Сравнительный анализ нейронных сетей в задаче классификации побочных эффектов на уровне сущностей в англоязычных текстах

И.С. Алимова <alimovailseyar@gmail.com>
Е.В. Тутубалина <tutubalinaev@gmail.com >
Казанский (Приволжский) федеральный университет,
420008, Казань, ул. Кремлёвская, 18

Аннотация. В данной работе представлено экспериментальное исследование эффективности ряда моделей нейронных сетей для задачи классификации побочных эффектов на уровне сущностей. Задача анализа тональности на уровне аспектных терминов, в которых необходимо определить мнение по отношению к конкретному аспекту, активно исследуется в течении последнего десятилетия. Для решения данной задачи в прошедшие годы было предложено несколько архитектур нейронных сетей. Несмотря на то, что модели, основанные на этих архитектурах, имеют много общего, есть некоторые компоненты, которые отличают их друг от друга. В данной статье была исследована применимость разработанных для аспектно ориентированного анализа тональности нейросетевых моделей для классификации побочных эффектов. Для оценки эффективности данных методов были проведены обширные эксперименты на различных англоязычных текстах биомедицинской тематики, включающих в себя записи клинических карточек, научную литературу и данные из социальных сетей. Также мы сравнили предлагаемую модель с одной из наилучших на данный момент моделей, основанной на методе опорных векторов и большом наборе признаков.

Ключевые слова: побочный эффект; обработка естественного языка; анализ социальных медиа; машинное обучение; глубокое обучение; нейронные сети

DOI: 10.15514/ISPRAS-2018-30(5)-11

Для цитирования: Алимова И.С., Тутубалина Е.В. Сравнительный анализ нейронных сетей в задаче классификации побочных эффектов на уровне сущностей в англоязычных текстах. Труды ИСП РАН, том 30, вып. 5, 2018 г., стр. 177-196. DOI: 10.15514/ISPRAS-2018-30(5)-11

1. Введение

В настоящее время в связи с бурным развитием сети интернет и электронных коллекций научных публикаций имеются обилие неструктурированной

информации, представленной текстами на естественном языке. В число активно развивающихся направлений обработки текстовой информации входят задачи медицинской науки, в частности, задачи фармакологии и персонализированной медицины. Всё более востребованной становится задача автоматической обработки текстов медицинской направленности с целью извлечения структурированных данных, которые затем используются при решении различного рода проблем: поиска информации о побочных реакциях лекарственных препаратов, использовании лекарств с нарушением предписаний инструкции, определении действия лекарственных препаратов по отношению к системам организма, излечения новых отношений между лекарствами и симптомами для построения гипотез о перепрофилировании препарата.

Для выявления новых побочных эффектов, не указанных в инструкции по применению препарата, все большую популярность приобретает подход с применением текстов медицинской тематики: электронных карточек пациентов, научной литературе, записей пациентов в социальных сетях и медицинских форумах. Обработка такого объема информации невозможна вручную, поэтому активно применяются методы автоматической обработки естественного языка [1-5].

Классификацию побочных эффектов можно рассматривать в двух направлениях: (i) на уровне сообщения и (ii) на уровне сущности. В первом случае необходимо определить наличие упоминания побочного эффекта во фрагменте текста, например, предложении или тексте твита. Данный тип классификации необходим для очистки коллекции текста от нерелевантных документов. Во втором случае классификация применяется к результатам работы алгоритмов извлечения именованных сущностей. В данной работе мы сосредоточились на второй задаче.

Одна из разновидностей задач классификации относительно сущностей - это аспектно-ориентированный анализ тональности. В аспектно-ориентированном анализе тональности определяется отношение пользователя не только к объекту в целом, но и к отдельным его частям или аспектам. Существующие работы показали успешность применения ряда архитектур нейронных сетей, основанных на сетях с короткой долгосрочной памятью (англ. long short-term memory; LSTM). В данной статье разработанные методы были адаптированы и применены для задачи классификации побочных эффектов.

Исследования были начаты с простых моделей, использующих только LSTM, далее архитектуры расширялись механизмами внимания и дополнительной

памятью. В качестве моделей были взяты следующие архитектуры нейронных сетей:

i. сеть с короткой долгосрочной памятью (англ. long short-term memory; LSTM) - базовая модель, которая использует все предложение, закодированное векторным представлением слов, в качестве входа;
ii. модель с заданной целью (англ. Target-Dependent LSTM; TD-LSTM) [6] которая использует два слоя LSTM для моделирования правого и левого контекста относительно сущности;
iii. сеть с механизмом интерактивного внимания (Interactive Attention Network; IAN) [7], которая состоит из двух слоев LSTM для представления предложения и целевой сущности и слоев с перекрестным вниманием, объединенные выходы которых передаются слою с логистической функцией для принятия решения о классификации;
iv. сеть с глубокой памятью (Deep Memory Network; MemNet) [8], которая применяет несколько раз механизм внимания к входному слою векторного представления слов, выход последнего из которых передается в слой с логистической функцией для предсказания класса;
v. сеть с рекуррентным механизмом внимания к памяти (Recurrent Attention Memory; RAM) [9] расширяет модель MemNet дополнительными слоями LSTM и многократным применением механизма внимания к выходам этих слоев.

Описанные модели применялись в задаче классификации мнений для отзывов пользователей о ресторанах и ноутбуках, однако работ по применению моделей к классификации побочных эффектов на уровне сущностей из различных источников текста фармаконадзора найдено не было. В рамках данного исследования были проведены обширные эксперименты на пяти базовых наборах данных, которые состоят из текстов аннотаций биомедицинских статей, электронных карточек пациентов и текстов из социальных сетей. Проведено сравнение эффективности описанных нейронных сетей и метода на основе опорных векторов с точки зрения стандартных метрик качества классификации.

