ОНЛАЙН-ИССЛЕДОВАНИЯ

Е.Л. Артемова, А.А. Максименко, Д.А. Охрименко
(*Москва*)

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ КОНТЕНТА КОРРУПЦИОННОЙ ТЕМАТИКИ В РУССКОЯЗЫЧНЫХ И АНГЛОЯЗЫЧНЫХ ИНТЕРНЕТ-СМИ[1]

В статье предпринята попытка классификации коррупционного медиаконтента русскоязычных и англоязычных интернет-СМИ с помощью методов машинного обучения. Данный методологический аспект является весьма актуальным и перспективным, поскольку, согласно полученным нами ранее данным, используемые в зарубежных публикациях механизмы коррупционного мониторинга, основанные на использовании передовых информационных технологий, обладают неоднозначной потенциальной эффективностью и не всегда адекватно интерпретируются. В работе показаны принципы и основания для выделения идентификационных параметров, а также подробно описана схема разметки собранного новостного массива. В ходе автоматической обработки текстов, проходившей в два этапа (векторизация текста и использование модели обучения), удалось решить 4 основные задачи: выделение значимой цитаты из новостной статьи для идентификации текста коррупцион-

Екатерина Леонидовна Артемова – кандидат технических наук, доцент факультета компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: elartemova@hse.ru
Александр Александрович Максименко – доктор социологических наук, кандидат психологических наук, доцент, эксперт проектно-учебной лаборатории антикоррупционной политики, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: Maximenko.Al@gmail.com
Дмитрий Андреевич Охрименко – студент факультета компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: ohrimenko@hse.ru

ной тематики; предсказание типа новостного сообщения; предсказание статьи УК РФ, по которой определяется ответственность за описанное коррупционное правонарушение, а также предсказание типа взаимоотношений в коррупционных правонарушениях. Полученные результаты продемонстрировали, что современные методы автоматической обработки текстов успешно справляются с идентификацией и классификацией коррупционного контента как на русском, так и на английском языках.

Ключевые слова: коррупционный контент; коррупционные правонарушения; машинное обучение; автоматическая обработка текстов; интернет-СМИ; русскоязычные медиа; англоязычные медиа

DOI: 10.19181/4m.2021.52.5

Введение

Ранее нами на основе ряда зарубежных эмпирических исследований по использованию искусственного интеллекта и машинного обучения в вопросах выявления и противодействия коррупции было показано, что освещаемые в зарубежных источниках механизмы коррупционного мониторинга, основанные на использовании передовых информационных технологий и алгоритмов, обладают разной потенциальной эффективностью и не всегда релевантно интерпретируются коллегами [1]. Не ставя под сомнение перспективность этого направления, попробуем оценить возможности машинного обучения в классификации новостного контента на примере англоязычных и русскоязычных текстов, содержащих коррупционную тематику, публикуемых интернет-СМИ.

Современные методы автоматической обработки текстов, основанные на моделях машинного обучения, успешно применяются в прикладных исследованиях. К наиболее востребованным прикладным областям можно отнести биомедицинское направ-

ление [2; 3; 4; 5], в рамках которого методы обработки текстов применяют как для анализа научных публикаций [6; 7; 8], историй болезни [9; 10; 11], так и для анализа социальных медиа [12; 13; 14]. Другим востребованным направлением считается поддержка юридической деятельности (LegalTech) [15; 16; 17]. Естественно, особое внимание уделяется анализу социальных медиа и интернет-СМИ, отражающим общественное мнение [18; 19]. Так, например, недавняя серия работ посвящена анализу общественного отношения к разного рода ограничениям, вызванным пандемией коронавируса [20; 21; 22]. Другие темы, привлекающие интересы исследователей социальных сетей, – это отношение пользователей к феминизму [23; 24; 25], проблемам экологии [26; 27; 28], психологическим проблемам [29; 30; 31]. Отметим, что большинство подобных исследований выполняются в англоязычной среде.

