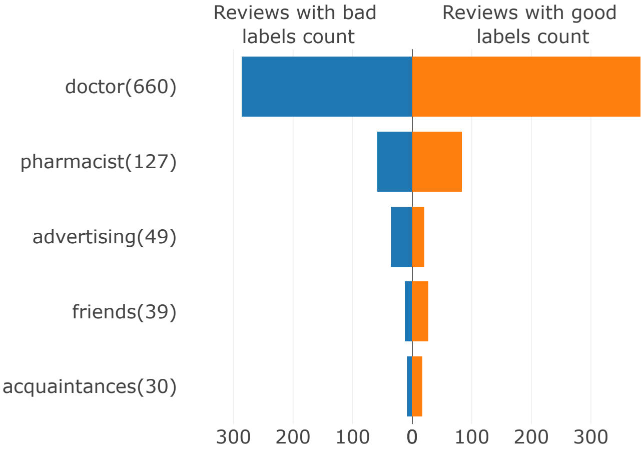

Figure 4: Distribution of the tonality for the different sources.

each word get 1 of 17 different parts of speech. They are represented as a one-hot vector and used as an input for the neural network model.

Word vector representations. The main idea is to represent a word by a vector in a special space where words with similar meanings are close to each other. A set of experiments was conducted to choose the best model for word vector representation. The following models were compared: FastText [2], ELMo (Embeddings from Language Model) [39], and BERT (Bidirectional Encoder Representations from Transformer) [9]. The key idea of the FastText model is based on the Word2Vec model principles: word distributions are predicted by their context, but FastText uses character trigrams as a basic vector representation. Each word is represented as a sum of trigram vectors that are the base for continuous bag of words or skip-grams algorithms [30]. 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 (like 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 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: words masking and logic check of the next sentence. The first strategy implies replacing 15% of the words on a token "MASK" which is later used as a target for the neural network to predict actual words. In the