2. Обзор существующих подходов

В исследованиях применяются различные подходы для выявления побочных реакций в текстах. Наиболее широко используемый метод - это подход, основанный на словарях [10-15]. Словари состоят из списков побочных реакций, извлеченных из инструкций по применению лекарств, записей о клинических испытаниях, отзывах пользователей в социальных сетях. Первые работы были ограничены в количестве исследуемых лекарств и целевых побочных эффектов из-за ограничений терминов в словарях. Для преодоления этого ограничения стали использоваться методы на основе правил [16-17]. Основная идея этих методов заключается в том, чтобы выделить наиболее распространенные конструкции предложений, которые могут

свидетельствовать об описании побочных реакций. Однако разработка правил является длительным и трудоемким процессом, для этого требуется наличие специалиста в данной предметной области, при этом данный подход не масштабируем для новых коллекций документов.

Большинство работ описывают исследования с использованием методов машинного обучения. Например, в работах [18-23] используется метод опорных векторов (SVM), в статьях [16, 24] применяется метод условных случайных полей (CRF), а в работе [25] метод случайного леса (Random Forest). В качестве признаков, подаваемых на вход алгоритмам машинного обучения, используются: n-граммы, части речи, принадлежность к семантическим типам из унифицированного языка медицинских систем (UMLS), количество слов с отрицанием, принадлежность рассматриваемого термина к словарям побочных реакций, наличие в тексте названия лекарства, вектора word2vec, вектора кластеризации.

В 2016-м и 2017-м годах проводились соревнования по поиску побочных эффектов в сообщениях из Твиттера [26-27]. В рамках соревнования присутствовали задачи классификации на уровне всего твита и на уровне сущности. Победители первого соревнования использовали комбинацию из девяти классификаторов, основанных на модели случайного леса [27] со следующим набором признаков: 1, 2, 3 - граммы, появление вместе лекарства и побочного эффекта, наличие отрицания и оценка тональности. В качестве набора данных для каждого классификатора на вход подавались все положительные примеры и такое же количество случайным образом выбранных отрицательных примеров, что позволило участникам решить проблему несбалансированности классов. Описанная система получила 41.95% F-меры. В соревновании 2017-го года в задаче классификации на уровне твитов первое место заняла система, использовавшая метод опорных векторов в качестве модели [28]. Однако, в отличие от предыдущего года, набор признаков был более обширным, модель получила 43.5% F-меры и таким образом улучшила результаты предыдущего соревнования на 1.55%. В задаче классификации на уровне сущностей лучшие результаты показала система, использовавшая ансамбль сверточных нейронных сетей [29]. Система достигла 69.3% F-меры.

В 2016-м году появляются первые работы по классификации текстов на наличие побочных эффектов, основанные на нейронных сетях. В работе [30] применялись сверточная рекуррентная нейронная сеть и сверточная сеть с вниманием. Эксперименты проводились на двух наборах данных: твитов из соревнования 2016-го года, описанном в данном разделе выше [26] и отчетов системы MEDLINE [31]. Сверточная рекуррентная нейронная сеть показала 51% F-меры на корпусе твиттеров и 87% F-меры на корпусе MEDLINE, модель с вниманием показала 49% и 83% F-меры соответственно. Таким образом, был получен прирост на 7.5% в сравнении с результатами соревнования.

Методы по анализу тональности активно применяются в предметной области медицины и в текстах других тематик [32-35]. В области аспектно

ориентированного анализа тональности активно применяются нейронные сети [36]. Танг и др. представили архитектуру нейронной сети TD LSTM (Target-Dependent LSTM; TD_LSTM) [6] и сеть с памятью MemNet (Deep Memory Network; MemNet) для классификации на уровне аспекта [8]. Предложенные модели демонстрируют сравнимые с существующими методами результаты. Чен и др. использовали рекуррентную сеть с вниманием (Recurrent Attention Memory; RAM) [9]. Модель применяет механизм внимания несколько раз для охвата тональных признаков, находящихся на большом друг от друга расстоянии. RAM превзошла результаты описанных ранее моделей на четырех корпусах из разных предметных областей. Ма и др. предложили сеть с интерактивным вниманием (Interactive Attention Network; IAN), которая генерирует отдельные представления для контекста и аспекта и применяет для них перекрестное внимание [7]. Модель показала высокую эффективность по сравнению с различными модификациями нейронной сети с длинной короткой памятью (Long Short Term Memory; LSTM).

На основе анализа предметной области можно сделать вывод, что сравнительно мало работ посвящено применению нейронных сетей в задаче классификации побочных эффектов. Большинство работ используют методы машинного обучения, которые ограничены линейностью модели и необходимостью поиска оптимальных признаков вручную [2,12,18,21,25,27,37,38]. Кроме того большинство методов извлекали признаки непосредственно из классифицируемой сущности, уделяя малое внимание контексту или используя маленький контекст размером в 4-5 слов слева и справа относительно сущности [21,25,39,40]. Стоит также отметить, что в большинстве работ проводились исследования на одном корпусе данных.

3. Корпуса

Эксперименты по оценке эффективности методов классификации проводились на четырех существующих англоязычных корпусах: CADEC, Твиттер, MADE, Twimed. Общая статистика для всех корпусов представлена в табл. 1. В таблице класс ‘ADR’ обозначает класс с побочным эффектом, соответственно, класс ‘non-ADR’ обозначает его отсутствие. Как видно из статистики, корпуса CADEC и MADE содержат большее кол-во аннотаций, чем остальные корпуса.

3.1 CADEC

Корпус CADEC состоит из размеченных отзывов пользователей о лекарственных препаратов с форума askapatient.com [41]. В корпусе размечены 5 видов аннотаций: лекарство (drug), побочный эффект (adverse), заболевание (disease), симптом (symptom) и другие медицинские термины, не вошедшие в описанные категории (finding). Аннотацией лекарство отмечены все названия лекарственных препаратов в тексте. Все побочные эффекты, связанные с лекарством, отмечены аннотацией побочный эффект. Аннотацией заболевание обозначены показания к применению. Симптом обозначает сопутствующие

признаки болезни. Аннотации заболевание и симптом были сгруппированы вместе с аннотацией, обозначающей другие медицинские термины в одну группу.