Методы обработки текстов, используемые в прикладных исследованиях, включают в себя: 1) методы классификации по теме, позволяющие обнаружить тексты, посвященные определенной тематике [32]; 2) методы, направленные на извлечение специфических терминов, упоминаний важных для предметной области событий, деятелей и др. [33], 3) методы информационного поиска, с помощью которых обнаруживаются релевантные вопросам исследования тексты [34]; 4) смежные с методами информационного поиска технологии вопросно-ответных систем, позволяющие находить важные фрагменты в рассматриваемом тексте [35], 5) методы описательного исследования текстов, в том числе тематическое моделирование [36]. Перечисленные методы опираются на различные модели машинного обучения и на различные формальные постановки задач. Так, например, классификация текстов предполагает обучение классификаторов по размеченным текстам, извлечение именованных сущностей – обучение моделей разметки последовательности. В свою очередь, тематическое моделирование основано на методах обучения без учителя. Любые методы, использующие обучение с учителем, предполагают разметку

текстовых данных: каждому тексту (или слову, фрагменту текста) приписывают метки класса. В дальнейшем модель классификации обучается по входному тексту предсказывать метки класса. Методы, использующие обучение без учителя, не предполагают дополнительной разметки текстовых данных и могут быть использованы для обработки сырого массива текстов.

Популярные на текущий момент методы анализа данных – особенно неструктурированных данных, таких как текстовые данные, – предполагают в основном использование нейронных сетей и методов глубокого обучения (deep learning) [37; 38; 39; 40; 41; 42]. Нейросетевые модели успешно справляются как с задачами кластеризации, так и с задачами классификации текстов. Среди методов анализа данных, применимых исключительно к обработке текстов, доминирует подход, предполагающий использование предобученных нейросетевых языковых моделей [43]. Такие языковые модели обучаются на большом неразмеченном объеме текстов, получая общие знания о языке и лингвистических феноменах. В последующем языковую модель дообучают для решения конкретной задачи. Отметим, что такой подход обладает рядом существенных преимуществ по сравнению с предшествовавшими подходами: не только качество решения целевой задачи становится выше, но и размеченных данных для получения высокого уровня качества требуется меньше [42; 43].

Данная статья посвящена анализу новостных сообщений. В ходе работы был создан и проанализирован массив из 3000 новостных сообщений, посвященных коррупционным правонарушениям, на английском и на русском языках. Собранные тексты были размечены по четырем параметрам: 1) выделена цитата, на основании которой можно сделать вывод о том, что текст посвящен коррупционным правонарушениям, 2) отмечен тип новостного сообщения (журналистское расследование или новость о судебном деле), 3) обозначена статья Уголовного кодекса Российской Федерации, по которой определяется ответственность за описанное коррупп-

ционное правонарушение, 4) обозначен тип взаимоотношений, описываемых в тексте коррупционных правонарушений. Задача предсказания каждого из параметров разметки может быть сформулирована формально как задача классификации. Были использованы методы машинного обучения для предсказания каждого из параметров разметки.

Цель исследования состояла в апробации метода предобученных нейросетевых моделей с использованием алгоритмов, позволяющих расширить объем обучающих данных, увеличив таким образом в дальнейшем способность выявлять типичные коррупционные практики в новостном контенте.

Актуальность данной проблемы состоит в преодолении сложностей, связанных с нехваткой размеченных данных для более системного и комплексного анализа проявлений коррупции в российском и западном обществах. Кроме того, антикоррупционная тематика отличается крайним разнообразием и акцентами в установках журналистов, ведущих как собственные расследования, так и освещающих данную тему в интересах граждан и государства.

В настоящей статье обсуждается подход к сбору и разметке текстов, приводится формальная постановка задач и подходы к их решению. Кроме того, представлены результаты исследования и их обсуждение. В заключение оговорены ограничения и допущения относительно использованных в настоящей статье методов, а также очерчены направления будущих исследований.