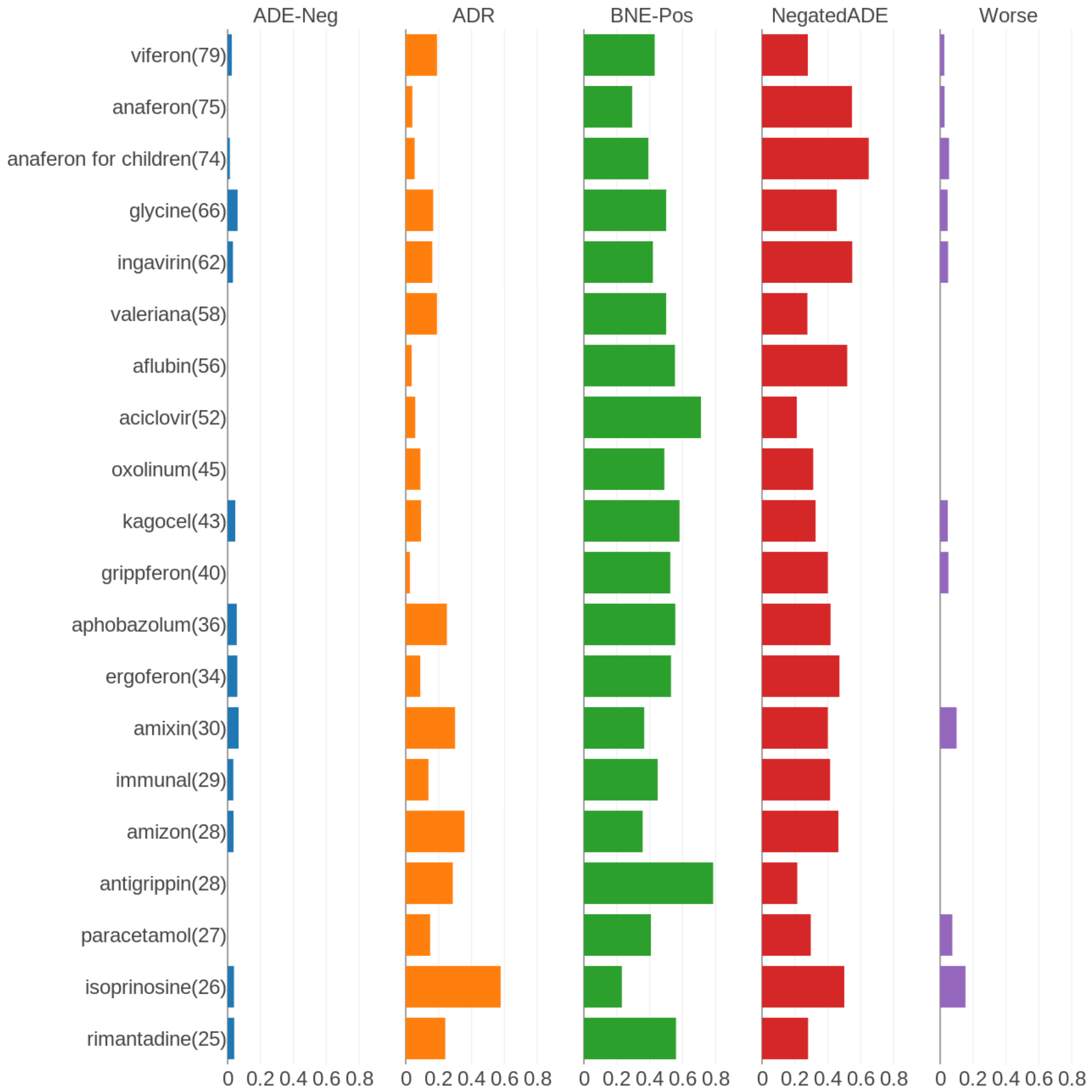

Figure 5: Distributions of labels of effects reported by reviewers after using drugs. Top 20 drugs by the entries count are presented. The number in brackets is the number of entries for a drug.

second learning strategy, the neural network should determine if two input sentences are logically sequenced or are just a set of random phrases. In BERT training, both strategies used simultaneously so as to minimize their combined loss function.

Word characters coding. For these we used convolution based neural network, CharCNN [22]. First, each word is represented as a character sequence. The number of characters is a hyperparameter, which in this study has 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. The training dataset is used to make a character vocabulary that also includes special characters «PADDING» and «UNKNOWN», the latter allowing for possible future occurrence of characters not present in the training set. For coding each character embedding layer [11] is used, which replaces every character from vocabulary appeared in a word to a corresponding real vector. In the beginning, the real vectors are initialized with values from random uniform distribution in the range of [-0.5; 0.5]. The size of real vectors is 30. Further, the matrix of coded characters of word is processed by convolution layer (with 30 filters and ker-

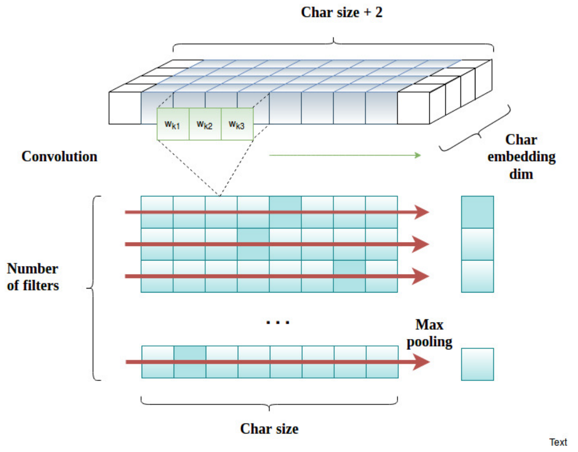

Figure 6: The scheme of character feature extraction on base of char convolution neural network. Each input vector after the embedding layer is expanded with two extra padding object (white boxes), $w_{(k1)}, w_{(k2)}, w_{(k3)}$ - weights of convolution filter k.

nel size = 3) [10] and global maxpooling function that provided maximization function of all values for each filter [3].

Common features. They are represented as 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?
- Does more than a half of the word consist of numbers?
- Does the entire word consist of numbers?
- Are all letters Latin?

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 [52]. Emotional words are taken from the dictionary [59] which contains 37 emotion categories, such as «Anxiety», «Inspiration», «Faith», «Attraction», etc. On the basis of the n available dictionaries, an n-dimensional binary vector is formed for each word, where each vector component reflects the presence of the word in a certain dictionary.

In addition, this word feature vector is concatenated with emotional features of the whole text. These features are LIWC and psycholinguistic markers.

The former is a set of specialized English Linguistic Inquiry and Word Count (LIWC) dictionaries [54], adapted for the Russian language by linguists [28]. The LIWC values are calculated for each document based on the occurrence of words in specialized psychosocial dictionaries.

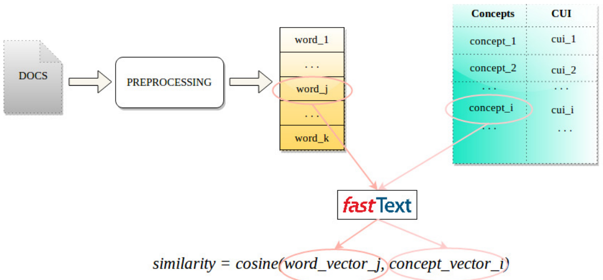

Figure 7: The matching scheme between words of corpus and concepts of UMLS.