3.2 Twitter

Корпус Twitter содержит твиты пользователей на тему здоровья [42]. В каждом твите отмечены побочные эффекты или сущности, обозначающие заболевание. Политика Твиттера не позволяет хранить и распространять твиты в открытом доступе. Создатели корпуса предоставляют только идентификатор пользователя и твита по которым можно загрузить исходный текст. В связи с этим часть твитов не удалось загрузить. Во время предобработки текста были удалены все ссылки, упоминания пользователей и ретвиты.

3.3 MADE

Корпус MADE состоит из обезличенных записей электронных карточек пациентов, больных раком. Корпус был создан для соревнования по обработке естественного языка, в задачи которого входило извлечение медицинских терминов, побочных эффектов и отношений между ними [43]. Аннотации, связанные с заболеваниями ‘SSLIF’ и ‘Indication’, были объединены в класс ‘non-ADR’.

3.4 Twimed

Корпус Twimed состоит из двух частей: твитов пользователей и текстов статей с ресурса PubMed [44]. Корпус содержит аннотации: болезнь, симптом и лекарство. Если отношение между лекарством и болезнью было размечено как негативное, то болезнь отмечалась как побочный эффект.

Табл. 1. Суммарная статистика по корпусам
Tab. 1. Summary statistics of corpora

Корпус	Источник	Кол-во документов	Кол-во ADR	Кол-во non-ADR	Максимальная длина предложения	Средняя длина предложения
CADEC [41]	Отзывы на форуме	1231	5770	550	236	28
MADE [43]	Электронные карточки пациентов	876	1506	37077	173	21
TwiMed-Pubmed [44]	Аннотации статей	1000	264	983	150	39
TwiMed-Twitter [44]	Твиттер	637	329	308	42	27
Twitter [42]	Твиттер	645	569	76	37	22

4 Архитектуры нейронных сетей

В данном разделе описаны архитектуры сравниваемых нейронных сетей.

4.1 LSTM

Классическая нейронная сеть, являющаяся разновидностью рекуррентных нейронных сетей, была представлена в [45]. Сеть состоит из трех слоев: входного, слоя с короткой долгосрочной памятью (LSTM) и выходного. В первом слое сети (Embedding) происходит кодирование входного текста в векторное представление и передаются в слой LSTM. Данный слой считывает пословно входное предложение и сохраняет скрытые состояния с помощью. После прочтения всего предложения скрытые состояния передаются в качестве признака в выходной классифицирующий слой с функцией softmax.

4.2 TD_LSTM

Данная модель была предложена в работе [6] и является расширением предыдущей модели. Модель состоит из двух частей, каждая из которых обрабатывает левый и правый контексты соответственно. Аналогично с предыдущей моделью входные тексты попадают в слой векторного представления слов, выходы которого передаются в LSTM слой. Вектора скрытых состояний LSTM слоев для левого и правого контекстов конкатенируются в один вектор. К полученному вектору, так же как и в предыдущей модели, применяется слой с функцией softmax и вычисляется класс с наибольшей вероятностью. Схема архитектуры данной сети представлена на рис. 1.

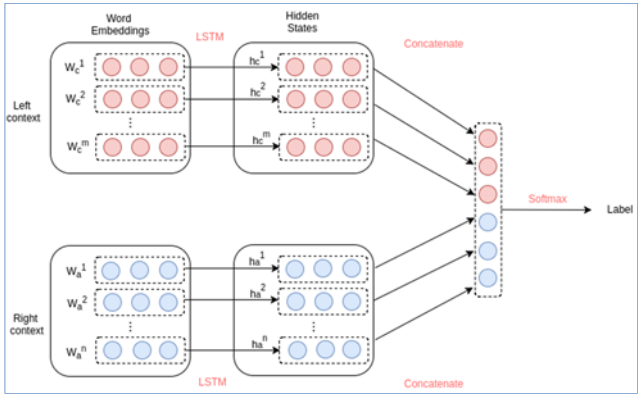

Рис. 1. Архитектура модели TD_LSTM
Fig. 1. The overall architecture of TD LSTM

4.3 IAN

Модель с механизмом интерактивного внимания была представлена в работе [7]. Сеть состоит из двух частей, каждая из которых строит представление контекста и классифицируемой сущности с помощью векторного представления слов и LSTM слоя. Полученные вектора усредняются и используются для вычисления вектора внимания. В первом слое внимания используется вектор контекста и усредненный вектор сущности, во втором - вектор сущности и усредненный вектор контекста. Полученные на выходе вектора конкатенируются и передаются слою с функцией активации softmax для классификации. Схема архитектуры сети представлена на рис.2.

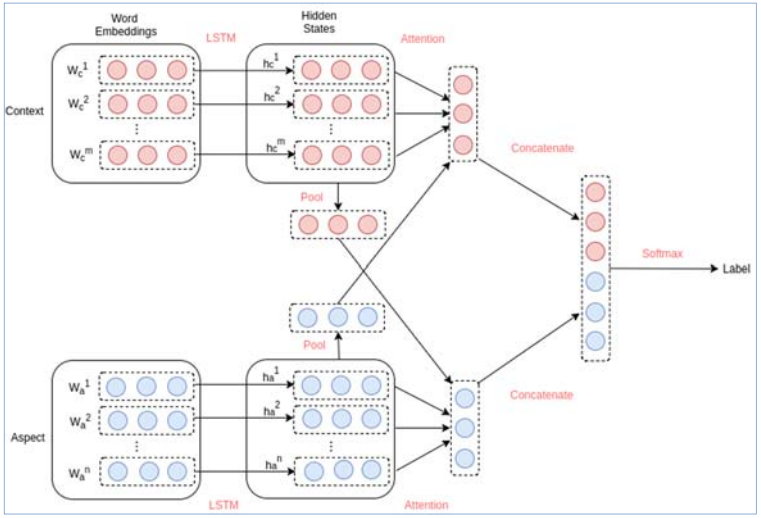

Рис. 2.Архитектура модели IAN
Fig. 2. The overall architecture of IAN

4.4 RAM

Сеть с рекуррентным механизмом внимания к памяти была представлена Ченом с соавторами [9]. Сеть состоит из трех главных частей: первая посвящена обработке контекста с использованием двунаправленной LSTM, полученные вектора сохраняются в память; вторая отвечает за представление классифицируемой сущности и также использует двунаправленную LSTM, на выходе получается среднее значение всех векторов скрытого состояния слов сущности; третья часть применяет механизмы внимания к полученным выходным данным второй части и сохраненным данным первой части. Выходной вектор внимания подается на вход слою с управляемыми рекуррентными блоками (Gated Recurrent Unit; GRU). На следующей итерации на вход слою с вниманием подается выход из GRU и вектора, сохраненные в