Материалы исследования

Реализованный авторами проект предусматривал поиск не менее 2000 новостей по теме коррупции в русскоязычных (1 160) и англоязычных (1 170) СМИ по заданным ключевым словам за последние 5 лет, их разметку и последующий анализ методами автоматической обработки текстов. На основе размеченного массива происходило обучение классификационных моделей соотнесению обнаруженных новостей с принятой схемой разметки.

Алгоритм пошагово может быть описан следующим образом: новостные тексты подвергались простой стандартной предобработке (удаление html-разметки, разбиение текстов на слова и предложения); все множество текстов было разделено на обучающую и тестовую части; на обучающих текстах дообучалась нейросетевая языковая модель. Качество итоговой модели определялось на тестовых данных.

Основная проектная идея состояла в ознакомлении с алгоритмами работы машинного обучения по классификации новостного контента по теме коррупции в русскоязычных и англоязычных СМИ. В ходе проекта решалась проблема классификации новостного контента по теме коррупции с последующей разметкой собранного медийного массива, удобного для машинного обучения.

В рамках проектной работы 10 студентов НИУ «Высшая школа экономики»[1] производили поиск коррупционного контента, опубликованного в следующих высокорейтинговых интернет-СМИ с 2010 года (см. табл. 1).

Для поиска в интернет-СМИ статей соответствующей тематики студентами использовались следующие ключевые слова: коррупция, взятка, получение взятки, дача взятки, коррупционное правонарушение, противодействие коррупции, злоупотребление служебным положением. В связи с этим выборку можно охарактеризовать как простую случайную.

При выборе площадок интернет-СМИ авторы руководствовались возможностью охвата средств массовой информации широкого спектра тем и мнений общественно-политического поля (освещающих происходящие события как с провластной, так и с оппозиционной точек зрения).

[1] Авторы выражают особую признательность за помощь в сборе материала в русскоязычных и англоязычных интернет-СМИ следующим студентам: А.А. Абакутиной, А.К. Аксёнову, Г.Д. Александрову, Ф.О. Васильеву, А.Г. Гарнову, В.В. Дубковской, М.С. Козловскому, С.Ф. Лазаревой, Е.А. Лымарь, А.А. Мурач, Н.Н. Непочатову, А.В. Печеному, В. Саве, О.А. Сарычевой, М.Ю. Сапрыкину, П.А. Севериновой, Ю.А. Смирновой, И.Д. Соловьеву, В.А. Шатылович.

Таблица 1

ИСТОЧНИКИ ДАННЫХ ДЛЯ СБОРА НОВОСТНЫХ СООБЩЕНИЙ

Язык / Критерии проекта	Русскоязычные СМИ	Англоязычные СМИ
Новостные интернет-СМИ (топ-10 ведущих)	RBC.ru Meduza.io gazeta.ru lenta.ru fontanka.ru kp.ru ria.ru regnum.ru m24.ru life.ru aif.ru tvrain.ru	Bbc.com Washingtonpost.com Theguardian.com Cnn.com Globo.com Lefigaro.fr Foxnews.com Bloomberg.com Reddit.com Dailymail.co.uk Telegraph.co.uk Dw.com

Обнаружив подобную новость, студенты в отдельную таблицу выносили заголовок новости, новостную цитату, по которой можно было понять, что новость носит коррупционный характер, а также самостоятельно определяли тип правонарушения (с последующей визуальной перепроверкой: 1) бытовая коррупция – взаимоотношения между физическим лицом и чиновником; 2) институциональная коррупция – взаимоотношения между предпринимателем и чиновником; 3) корпоративная коррупция – взаимоотношения между физическим лицом и организацией; 4) бюрократическая коррупция – взаимоотношения между чиновниками, а также вид правонарушения, сопоставляя соответствующую статью УК РФ с видом описанного представителями СМИ того или иного правонарушения по ранее определенной схеме наказаний за коррупционные деяния (статья 159 «Мошенничество»; статья 160 «Присвоение или растрата»; статья 204 «Коммерческий подкуп»; статья 285 «Злоупотребление должностными полномочиями»; статья 285.1 «Нецеле-