Psycholinguistic text markers [47] reflect the level of the emotional intensity of the text. They are calculated as the ratio 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; the number of question marks, exclamation points, and average sentence length. The combination of these features are referred to as "ton" in Table 12.

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

1. Word vectors formed on base of the MESHRUS thesaurus as described in Section 3.3. The two approaches described in that section are referred to as MESHRUS and MESHRUS-2. The resulting CUI codes are encoded with one-hot representation.
2. Vidal. For each word, a binary vector is formed, which reflects belonging to categories from the Vidal medication handbook [55]: 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.

3.4.2. Model topology

The features described above are passed to the supervised model based on Long short-term memory (LSTM [16]), depicted in Fig. 3.4.1. At the output of the model, we put either a fully connected layer [7] or

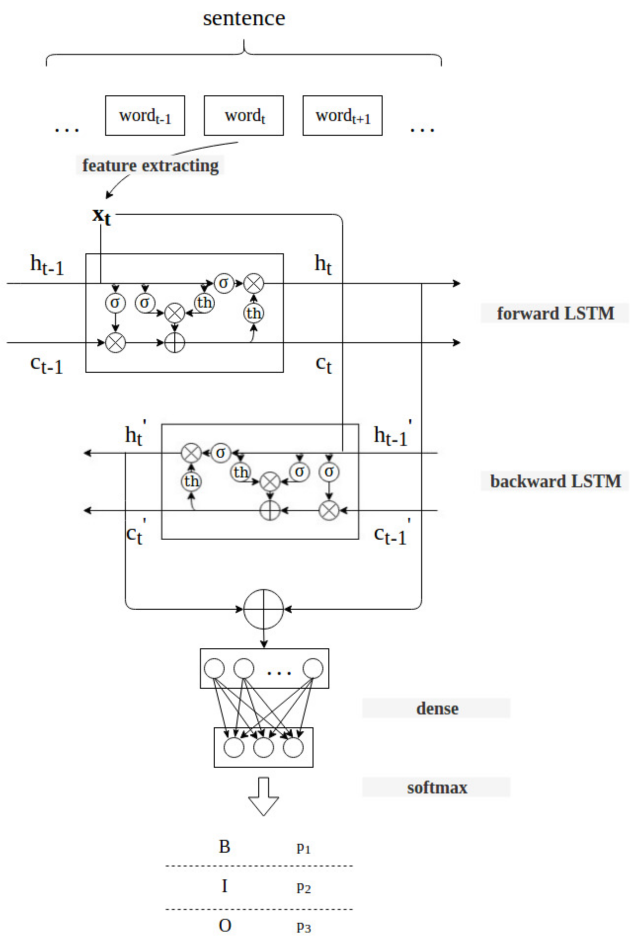

Figure 8: The main architecture of the network. Input data goes to bidirectional LSTM, where the hidden states of forward LSTM and backward LSTM get concatenated, and the resulting vector goes to fully-connected layer with size 3 and SoftMax activation function. The output p_1, p_2, and p_3 are the probabilities for the word to belong to the classes B, I, and O, i. e. to have B, I, or O tag.

conditional random fields (CRF [24]), 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).

LSTM. LSTM is a modification of a recurrent neural network (RNN) which computes on each time step t a new hidden state h_t on base of the previous hidden state h_{t-1} and the input vector x_t processed with an activation function (e.g. hyperbolic tangent function). Though RNN is able to process long input sequences, its training is complicated due to "gradient vanishing" which occurs when propagating the error back through many time steps on each of which the activation function was applied. In order to address the problem, LSTM networks have an additional item – memory cell c_t which is a linear combination of h_{t-1} and x_t. The LSTM memory cell in its processing interacts with 3 "gates": a) f_t controls which part of the previous cell memory should be "forgotten", b) i_t controls which part of the input should be saved in the memory cell, c) o_t controls which part of memory cell will be outputted on each step as the cell hidden state. The value of a memory cell after receiving a new input x_t is computed as follows:

$$\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{ix}h_{t-1} + W_{ic}c_t + b_i),\\
f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f),\\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c),
\end{aligned}$$

where σ is an activation function (e.g. sigmoid), tanh is another activation function applied element-wise to its argument vector, and $\circ$ is the Hadamard product. On each step, the h_t value is modified with the third gate o_t:

$$\begin{aligned}
o_t &= \sigma(W_{ox}x_t + W_{oc}h_{t-1} + W_{oc}c_{t-1} + b_o),\\
h_t &= o_t \cdot tanh(c_t)
\end{aligned}$$

Multiple LSTM layers could be used in series to increase the capacity and performance of an LSTM-based network (stacked LSTM topology). In that way in the research topology with 3 sequential LSTM layers show quality increasing in comparison with a single-layer LSTM.