памяти. Это позволяет применить механизм внимания к сохраненным в память данным несколько раз и извлечь больше необходимой для классификации информации. Вектор, полученный в результате нескольких подобных итераций, передается в полносвязный слой с классификатором. Архитектура RAM представлена на рис. 3.

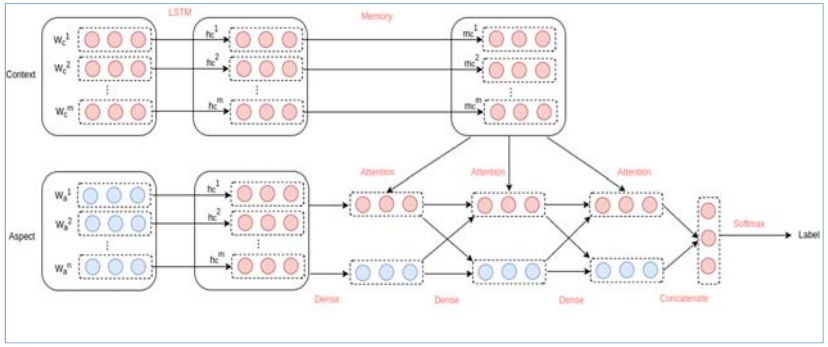

Рис. 3. Архитектура модели RAM
Fig. 3. The overall architecture of RAM

4.5 MemNet

Модель MemNet была представлена Тангом с соавторами [8]. Данная модель состоит из двух главных частей: модуля памяти, который хранит в себе входные данные для контекста в виде распределенного представления слов и механизма внимания. На вход слою с вниманием подаются сущность в виде векторного представления слов и вектора, сохраненные в памяти. Выход из слоя памяти суммируется с векторами памяти и подается в следующий слой с механизмом внимания. Архитектура RAM представлена на рис. 4.

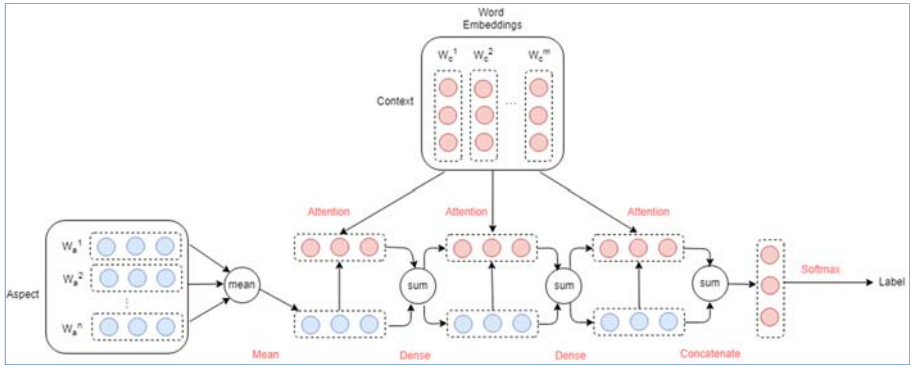

Рис. 4. Архитектура модели MemNet
Fig. 4. The overall architecture of MemNet

5. Эксперименты

В этом разделе мы представляем сравнение эффективности описанных нейронных сетей с моделью на основе метода опорных векторов (Support vector Machine; SVM) с большим набором признаков, чтобы ответить на ключевые вопросы исследования, изложенные во введении.

5.1 Метод на основе SVM

Мы сравнили наши подходы с классификатором предложенном в работе [37]. Данный метод основан на SVM с линейным ядром. Набор экспериментов показал, что признаки на основе униграмм, бинрамм, частей речи, тональности, векторов кластера и семантических типов из словаря UMLS являются наиболее эффективными для классификации побочных эффектов. Признак на основе частей речи состоит и количества существительных, глаголов, наречий и прилагательных. Для тонального признака использовались словари: SentiWordNet [46], MPQA Subjectivity Lexicon [45], Bing Liu's словарь [47]. Признак на основе кластерного представления использовал кластера из [38], полученные с использованием иерархического алгоритма кластеризации Брауна. Последний признак представляет собой количество токенов из каждого семантического типа словаря UMLS.

Оценка эффективности данного метода показала его превосходство в сравнении с предыдущими подходами, основанными на методах машинного обучения и сверточной нейронной сети.

5.2 Параметры моделей

Мы использовали векторное представление слов, обученное на записях из социальных медиа [38]. Векторное представление слов было получено с использованием модели word2vec, обученной на неразмеченном корпусе, состоящем из 2.5 миллиона англоязычных отзывов пользователей о лекарственных препаратах. Длина векторов 200. Статистика покрываемости корпусов словами из модели векторного представления слов: CADEC - 93.5%, Twitter - 80.4%, MADE - 62.5%, TwiMed-Twitter - 81.2%, TwiMed-Pubmed - 76.4%. Для слов, отсутствующих в модели, генерируется вектор случайных чисел с нормальным распределением и значениями, ранжирующимися в рамках значений векторов модели векторного представления слов. Мы использовали 15 эпох для обучения каждой модели на каждом из корпусов, размер входного блока 128 для корпусов CADEC и MADE и 32 для остальных корпусов, количество скрытых состояний 300, шаг обучения (learning rate) 0.01, l2 регуляризация со значением 0.001. В ходе экспериментов модель с данным набором параметров показала наиболее высокий результат. Для реализации модели был использован публично доступный код из репозитория[1].