вое расходование бюджетных средств»; статья 285.3 «Внесение в единые государственные реестры заведомо недостоверных сведений»; статья 286 «Превышение должностных полномочий»; статья 289 «Незаконное участие в предпринимательской деятельности»; статья 290 «Получение взятки»; статья 291 «Дача взятки»; статья 291.1 «Посредничество о взяточничестве»; статья 292 «Служебный подлог»; статья 304 «Провокация взятки либо коммерческого подкупа»; статья 309 «Подкуп или принуждение к даче показаний или уклонению от дачи показаний либо к неправильному переводу») (см. табл. 2).

Разметка осуществлялась по аналогии с задачей поиска документа по запросу (найти статью или фрагмент, соответствующий определенной тематике). Задача поиска по запросу и определения релевантного документа не очень сложна для разметчиков, но при этом позволяет получить данные высокого качества: найденные документы действительно будут обладать высокой релевантностью к заданному запросу. Временны́е затраты на такую разметку тоже невысоки: каждый источник данных может обработать один разметчик. Отметим, однако, что для оценки надежности такой разметки неприменимы стандартные коэффициенты надежности (Каппа Коэна, Альфа Кронбаха). Использование этих коэффициентов предполагает, что: а) существует фиксированное количество документов, б) релевантность каждого документа определяют как минимум два разметчика. В предложенном подходе к разметке выполнение этих требований является чрезмерно трудозатратным.

Рис. 1–3 и табл. 3 представляют описательные статистики составленного набора текстов. Рис. 1 демонстрирует абсолютное число новостных сообщений, посвященных судебным делам или являющихся журналистскими расследованиями. В большей части собранных новостных сообщений обсуждаются новые или завершенные судебные дела, при этом доля журналистских расследований существенно ниже.

Таблица 2

ПРИМЕРЫ РАЗМЕЧЕННЫХ НОВОСТНЫХ СООБЩЕНИЙ, СОДЕРЖАЩИХ КОРРУПЦИОННЫЙ КОНТЕНТ

Заголовок новости	Новостная цитата, позволяющая идентифицировать коррупционный контент	ЖР/СД (предъявление обвинения или вынесение судебного решения)	Вид правонарушения	Тип взаимоотношений
Мэр камчатского города отстранен от должности по обвинению в коррупции	Чиновнику вменяется мошенничество и превышение служебных полномочий. В 2008 г. Иванов обманом вынудил местного жителя перевести на лицевой счет поселка более 100 тыс. руб., которыми затем распорядился по собственному усмотрению	СД	ст. 159 «Мошенничество»	Между физическим лицом и чиновником (бытовая коррупция)
В Москве за взятку в 4 млн руб. арестован подполковник милиции	В Москве при получении взятки в размере 4 млн руб. задержан подполковник милиции, заместитель начальника службы тыла МВД Татарстана. Возбуждено уголовное дело по статье «получение взятки»	СД	ст. 290 «Получение взятки»	Между предпринимателем и чиновником (институциональная коррупция)

Примечание. ЖР – журналистское расследование; СД – судебное дело.

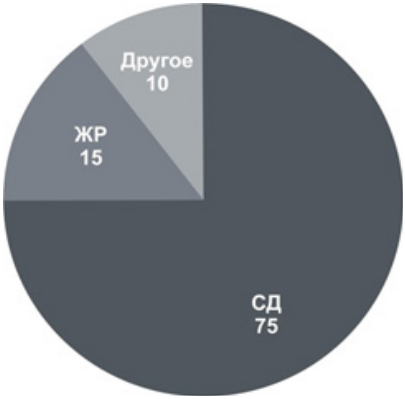

Рис. 1. Доля новостных сообщений, посвященных судебным делам или журналистским расследованиям (*в %*)

Примечание. ЖР – журналистское расследование, СД – судебное дело.