CRF. It is an implementation of hidden Markov models (HMM), a graph model for the representation of joint probabilities of several random values. CRF [24] is defined as follows. Let X be a random variables over a sequence to be classified, Y – random variables mapping into a sequence labels. Also, let $G = (V, E)$ be such a graph that $Y = (Y_v)_{v \subset V}$, so Y is indexed by vertices of G, then (X, Y) is a conditional random field in case when each random variable Y_v (that depends of X) has the Markov property with respect to the graph. Here E reflects all dependencies between variables in a random field (X, Y). CRF represents the following distribution of random variables set:

$$p(\bar{y}|\bar{x}; w) = \frac{exp(\sum_i \sum_j w_j f_j(y_i, y_{i-1}, \bar{x}, i))}{\sum_{y' \subset Y} exp(\sum_i \sum_j w_j f_j(y_i, y_{i-1}, \bar{x}, i))} (*),$$

where f_j – features functions, w_j – weights for feature function j. The task is to find y^* values to maximize the equation

$$y^* = \underset{y}{\text{argmax}} \max P(\bar{y}|\bar{x}; w).$$

3.5. Quality Metrics

In order to assess partially correct determination of mention boundaries, we employ two evaluation metrics:

1. Exact mention matching F_1^{exact};
2. Partial matching F_1^{partial}.

Table 10
Accuracy (%) of recognizing ADR, Medication and Disease entities in our corpus by models with different embeddings.

Embedding type	dim	ADR		Medication		Disease	
		$f_{1_{partial}}$	$f_{1_{exact}}$	$f_{1_{partial}}$	$f_{1_{exact}}$	$f_{1_{partial}}$	$f_{1_{exact}}$
FastText	300	33.6 ± 3.3	22.4 ± 1.6	77.2 ± 0.6	70.4 ± 1.1	59.3 ± 1.7	44.1 ± 1.7
ELMo	1024	46.6 ± 4.2	24.3 ± 1.7	85.5 ± 0.6	73.4 ± 1.5	70.8 ± 1.1	46.4 ± 0.6
BERT	768	30.6 ± 5.8	22.1 ± 2.4	78.8 ± 6.8	71.4 ± 3.3	63.6 ± 2.8	45.5 ± 3.2

F_1^{exact}. For every entity (in our case, ADR, Medication, and Disease) from the ground truth set we calculate precision, recall and F_1 as follows:

$$\text{precision} = \sum_{e_s \in E_s} \frac{[e = e_s]}{|E_s|}$$

$$\text{recall} = \sum_{e \in E} \frac{[e = e_s]}{|E|}$$

$$F_1^{\text{exact}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

where e_s is a predicted mention, e is the corresponding ground truth mention, E is the ground truth set of mentions, E_s is the set of mentions predicted by the model, $|E|$ is the number of items in E, $[e = e_s]$ is the Iverson bracket which is 1 if the mentions e and e_s are equal, and 0 otherwise.

While comparing mentions by equality, if an O-tag precedes an I-tag, the latter is replaced with a B-tag.

F_1^{partial}. For every i-th sentence from the test dataset we calculated the values of precision_i, recall_i, and F_{1i} using the following equations:

$$\text{precision}_i = \frac{|t_i \cap t_{si}|}{|t_{si}|}$$

$$\text{recall}_i = \frac{|t_i \cap t_{si}|}{|t_i|}$$

$$F_{1i} = 2 \cdot \frac{\text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i}$$

where t_{si} is the list of tokens of i-th sentence that were recognized by the model as parts of mentions, t_i is the list of tokens belonging to ground truth mentions of i-th sentence, and $|t_i|$ is the list length (the number of tokens in t_i). The final F_1^{partial} is calculated as the mean of F_{1i} over all i in the set T of sentences in the text that contain any mentions.:

$$F_1^{\text{partial}} = \frac{1}{|T|} \sum_{i=1}^{|T|} F_{1i}.$$

3.6. Experiments

3.6.1. Finding the best embedding.

We considered the following embedding models: FastText, ELMo, and BERT. Two corpora were used to train the FastText model – a corpus of reviews from Otzovik.com from the category "medicines" and a corpus of reviews from the category "hospitals" [9], also we used vectors pretrained on the Commoncrawl corpus[10]. The ELMo model which had been preliminarily trained on the Russian WMT News [21] was taken from the DeepPavlov [11] [5] open-source library. The pretrained multilingual BERT model was taken from the Google repository [12] and subsequently fine-tuned on the above-mentioned corpora of drug and hospital reviews. These pretrained models were used as input to our neural network model presented in Fig. 3.4.1. The dataset was split into 5 folds for cross-validation. On each fold, the training set was split into training and validation sets in the ratio 9:1. Training was performed for a maximum of 70 epochs, with early stopping by the validation loss. Cross entropy was used as the loss function, with nAdam as the optimizer and cyclical learning rate mechanism [50]. The results of the test experiments are given in Table 10.

3.6.2. Comparing the numerical results of our model on the CADEC corpus to the known literature results.

In this case, models were trained on the CADEC corpus of drug reviews, from which the following objects are extracted: ADR, Drug, Symptoms, Findings, Disease. The work [29] was devoted to the extraction of Disease and Drug entities, while the Disease entity was presented there as a combination of the tags ADR, Disease, Findings and Symptoms. The model was based on CRF with word2vec embeddings, called HealthVec, pretrained on the Health Dataset [29]. Part-of-speech tags, word shape features, syntactic relations and dictionaries were used as features. Learning was preformed with 5-fold cross-validation.

The second part of Table 11 presents a comparison of our model, employing LSTM and various features, with that model. F_1^{partial} and F_1^{exact} calculated for the Disease and Drug entities are given as final estimates.

The work [56] was devoted to extracting only the

[9] Reviews were taken from the Otzovik website from the categories "hospitals" and "medicines" - https://otzovik.com/health/

[10] http://commoncrawl.org/

[11] https://deeppavlov.readthedocs.io/en/master/intro/pretrained_vectors.html

[12] https://github.com/google-research/bert/

Table 11
The accuracy of our model on the CADEC corpus compared to other models.

	Method	F_1^{partial}	F_1^{exact}
ADR			
Our model	ELMo, BERT, CRF, pos, 3-layer LSTM	68.8 ± 1.5	78.84 ± 2.8
Tutubalina E. et al., 2017 [56]	3-layer GRU, CNN, CRF	70.65	79.78
	3-layer LSTM, CNN, CRF	69.65	81.15
Disease + Drug			
Our model	ELMo, BERT, CRF, pos, 3-layer LSTM	75.6 ± 1.4	86.3 ± 0.7
Miftahutdinov Z. Sh. et al., 2017 [29]	CRF, HealthVec, all features	69.1	79.4
	3-layer LSTM, HealthVec	67	81.2

Table 12
Entity recognition accuracy (%) on our corpus of the models with different features and topology.

Topology and features	ADR		Medication		Disease	
	F_1^{partial}	F_1^{exact}	F_1^{partial}	F_1^{exact}	F_1^{partial}	F_1^{exact}
ELMo + ton	44.9 ± 6.8	26.6 ± 3.9	85.6 ± 0.4	73.5 ± 0.5	70.8 ± 0.7	47.3 ± 1.0
ELMo + MESHRUS	48.6 ± 4.3	27.4 ± 2.2	85.2 ± 0.7	73.3 ± 1.5	71.3 ± 0.4	46.5 ± 1.2
ELMo + PoS	46.5 ± 6.5	26.2 ± 3.0	85.6 ± 0.7	72.9 ± 0.6	71.5 ± 0.9	46.6 ± 0.9
ELMo + BERT	26.2 ± 13.3	18.7 ± 9.8	84.6 ± 1.1	74.1 ± 1.1	67.6 ± 2.5	47.9 ± 1.6
ELMo + Vidal	47.1 ± 2.8	26.8 ± 1.0	85.6 ± 0.6	73.2 ± 1.1	71.5 ± 0.9	45.8 ± 1.2
ELMo + CRF	**51.7 ± 6.0**	28.8 ± 2.7	85.6 ± 0.8	73.2 ± 1.1	71.6 ± 0.8	46.9 ± 0.4
3-layer LSTM, ELMo	44.6 ± 7.4	28.2 ± 5.1	**86.7 ± 0.5**	74.7 ± 0.7	**73.4 ± 1.2**	51.5 ± 1.8
ELMo + MESHRUS-2	49.7 ± 2.7	27.4 ± 0.9	85.6 ± 0.8	73.1 ± 0.4	71.0 ± 1.0	46.7 ± 1.4
ELMo, ton, PoS, MESHRUS, MESHRUS-2, Vidal, CRF, 3-layer LSTM	48.7 ± 5.6	**32.4 ± 4.7**	86.5 ± 0.6	74.6 ± 1.1	73.2 ± 0.7	52.3 ± 1.4
ELMo, BERT, ton, 3-layer LSTM	38.1 ± 7.0	28.0 ± 1.9	86.1 ± 0.7	75.6 ± 0.4	69.2 ± 3.7	**54.2 ± 0.9**
ELMo, BERT, ton, PoS, MESHRUS, MESHRUS-2, Vidal, CRF, 3-layer LSTM	40.3 ± 7.7	30.2 ± 4.6	85.9 ± 1.1	**76.3 ± 2.1**	71.4 ± 1.3	52.2 ± 2.7