[1] https://github.com/songyouwei/ABSA-PyTorch

5.3 Результаты

Все модели были оценены на 5-фолдовой кросс валидации с помощью стандартных метрик оценки качества классификации: точность (P), полнота (R) и F-мера – среднее гармоническое между точностью и полнотой. Результаты экспериментов приведены в табл. 2-6. В таблицах класс ‘ADR’ обозначает класс с побочным эффектом, соответственно, класс ‘non-ADR’ обозначает его отсутствие.

Из результатов видно, что на всех корпусах, кроме Twitter лучшие результаты по макро F-мере показала модель IAN. Наиболее значимый прирост качества по сравнению с другими моделями был получен на корпусах Twimed-Twitter и Twitter-Pubmed, где модель IAN достигла 81.9% и 87.4% макро F-меры соответственно. На корпусе Twitter лучшие результаты показала модель RAM с макро F-мерой 83.4%.

Исходя из полученных результатов, можно сделать вывод, что разделение входного предложения на правый и левый контекст относительно выделенной сущности может улучшить качество классификации для корпусов, состоящих из твитов. Это следует из того, что TD_LSTM с результатами макро F-меры 75.8% и 70.3% на корпусах Twitter и Twimed-Twitter соответственно превзошли модель LSTM с результатами 61.3% и 70% макро F-меры. Для остальных корпусов разделение контекста не смогло улучшить результатов. На корпусе Twimed-Pubmed LSTM превзошла модель TD_LSTM на 7% по метрике F-меры. На остальных корпусах результаты сравнимы и отличаются всего на 2%.

Сравнение результатов работы моделей RAM и MemNet показывают, что наличие LSTM слоя перед слоем с памяти оказалось эффективно только на одном корпусе Twitter, где RAM показала существенно высокие результаты F-меры (83.4%) по сравнению с MemNet (76.3%).

Превосходство IAN по сравнению с RAM и MemNet на четырех из пяти корпусов также показывает, что наличие дополнительной памяти далеко не всегда дает преимущество.

Табл. 2. Результаты классификации на корпусе Twitter
Tab. 2. Classification results of the compared methods for Twitter corpus

Модель	Класс non-ADR			Класс ADR			Макро		
	P	R	F	P	R	F	P	R	F
SVM	.602	.520	.554	.602	.520	.554	.769	.736	.749
IAN	.654	.627	.634	.951	.957	.954	.802	.792	.794
RAM	**.779**	.653	**.705**	**.955**	**.973**	**.964**	**.867**	**.813**	**.834**
MemNet	.559	**.667**	.590	.954	.918	.935	.757	.792	.763
TD-LSTM	.606	.547	.570	.940	.952	.946	.773	.749	.758
LSTM	.388	.427	.392	.920	.889	.903	.618	.621	.613

Табл. 3. Результаты классификации на корпусе CADEC
Tab. 3. Classification results of the compared methods for CADEC corpus

Модель	Класс non-ADR			Класс ADR			Макро		
	P	R	F	P	R	F	P	R	F
SVM	.659	.620	.638	.964	.969	.967	.811	.795	.802
IAN	**.699**	**.637**	**.662**	**.966**	.972	**.969**	**.832**	**.805**	**.815**
RAM	.696	.406	.506	.946	**.981**	.963	.821	.694	.734
MemNet	.575	.570	.559	.960	.955	.957	.767	.762	.758
TD-LSTM	.630	.557	.582	.958	.967	.962	.794	.762	.772
LSTM	.664	.554	.602	.958	.973	.966	.811	.764	.784

Табл. 4. Результаты классификации на корпусе MADE
Tab. 4. Classification results of the compared methods for MADE corpus

Модель	Класс non-ADR			Класс ADR			Макро		
	P	R	F	P	R	F	P	R	F
SVM	**.984**	.981	.982	.551	**.582**	.562	.767	**.782**	.772
IAN	.982	**.991**	**.986**	**.740**	.524	**.585**	**.861**	.758	**.786**
RAM	.980	.989	.985	.615	.486	.538	.798	.737	.761
MemNet	.979	**.991**	.985	.684	.447	.535	.832	.719	.760
TD-LSTM	.980	.988	.984	.606	.470	.515	.793	.729	.750
LSTM	.981	.989	.985	.636	.510	.557	.809	.749	.771

Табл. 5. Результаты классификации на корпусе Twimed-Twitter
Tab. 5. Classification results of the compared methods for Twimed-Twitter corpus

Модель	Класс non-ADR			Класс ADR			Макро		
	P	R	F	P	R	F	P	R	F
SVM	.779	.707	.739	**.752**	.810	.778	.766	.758	.758
IAN	**.802**	**.825**	**.813**	**.836**	.813	**.824**	**.819**	**.819**	**.819**
RAM	.799	.736	.764	.773	**.823**	.796	.786	.779	.780
MemNet	.772	.821	.789	.823	.791	.801	.798	.806	.795
TD-LSTM	.731	.711	.717	.741	.751	.742	.736	.731	.730
LSTM	.669	.757	.709	.743	.649	.691	.706	.703	.700

Табл. 6. Результаты классификации на корпусе Twimed-Pubmed
Tab. 6. Classification results of the compared methods for Twimed-Pubmed corpus

Модель	Класс non-ADR			Класс ADR			Макро		
	P	R	F	P	R	F	P	R	F
SVM	.925	.955	0.939	.799	.681	.728	.862	.818	.834
IAN	**.936**	**.977**	**.956**	.878	.738	0.792	**.907**	**.858**	**.874**
RAM	.917	.916	.916	.675	.669	0.662	.796	.792	.789
MemNet	.929	.912	.917	.736	.748	0.705	.833	.830	.811
TD-LSTM	.495	.493	.487	**.932**	**.930**	**0.931**	.714	.712	.709
LSTM	.929	.949	.939	.786	.707	0.740	.858	.828	.839

6 Заключение

В данной статье была исследована применимость общепринятых архитектур нейронных сетей в области аспектно-ориентированного анализа тональности к задаче классификации побочных эффектов. Для оценки эффективности данных моделей были проведены обширные эксперименты на пяти общедоступных текстовых корпусах. Согласно полученным результатам, для четырех из пяти корпусов наилучшие результаты показала модель IAN и на одном корпусе RAM. Также можно сделать вывод, что базовые архитектуры не уступают результатам работы существующего метода на основе SVM, а сети с дополнительной памятью и механизмом внимания превосходят их, что доказывает применимость данных архитектур к задаче классификации побочных эффектов.