Рис. 2 показывает количество новостных сообщений, описывающих различные виды правонарушений в соответствии с УК РФ. Чаще всего появляются новости о делах по ст. 290 «Получение взятки» (порядка 500 упоминаний), остальные статьи УК РФ упоминаются существенно реже. Так, второй и третьей по частоте являются ст. 159 «Мошенничество» и ст. 291 «Дача взятки» – упомянуты порядка 60 раз каждая.

Рис. 3 отражает характер коррупционного взаимоотношения, обсуждаемого в новостной статье. Чаще всего в новостных статьях описывают случаи институциональной коррупции (более 350 случаев), существенно реже – бытовой коррупции (порядка 180 случаев). Оставшиеся из выделенных нами типов правонарушений встречаются реже (около 50 случаев каждый). Примерно в десятой доли (около 100 случаев) собранных новостных сообщений тип коррупционного взаимодействия не удалось установить.

Табл. 3 представляет описательные статистики, посчитанные по собранным коллекциям новостных сообщений на русском и английском языках. Обе коллекции имеют сопоставимые размеры (1160 и 1170 сообщений). Средние длины заголовков на обоих языках составляют примерно 10 слов. В среднем в собранных кол-

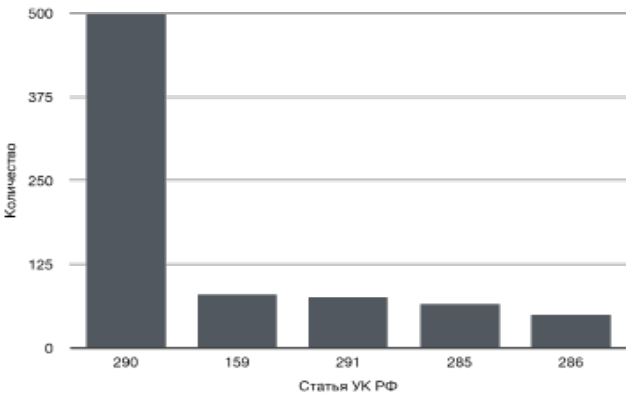

Рис. 2. Количество упоминаний различных видов правонарушений в соответствии со статьями УК РФ (5 самых частых статей УК РФ), по данным российских интернет-СМИ

Примечание. Ст. 290 – «Получение взятки», ст. 159 – «Мошенничество», ст. 291 – «Дача взятки», ст. 285 – «Злоупотребление должностными полномочиями», ст. 286 – «Превышение должностных полномочий».

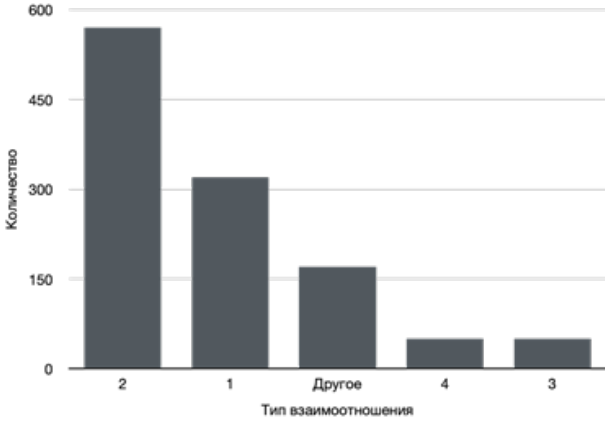

Рис. 3. Количество новостных статей, описывающих различные типы коррупционных взаимоотношений

Примечание. 1 – бытовая коррупция; 2 – институциональная коррупция; 3 – корпоративная коррупция; 4 – бюрократическая коррупция.

лекциях цитаты, позволяющие идентифицировать коррупционный контент, на русском языке несколько короче, чем на английском. Часто встречающиеся в выделенных цитатах слова на русском и на английском языках также похожи: взятка (bribe), задержание (arrested), суд (court) и др.