ADR entity with the help of a model combining recurrent neural networks and CRF. The corpus was split into a training set and a testing set in the proportion 70:30. The first part of Table 11 compares our model to that work.

3.6.3. Choosing the best combinations of model topology with the selected features.

Next, we provide a set of experiments on the choice of topology: replacing the last fully-connected layer with a CRF layer, or changing the number of biLSTM layers. This was studied in combination with adding emotion markers, PoS and MESHRUS, MESHRUS-2 and Vidal dictionaries, as shown in Table 12.

4. Results

4.1. Results of embeddings comparison experiments.

These results are presented 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. Consequently, in further experiments on adding features and changing the topology we use the ELMo embedding as the basic approach.

4.2. Results of comparing our model to the literature.

The purpose of presenting the comparative data in Table 11 is to confirm the general quality of our model. For this, we present the results of our model on the CADEC corpus and compare them to the recent works [56, 29]. Even though after having to exclude the Russian-specific corpus features we did not perform the time-consuming extraction of analogous English features from CADEC, our model showed results comparable to the modern computational results on CADEC. This confirms the applicability of our model to evaluating the state of the art for the developed corpus.

4.3. Results of choosing the best model topology and input feature set.

For our corpus, as shown in Table 12, various changes in features and topology were added to the basic model with ELMo embedding. First of all, we focused on the metric F_1^{exact}, since it reflects the quality of the model better. Adding normalization gave the greatest increase in the least-represented class ADR. As a result, a combination of dictionary features, emotion markers, 3-layers LSTM and CRF can achieve the highest quality increase in ADR and Disease entities. For Medication, the best result was shown by a combination of ELMo, BERT embeddings, emotion markers, and 3-layer LSTM.

5. Discussion

Currently there are a significant diversity of corpora in different languages to analyze the safety and effectiveness of drugs. However, for the Russian language, such corpus is the first. In our opinion, a conclusion about its quality can only be made on base of calculational analysis by advanced models. With this in mind, we solved simultaneously two interconnected tasks: the Russian corpus annotation and the creation of the machine learning complex, testing it preliminarily on the commonly acknowledged CADEC corpus. This complex was then used to evaluate the quality of our corpus. It should be noted that it is quite difficult to compare models even on the same corpus, as the authors often use different metrics and varied data splits. Hence, from the very onset we adhered to metrics similar to the works with which we planned to compare the results.

The model has shown accuracy comparable to the up-to-date literature results, and thus proved itself applicable for establishing the state of the art for our newly-created corpus. The reasonable level of entity identification accuracy achieved on our corpus, in its turn, confirms the validity of the latter.

6. Conclusion

The basic result of this work is the creation of the first Russian tagged corpus of pharmacological texts which is approved by a complex of modern deep learning neuronet models with up-to-date language feature embeddings. The quality of this complex is confirmed by comparison with well-known literature results on CADEC.

The level of accuracy obtained by the developed complex on our corpus is comparable to those obtained on similar corpora of other languages and may be seen as the state of the art for the task in view. The relatively low accuracy results for adverse drug effects (ADR) can be explained by the low representativeness of such entity type in the current version of the corpus. The developed neuronet complex may be used as a base for the replenishment of the corpus by ADR. This, along with including new entities and relations, is a goal of further work.

Acknowledgments

This work has been supported by the Russian Science Foundation grant 20-11-20246 and 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/.

References

[1] Alvaro, N., Miyao, Y., Collier, N., 2017. Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations.

[2] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.