В дальнейшем планируются исследования по трем направлениям:

1) оценка влияние параметров предложенных архитектур нейронных сетей на качество классификации;
2) адаптация описанных моделей для классификации на уровне сообщений;
3) применение данных моделей для задачи классификации побочных эффектов на других языках.

7 Благодарности

Работа поддержана грантом РНФ № 18-11-00284.

Список литературы

[1]. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. Journal of Biomedical Informatics, vol. 36, issue ½, 2003, pp. 131–143.

[2]. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S et al. Utilizing social media data for pharmacovigilance: A review. Journal of Biomedical Informatics, vol. 54, 2015, pp. 202–212.
[3]. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N et al. Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review. Journal of Medical Internet Research, vol 17, no 7, 2015.
[4]. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. Drug Safety, vol. 37, 2014, pp. 777–790.
[5]. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical Pharmacology & Therapeutics, vol. 91, no. 6, 2012, pp. 1010–1021.
[6]. Tang D, Qin B, Feng X, Liu T. Effective LSTMs for Target-Dependent Sentiment Classification [Internet]. arXiv [cs.CL], 2015. Available at: http://arxiv.org/abs/1512.01100, accessed 15.11.2008
[7]. Ma D, Li S, Zhang X, Wang H. Interactive attention networks for aspect-level sentiment classification. arXiv preprint arXiv:1709 00893, 2017.
[8]. Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network. arXiv preprint arXiv:1605 08900, 2016;
[9]. Chen P, Sun Z, Bing L, Yang W. Recurrent attention network on memory for aspect sentiment analysis. In Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017, pp. 452–461.
[10]. Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, et al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. Journal of Biomedical Informatics, vol. 44, 2011, pp. 989–996.
[11]. Yang CC, Yang H, Jiang L, Zhang M. Social Media Mining for Drug Safety Signal Detection. In Proc. of the 2012 International Workshop on Smart Health and Wellbeing, 2012. pp. 33–40.
[12]. Liu X, Chen H. AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums / Lecture Notes in Computer Science, vol. 8040, 2013. pp. 134–150.
[13]. Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, Srinivasan R. A pipeline to extract drug-adverse event pairs from multiple data sources. BMC Medical Informatics and Decision Making, vol. 14, no. 13, 2014.
[14]. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, et al. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug Safety, vol. 37, 2014, pp. 343–350.
[15]. O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith KL, Gonzalez G. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. In Proc. of the AMIA Annual Symposium, 2014, pp. 924–933.
[16]. Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. In Proc. of the AMIA Annual Symposium, 2011, pp. 1019–1026.
[17]. Na J-C, Kyaing WYM, Khoo CSG, Foo S, Chang Y-K, Theng Y-L. Sentiment Classification of Drug Reviews Using a Rule-Based Linguistic Approach. Lecture Notes in Computer Science, vol. 7634, 2012. pp. 189–198.
[18]. Yun Niu et al. Analysis of polarity information in medical text. In Proc. of the AMIA Annual Symposium, 2005, pp. 570–574.

[19]. Leaman R. et al. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In Proc. of the 2010 workshop on biomedical natural language processing, 2010, pp. 117-125.
[20]. Yun Niu, Xiaodan Zhu et al. Predicting adverse drug events from personal health messages. In Proc. of the AMIA Annual Symposium, 2011, pp. 217-226.
[21]. Bian J., Topaloglu U., Yu F. Towards large-scale twitter mining for drug-related adverse events. In. Proc. of the 2012 International workshop on smart health and wellbeing, 2012, pp. 25-32.
[22]. Yang M., Wang X., Kiang M. Y. Identification of Consumer Adverse Drug Reaction Messages on Social Media. In Proc. of the Pacific Asia Conference on Information Systems, 2013.
[23]. Sarker A., Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics, vol. 53, 2015, pp. 196-207.
[24]. Aramaki E. et al. Extraction of adverse drug effects from clinical records. Studies in Health Technology and Informatics, vol. 160, №. Pt 1, 2010, pp. 739-743.
[25]. Rastegar-Mojarad M., Elayavilli R.K., Yu Y., Liu H. Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. In Proc. of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, 2016.
[26]. Sarker A, Nikfarjam A, Gonzalez G. Social Media Mining Shared Task Workshop. In Proc. of the Pacific Symposium on Biocomputing, 2016, pp. 581–592.
[27]. Sarker A, Gonzalez-Hernandez G. Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017. In Proc. of the 2nd Social Media Mining for Health Research and Applications Workshop, 2017, pp. 43-48.
[28]. Kiritchenko S, Mohammad SM, Morin J, de Bruijn B. NRC-Canada at SMM4H Shared Task: Classifying Tweets Mentioning Adverse Drug Reactions and Medication Intake. arXiv preprint arXiv:1805 04558. 2018.
[29]. Friedrichs J, Mahata D, Gupta S. InfyNLP at SMM4H Task 2: Stacked Ensemble of Shallow Convolutional Neural Networks for Identifying Personal Medication Intake from Twitter. arXiv preprint arXiv:1803 07718. 2018.
[30]. Huynh T, He Y, Willis A, Rüger S. Adverse drug reaction classification with deep neural networks. In Proc. of the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 877–887.
[31]. Gurulingappa H., Rajput A.M. et al. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics, vol. 45, 2012, pp. 885–892.
[32]. Serrano-Guerrero J., Olivas J.A. et al. Sentiment analysis: A review and comparative analysis of web services. Information Sciences, vol. 311, 2015, pp. 18–38
[33]. Rusnachenko N., Loukachevitch N. Using convolutional neural networks for sentiment attitude extraction from analytical texts. In Proc. of the Third Workshop on Computational linguistics and language science (to be published in CEUR Workshop Proceedings), 2018
[34]. Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting aspects, sentiment and categories of aspects in user reviews about restaurants and cars. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”, issue 14, vol. 2, 2015, pp. 22–34
[35]. Solovyev V., Ivanov V. Dictionary-based problem phrase extraction from user reviews. Lecture Notes in Computer Science, vol. 8655, 2014, pp. 225–232.