Таблица 3

ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ, СОСТАВЛЕННЫЕ ПО ДВУМ КОЛЛЕКЦИЯМ НОВОСТНЫХ СООБЩЕНИЙ

Описательные статистики	Новостные сообщения на русском языке	Новостные сообщения на английском языке
Количество текстов	1160	1170
Средняя длина заголовка (*в словах*)	9,92 ± 3,4	10,44 ±4,55
Средняя длина цитаты (*в словах*)	27,4 ±14,6	33,16 ±20.58
Топ-25 слов в цитатах (без стоп-слов)	Взятка, получение, РФ, УК, статья, размер, коррупция, крупный, получать, область, следствие, особо, дело, задержание, суд, доллары, управление, бывший, компания, сумма, МВД, начальник, район, полиция, рубли	Corruption, former, million, charges, scheme, president, money, state, officials, accused, court, federal, bribe, government, prosecutors, company, guilty, arrested, years, investigation, minister, exchange, charged, allegations, laundering

Методы исследования

Современный подход к автоматической обработке текстов предполагает выполнение нескольких этапов: 1) векторизацию текста – т.е. представление текста в виде вектора, формального математического объекта, который служит входом для следующего

этапа; 2) использование модели машинного обучения для решения целевой задачи. Представленная в 2018 г. предобученная языковая модель BERT и ее потомки позволяют совместить оба этапа [43]. Отметим, что популярность и эффективность подобного подхода принято объяснять двумя факторами: использованием современных нейросетевых архитектур, позволяющих достаточно быстро обрабатывать большие объемы текстовых данных, и эффективным предобучением на большом объеме текстовых данных, за счет которого нейросетевая модель учится понимать различные лингвистические феномены.

Пусть дано новостное сообщение t, состоящее из k слов. Модель векторизации представляет входной текст t в виде плотного вещественного вектора $x = (x_1, \ldots, x_n)$ фиксированной размерности n. Вектор x называют векторным представлением текста и используют в качестве входа в некоторую модель машинного обучения. Задача классификации формулируется как предсказание вероятности класса c из заданной системы классов $c \in C$ для данного текста x. Другими словами, требуется оценить $p(x,c)$.

В нашем случае в качестве модели векторизации выступают нейросетевые предобученные модели: адаптированная к русскому языковая модель ruBERT [37], использующая блоки Трансформер [38]. Оценка искомых вероятностей получается с использованием полносвязных слоев и нормализующей функции softmax, которая преобразует выходы полносвязных слоев в векторы значений, интерпретируемых как вероятности.

В данной работе были поставлены четыре задачи, которые могут быть сформулированы как задачи классификации:

1) выделение значимой цитаты из новостной статьи, на основании которой можно сделать вывод о том, что текст посвящен коррупционным правонарушениям (при этом задача определения того, является ли текст коррупционного содержания, по ключевой (значимой) цитате не ставилась);

2) предсказание типа новостного сообщения (журналистское расследование или новость о судебном деле);

3) предсказание статьи УК РФ, по которой определяется ответственность за описанное коррупционное правонарушение;

4) предсказание типа взаимоотношений в коррупционных правонарушениях.

Задачи 1 и 2 являются задачами бинарной классификации. Решение задачи 1 устроено следующим образом: мы последовательно перебираем все предложения из данного текста. Предложениям, входящим в значимую цитату, мы ставим, соответственно, «+», остальным предложениям – «-». Классификатор обучается предсказывать для входного предложения вероятность класса «+». Предлагаемое решение построено по аналогии с подходами к экстрактивному реферированию с обучением [39]. Для решения задачи 2 классификатор обучается предсказывать один из двух классов (журналистское расследование или новость о судебном деле) по двум текстовым фрагментам, объединяемым в один: по заголовку новости и по значимой цитате. Аналогичным образом мы подходим к решению задач 3 и 4, в которых классов больше 2. Отметим, что с технической точки зрения количество классов не имеет значения.