[3] Boureau, Y.L., Ponce, J., LeCun, Y., 2010. A theoretical analysis of feature pooling in visual recognition, in: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 111–118.

[4] Brown, E.G., Wood, L., Wood, S., 1999. The medical dictionary for regulatory activities (meddra). Drug safety 20, 109–117.

[5] Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., et al., 2018. Deeppavlov: open-source library for dialogue systems, in: Proceedings of ACL 2018, System Demonstrations, pp. 122–127.

[6] Campbell, K.E., Oliver, D.E., Shortliffe, E.H., 1998. The unified medical language system: toward a collaborative approach for solving terminologic problems. Journal of the American Medical Informatics Association 5, 12–16.

[7] Chiu, J.P., Nichols, E., 2016. Named entity recognition with bidirectional lstm-cnns. Transactions of the Association for Computational Linguistics 4, 357–370.

[8] Dai, X., Karimi, S., Paris, C., 2017. Medication and adverse event extraction from noisy text, in: Proceedings of the Australasian Language Technology Association Workshop 2017, pp. 79–87.

[9] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

[10] Dumoulin, V., Visin, F., 2016. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285 .

[11] Gal, Y., Ghahramani, Z., 2016. A theoretically grounded application of dropout in recurrent neural networks, in: Advances in neural information processing systems, pp. 1019–1027.

[12] Gupta, S., Gupta, M., Varma, V., Pawar, S., Ramrakhiyani, N., Palshikar, G.K., 2018. Multi-task learning for extraction of adverse drug reaction mentions from tweets, in: European Conference on Information Retrieval, Springer. pp. 59–71.

[13] Gurulingappa, H., Rajput, A.M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L., 2012. Development of a

benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of biomedical informatics 45, 885–892.

[14] Gurulingappa, H., M.R.A..T.L., 2012. xtraction of potential adverse drug events from medical case reports. Journal of biomedical semantics 3(1).

[15] Hanisch, D., Fundel, K., Mevissen, H.T., Zimmer, R., Fluck, J., 2005. Prominer: rule-based protein and gene entity recognition. BMC bioinformatics 6, S14.

[16] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

[17] Jagannatha, A., Liu, F., Liu, W., Yu, H., 2019. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). Drug safety 42, 99–111.

[18] Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G., 2016. Mimic-iii, a freely accessible critical care database. Scientific data 3, 160035.

[19] Johri, N., Niwa, Y., Chikka, V.R., 2014. Optimizing apache ctakes for disease/disorder template filling: Team hitachi in 2014 share/clef ehealth evaluation lab .

[20] Karimi, S., Metke-Jimenez, A., Kemp, M., Wang, C., 2015. Cadec: A corpus of adverse drug event annotations. Journal of biomedical informatics 55, 73–81.

[21] Koehn, P., 2019. Statmt - internet resource about research in the field of statistical machine translation. URL: www.statmt.org. accessed: 2019-05-24.

[22] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

[23] Kuhn, M., Letunic, I., Jensen, L.J., Bork, P., 2015. The sider database of drugs and side effects. Nucleic acids research 44, D1075–D1079.

[24] Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data .

[25] Laguna, J.Y., 2003. Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias .

[26] Li, F., Zhang, M., Fu, G., Ji, D., 2017. A neural joint model for entity and relation extraction from biomedical text. BMC bioinformatics 18, 198.

[27] Library, S.C.S.M., 2019. Russian translation of the medical subject headings. URL: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHRUS/. accessed: 2019-05-24.

[28] Litvinova, O., Seredin, P., Litvinova, T., Lyell, J., 2017. 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, pp. 43–52.

[29] Miftahutdinov, Z., Tutubalina, E., Tropsha, A., 2017. Identifying disease-related expressions in reviews using conditional random fields, in: Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog, pp. 155–166.

[30] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

[31] Miller, G., Britt, H., 1995. A new drug classification for computer systems: the atc extension code. International journal of bio-medical computing 40, 121–124.

[32] Mowery, D.L., Velupillai, S., South, B.R., Christensen, L., Martinez, D., Kelly, L., Goeuriot, L., Elhadad, N., Pradhan, S., Savova, G., et al., 2014. Task 2: Share/clef ehealth evaluation lab 2014, in: Proceedings of CLEF 2014.

[33] NEHTA, 2014. Australian medicines terminology v3 model–common v1.4, tech. rep. ep-1825:2014, national e-health transition authority.

[34] Organization, W.H., et al., 2004. International statistical classification of diseases and related health problems: tenth revision-version, 2nd ed.