[36]. Zhang L., Wang S., Liu, B. Deep learning for sentiment analysis. A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, issue 4, 2018.
[37]. Alimova I., Tutubalina E. Automated detection of adverse drug reactions from social media posts with machine learning. Lecture Notes in Computer, vol. 10716, 2017, pp. 3–15.
[38]. Miftahutdinov Z.S., Tutubalina E.V., Tropsha A.E. Identifying disease-related expressions in reviews using conditional random fields. Computational Linguistics and Intellectual Technologies: Papers from the Annual conference “Dialogue”, issue 16, vol. 1, 2017, pp 155–166
[39]. Korkontzelos I., Nikfarjam A. et al. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. Journal of biomedical informatics, vol. 62, 2016, pp. 148–158.
[40]. Dai H.-J., Touray M., Jonnagaddala J., Syed-Abdul S. Feature engineering for recognizing adverse drug reactions from twitter posts. Information, vol. 7, no. 27, 2016.
[41]. Karimi, S. Metke-Jimenez,A., Kemp M., Wang C.: Cadec. A corpus of adverse drug event annotations. Journal of biomedical informatics, vol. 55, 2015, pp. 73–81.
[42]. Nikfarjam A., Sarker A. et al. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association, vol. 22, no. 3, 2015, pp. 671–681
[43]. Nlp challenges for detecting medication and adverse drug events from electronic health records (made1.0) (2018). University of Massachusetts Lowell, Worcester, Amhers. Available at: https://bio-nlp.org/index.php/projects/39-nlp-challenges, accessed 15.11.2008.
[44]. Alvaro N., Miyao Y., Collier N. Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations. JMIR public health and surveillance, vol. 3, no. 2, 2017.
[45]. Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. In Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 347–
[46]. Baccianella S., Esuli A., Sebastiani F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh conference on International Language Resources and Evaluation, 2010, pp. 2200–2204 (2010)
[47]. Hu M., Liu B. Mining and summarizing customer reviews. In Proc, of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.

Entity-level classification of adverse drug reactions: a comparison of neural network models

I.S. Alimova <alimovailseyar@gmail.com>
E.V. Tutubalina <tutubalinaev@gmail.com >
Kazan Federal University,
18 Kremlyovskaya street, Kazan, 420008, Russian Federation

Abstract. This paper presents our experimental work on neural network models for entity-level adverse drug reaction (ADR) classification. Aspect-level sentiment classification, which aims to determine the sentimental class of a specific aspect conveyed in user opinions, have been actively studied for more than 10 years. In the past few years, several neural network models have been proposed to address this problem. While these models have a lot in common, there are some architecture components that distinguish them from each other. We investigate the applicability of neural network models for ADR classification. We conduct extensive experiments on various pharmacovigilance text sources including biomedical literature, clinical narratives, and social media and compare the performance of five state-of-the-art models as well as a feature-rich SVM in terms of the accuracy of ADR classification.

Keywords: adverse drug reactions; text mining; natural language processing; health social media analytics; machine learning; deep learning

DOI: 10.15514/ISPRAS-2018-30(5)-11

For citation: Alimova I.S., Tutubalina E.V. Entity-level classification of adverse drug reactions: a comparison of neural network models. Trudy ISP RAN/Proc. ISP RAS, vol. 30, issue 5, 2018, pp. 177-196 (in Russian). DOI: 10.15514/ISPRAS-2018-30(5)-11

References

[1]. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. Journal of Biomedical Informatics, vol. 36, issue ½, 2003, pp. 131–143.
[2]. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S et al. Utilizing social media data for pharmacovigilance: A review. Journal of Biomedical Informatics, vol. 54, 2015, pp. 202–212.
[3]. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N et al. Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review. Journal of Medical Internet Research, vol 17, no 7, 2015.
[4]. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. Drug Safety, vol. 37, 2014, pp. 777–790.
[5]. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical Pharmacology & Therapeutics, vol. 91, no. 6, 2012, pp. 1010–1021.
[6]. Tang D, Qin B, Feng X, Liu T. Effective LSTMs for Target-Dependent Sentiment Classification [Internet]. arXiv [cs.CL], 2015. Available at: http://arxiv.org/abs/1512.01100, accessed 15.11.2008