Из рис. 1 видно, что данные не сбалансированы: распределение числа текстов по разным классам далеко от равномерного. Мы решили использовать методы расширения обучающего набора данных (data augmentation) для того, чтобы компенсировать дисбаланс классов. Таким образом, процедура обработки новостных сообщений состояла из следующих шагов: шаг 1 – разбиение текстов на предложения, шаг 2 – векторизация каждого предложения с помощью модели ruBERT, шаг 3 – порождение псевдовекторов предложений для расширения обучающего множества. С этой целью мы использовали алгоритм, построенный по аналогии со SMOTE [40], чтобы расширить объем обучающих данных для задач 2–4. Этот алгоритм создает вектор, относящийся к данному классу и удовлетворяющий следующим условиям: ближайшие к нему вектора предложений принадлежат к этому классу и нет ни

одного близкого вектора предложений из другого класса. Заключительный шаг 4 предполагает обучение модели машинного обучения на основе логистической регрессии. Модели машинного обучения, используемые для рассматриваемых задач, обучались независимо друг от друга, то есть потребовалось обучить четыре экземпляра модели.

Результаты исследования

Полученные результаты (табл. 4) в целом доказывают, что современные методы автоматической обработки текстов успешно справляются с идентификацией и классификацией коррупционного контента как на русском, так и на английском языках. В качестве метрики качества в экспериментах использована доля правильных ответов (accuracy).

Таблица 4

ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Доля правильных ответов классификации предложений	Задача 1 (цитата)	Задача 2 (тип новостного сообщения)	Задача 3 (статья УК РФ)	Задача 4 (тип взаимоотношения)
Новости на русском языке (n = 1160)				
Независимые модели	0,94	0,73	0,62	0,64
+ расширение обучающих данных		0,96	0,82	0,84
Новости на английском языке (n = 1170)				
Независимые модели	0,96	0,69	0,63	0,64
+ расширение обучающих данных		0,89	0,77	0,78

В первую очередь отметим, что задача выделения значимой цитаты, позволяющей идентифицировать коррупционный контент, решается на высоком уровне качества (94 и 96% правильных ответов для русского и английского языков соответственно). Мы видим несколько причин достижения таких высоких показателей качества. Во-первых, задача выделения цитаты, характеризующей коррупционный контент, представляется относительно несложной. Так, судя по частотным словам, выделенным из цитат (см. табл. 3), существует вполне устойчивый набор лексических маркеров, на основании которого классификатор легко определяет нужные предложения. Во-вторых, мы рассматриваем упрощенную постановку задачи: модель классификации должна отличить незначимые предложения от значимых. Естественно, число незначимых предложений существенно выше числа значимых. Число правильно определенных незначимых предложений имеет существенный вклад в показатель качества. При этом мы считаем, что достигнутый уровень определения значимой цитаты уже позволяет использовать в дальнейшем обученную модель на практике для анализа больших объемов новостных сообщений, поскольку доля правильных ответов модели машинного обучения выше 90%. Три другие задачи представляются более сложными. Тем не менее качество их решения значительным образом превосходит пессимистичный сценарий случайного угадывания. Первые очевидные подходы к решению задач 2–4 предоставляют порядка 60–70% правильных ответов для обоих языков. За счет расширения набора обучающих данных синтетическими примерами удается достичь лучших результатов: тип новостного сообщения (задача 2) определяется верно в 96 и 89% случаев для русского и английского языков соответственно. Тип взаимоотношения (задача 4) верно определен в 84% случаях по текстам на русском и в 78% – по текстам на английском языке. Заметим, что даже сложная в силу большого числа классов задача 3 определения вида правонарушения также решается на приемлемом уровне (82% правильных ответов для русского и 77% для

английского языка). Разница между показателями качества для русского и английского языков может быть объяснена бóльшим тематическим разнообразием текстов на английском языке и более явными стилистическими различиями источников.