[35] Oronoz, M., Casillas, A., Gojenola, K., Perez, A., 2013. Automatic annotation of medical records in spanish with disease, drug and substance names, in: Iberoamerican Congress on Pattern Recognition, Springer. pp. 536–543.

[36] Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A.D., Casillas, A., 2015. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. Journal of biomedical informatics 56, 318–332.

[37] Patrick, J., Li, M., 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. Journal of the American Medical Informatics Association 17, 524–527.

[38] Perez, A., Weegar, R., Casillas, A., Gojenola, K., Oronoz, M., Dalianis, H., 2017. Semi-supervised medical entity recognition: A study on spanish and swedish clinical corpora. Journal of biomedical informatics 71, 16–30.

[39] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365 .

[40] Pradhan, S., Elhadad, N., South, B.R., Martinez, D., Christensen, L.M., Vogel, A., Suominen, H., Chapman, W.W., Savova, G.K., 2013. Task 1: Share/clef ehealth evaluation lab 2013., in: CLEF (Working Notes).

[41] Ratcliff, J.W., Metzener, D.E., 1988. Pattern-matching-the gestalt approach. Dr Dobbs Journal 13, 46.

[42] Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J.S., Roberts, I., Setzer, A., Tapuria, A., et al., 2007. The clef corpus: semantic annotation of clinical text, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association. p. 625.

[43] Roberts, A., Gaizauskas, R.J., Hepple, M., Guo, Y., 2008. Combining terminology resources and statistical methods for entity recognition: an evaluation., in: LREC.

[44] Rosminzdrav, 2019. State register of drugs. URL: https://grls.rosminzdrav.ru/. accessed: 2019-05-24.

[45] Sarker, A., Gonzalez, G., 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics 53, 196–207.

[46] Sarker, A., Nikfarjam, A., Gonzalez, G., 2016. Social media mining shared task workshop, in: Biocomputing 2016: Proceedings of the Pacific Symposium, World Scientific. pp. 581–592.

[47] Sboev, A., Gudovskikh, D., Rybka, R., Moloshnikov, I., 2015. A quantitative method of text emotiveness evaluation on base of the psycholinguistic markers founded on morphological features. Procedia Computer Science 66, 307–316.

[48] Shahpori, R., Doig, C., 2010. Systematized nomenclature of medicine–clinical terms direction and its implications on critical care. Journal of critical care 25, 364–e1.

[49] Shelmanov, A., Smirnov, I., Vishneva, E., 2015. Information extraction from clinical texts in russian, in: Computational Linguistics and Intellectual Technologies: Annual International Conference “Dialog.

[50] Smith, L.N., 2017. Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 464–472.

[51] Straka, M., Hajic, J., Strakov'a, J., 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing., in: LREC.

[52] Suero Montero, C., Munezero, M., Kakkonen, T., 2014. Investigating the role of emotion-based features in author gender classification of text. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8404 LNCS, 98–114. doi:10.1007/978-3-642-54903-8_9.

[53] Tang, B., Wu, Y., Jiang, M., Denny, J.C., Xu, H., 2013. Recognizing and encoding discorder concepts in clinical text using machine learning and vector space model. CLEF (Working Notes) 665.

[54] Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology 29, 24–54.

[55] Tolmachova, E., 2019. Spravochnik Vidal. Lekarstvenie preparati v Rossii. Vidal Rus.

[56] Tutubalina, E., Nikolenko, S., 2017. Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. Journal of healthcare engineering 2017.

[57] Uzuner, Ö., Solti, I., Cadag, E., 2010. Extracting medication information from clinical text. Journal of the American Medical Informatics Association 17, 514–518.

[58] Wang, W., 2016. 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.

[59] Website, 2019. Information retrieval system ”emotions and feelings in lexicographical parameters: Dictionary emotive vocabulary of the russian language”. URL: http://lexrus.ru/default.aspx?p=2876.

[60] Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J., 2006. Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic acids research 34, D668–D672.

[61] Wunnava, S., Qin, X., Kakar, T., Rundensteiner, E.A., Kong, X., 2018. Bidirectional lstm-crf for adverse drug event tagging in electronic health records, in: International Workshop on Medication and Adverse Drug Event Detection, pp. 48–56.

[62] 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., 2019a. The psytar dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. Data in brief 24, 103838.

[63] Zolnoori, M., Fung, K.W., Patrick, T.B., Fontelo, P., Kharrazi, H., Faiola, A., Wu, Y.S.S., Eldredge, C.E., Luo, J., Conway, M., et al., 2019b. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for ssri and snri medications. Journal of biomedical informatics 90, 103091.