[7]. Ma D, Li S, Zhang X, Wang H. Interactive attention networks for aspect-level sentiment classification. arXiv preprint arXiv:1709 00893, 2017.
[8]. Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network. arXiv preprint arXiv:1605 08900, 2016;
[9]. Chen P, Sun Z, Bing L, Yang W. Recurrent attention network on memory for aspect sentiment analysis. In Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017, pp. 452–461.
[10]. Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, et al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. Journal of Biomedical Informatics, vol. 44, 2011, pp. 989–996.
[11]. Yang CC, Yang H, Jiang L, Zhang M. Social Media Mining for Drug Safety Signal Detection. In Proc. of the 2012 International Workshop on Smart Health and Wellbeing, 2012. pp. 33–40.
[12]. Liu X, Chen H. AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums / Lecture Notes in Computer Science, vol. 8040, 2013. pp. 134–150.
[13]. Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, Srinivasan R. A pipeline to extract drug-adverse event pairs from multiple data sources. BMC Medical Informatics and Decision Making, vol. 14, no. 13, 2014.
[14]. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, et al. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug Safety, vol. 37, 2014, pp. 343–350.
[15]. O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith KL, Gonzalez G. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. In Proc. of the AMIA Annual Symposium, 2014, pp. 924–933.
[16]. Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. In Proc. of the AMIA Annual Symposium, 2011, pp. 1019–1026.
[17]. Na J-C, Kyaing WYM, Khoo CSG, Foo S, Chang Y-K, Theng Y-L. Sentiment Classification of Drug Reviews Using a Rule-Based Linguistic Approach. Lecture Notes in Computer Science, vol. 7634, 2012. pp. 189–198.
[18]. Yun Niu et al. Analysis of polarity information in medical text. In Proc. of the AMIA Annual Symposium, 2005, pp. 570–574.
[19]. Leaman R. et al. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In Proc. of the 2010 workshop on biomedical natural language processing, 2010, pp. 117-125.
[20]. Yun Niu, Xiaodan Zhu et al. Predicting adverse drug events from personal health messages. In Proc. of the AMIA Annual Symposium, 2011, pp. 217-226.
[21]. Bian J., Topaloglu U., Yu F. Towards large-scale twitter mining for drug-related adverse events. In. Proc. of the 2012 International workshop on smart health and wellbeing, 2012, pp. 25-32.
[22]. Yang M., Wang X., Kiang M. Y. Identification of Consumer Adverse Drug Reaction Messages on Social Media. In Proc. of the Pacific Asia Conference on Information Systems, 2013.
[23]. Sarker A., Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics, vol. 53, 2015, pp. 196-207.

[24]. Aramaki E. et al. Extraction of adverse drug effects from clinical records. Studies in Health Technology and Informatics, vol. 160, №. Pt 1, 2010, pp. 739-743.
[25]. Rastegar-Mojarad M., Elayavilli R.K., Yu Y., Liu H. Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. In Proc. of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, 2016.
[26]. Sarker A, Nikfarjam A, Gonzalez G. Social Media Mining Shared Task Workshop. In Proc. of the Pacific Symposium on Biocomputing, 2016, pp. 581–592.
[27]. Sarker A, Gonzalez-Hernandez G. Overview of the Second Social Media Mining for Health (SMM4H) Shared Tasks at AMIA 2017. In Proc. of the 2nd Social Media Mining for Health Research and Applications Workshop, 2017, pp. 43-48.
[28]. Kiritchenko S, Mohammad SM, Morin J, de Bruijn B. NRC-Canada at SMM4H Shared Task: Classifying Tweets Mentioning Adverse Drug Reactions and Medication Intake. arXiv preprint arXiv:1805 04558. 2018.
[29]. Friedrichs J, Mahata D, Gupta S. InfyNLP at SMM4H Task 2: Stacked Ensemble of Shallow Convolutional Neural Networks for Identifying Personal Medication Intake from Twitter. arXiv preprint arXiv:1803 07718. 2018.
[30]. Huynh T, He Y, Willis A, Rüger S. Adverse drug reaction classification with deep neural networks. In Proc. of the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 877–887.
[31]. Gurulingappa H., Rajput A.M. et al. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics, vol. 45, 2012, pp. 885–892.
[32]. Serrano-Guerrero J., Olivas J.A. et al. Sentiment analysis: A review and comparative analysis of web services. Information Sciences, vol. 311, 2015, pp. 18–38
[33]. Rusnachenko N., Loukachevitch N. Using convolutional neural networks for sentiment attitude extraction from analytical texts. In Proc. of the Third Workshop on Computational linguistics and language science (to be published in CEUR Workshop Proceedings), 2018
[34]. Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting aspects, sentiment and categories of aspects in user reviews about restaurants and cars. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”, issue 14, vol. 2, 2015, pp. 22–34
[35]. Solovyev V., Ivanov V. Dictionary-based problem phrase extraction from user reviews. Lecture Notes in Computer Science, vol. 8655, 2014, pp. 225–232.
[36]. Zhang L., Wang S., Liu, B. Deep learning for sentiment analysis. A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, issue 4, 2018.
[37]. Alimova I., Tutubalina E. Automated detection of adverse drug reactions from social media posts with machine learning. Lecture Notes in Computer, vol. 10716, 2017, pp. 3–15.
[38]. Miftahutdinov Z.S., Tutubalina E.V., Tropsha A.E. Identifying disease-related expressions in reviews using conditional random fields. Computational Linguistics and Intellectual Technologies: Papers from the Annual conference “Dialogue”, issue 16, vol. 1, 2017, pp 155–166
[39]. Korkontzelos I., Nikfarjam A. et al. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. Journal of biomedical informatics, vol. 62, 2016, pp. 148–158.
[40]. Dai H.-J., Touray M., Jonnagaddala J., Syed-Abdul S. Feature engineering for recognizing adverse drug reactions from twitter posts. Information, vol. 7, no. 27, 2016.

[41]. Karimi, S. Metke-Jimenez,A., Kemp M., Wang C.: Cadec. A corpus of adverse drug event annotations. Journal of biomedical informatics, vol. 55, 2015, pp. 73–81.
[42]. Nikfarjam A., Sarker A. et al. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association, vol. 22, no. 3, 2015, pp. 671–681
[43]. Nlp challenges for detecting medication and adverse drug events from electronic health records (made1.0) (2018). University of Massachusetts Lowell, Worcester, Amhers. Available at: https://bio-nlp.org/index.php/projects/39-nlp-challenges, accessed 15.11.2008.
[44]. Alvaro N., Miyao Y., Collier N. Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations. JMIR public health and surveillance, vol. 3, no. 2, 2017.
[45]. Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. In Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 347–
[46]. Baccianella S., Esuli A., Sebastiani F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh conference on International Language Resources and Evaluation, 2010, pp. 2200–2204 (2010)
[47]. Hu M., Liu B. Mining and summarizing customer reviews. In Proc, of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.