Перспективные направления исследований

Полученные результаты вселяют надежду, что расширение обучающего набора данных и использование дополнительных приемов машинного обучения позволят в дальнейшем улучшить качество решения всех четырех задач. Перечислим некоторые направления развития исследований, которые могут привести к повышению качества. Во-первых, задачи 1, 2 и 4 позволяют использовать одновременно данные на английском и на русском языках. Обученная на данных обоих языков мультиязыковая модель может предоставлять более высокие показатели качества. В этом состоит кумулятивный полилингвистический и кросс-культурный эффект анализа данной проблематики. Во-вторых, вне зависимости от языка задачи 1–4 могут быть решены одновременно с использованием общей модели векторизации с независимыми полносвязными слоями для решения конкретной задачи. В этом случае более эффективное использование данных позволит сократить время на обучение модели и даст возможность подобрать оптимальные параметры обучения. В-третьих, текущие результаты показывают важность эффективных алгоритмов расширения обучающих данных. Использование альтернативных методов (в том числе на основе генерирующих моделей [41]) кажется перспективным направлением для повышения качества.

Идентификация и классификация коррупционного контента в интернет-СМИ ставит еще одну актуальную перспективную задачу по изучению особенностей работы журналистов в сфере освещения коррупционных правонарушений и проведения журналистских расследований. Анализируемые нами новости выделенного

пула интернет-СМИ сводились к поиску коррупционного контента наиболее известных федеральных изданий, имеющих возможность обращаться к данной теме как в жанре инициативных журналистских расследований, так и в формате оповещения населения об оперативно-розыскной деятельности соответствующих ведомств в изучаемом нами направлении. При этом работа журналистов в региональных интернет-СМИ осталась вне фокуса нашего анализа. Нужно признать это существенным ограничением нашей выборки и в то же время перспективной задачей дальнейших исследований. Производство медиаконтента коррупционной направленности, расположенное в плоскости сочетания профессиональной этики и экономических интересов различных стейкхолдеров, способно наряду с общественно значимыми и социально полезными задачами создавать имитационные практики широкого спектра симуляций, предпринимать определенные маневры в ходе конкурентной борьбы за рынки рентных и финансовых потоков. В связи с этим анализ среды интернет-СМИ необходимо сопровождать как поиском новых маркеров коррупционных правонарушений с апробацией работы построенных на основе выделенных параметров новых моделей обучения, так и с имплементацией этических стандартов в работе журналистов, разработкой ресурсов, позволяющих не только определять степень фейковости медиаконтента, но и оценивать тенденциозность и конечных бенефициаров используемых в СМИ материалов.

Полученные результаты подтверждают, что современные модели машинного обучения справляются с рассматриваемыми задачами на приемлемом уровне качества. Собранная коллекция текстов планируется к публикации в открытом доступе (на сайте проектно-учебной лаборатории антикоррупционной политики НИУ «Высшая школа экономики»[1]) и, вероятно, вызовет интерес к возможностям методов обработки текстов у исследователей-соци-

[1] URL: https://lap.hse.ru

ологов, а также привлечет внимание специалистов по машинному обучению к новым предметным областям.

Настоящее исследование имеет ограничение в связи с использованием модели векторизации, свойственное данному методу, которое заключается в использовании достаточных вычислительных мощностей (в том числе графических процессоров). Кроме того, определенные погрешности могли быть связаны с поиском и классификацией цитат, свидетельствующих о новостных сообщениях, содержащих коррупционный контент. При этом использование алгоритмов для контроля «живого кодирования» будет являться одной из задач наших будущих исследований по данной проблематике.

ЛИТЕРАТУРА

1. *Крылова Д.В., Максименко А.А.* Возможности использования искусственного интеллекта в вопросах выявления и противодействия коррупции (обзор международного опыта) // Государственное управление. Электронный вестник. 2021. № 84. С. 245–255.

2. *Cruz J.A., Wishart D.S.* Applications of machine learning in cancer prediction and prognosis // Cancer informatics. 2007. Vol. 2. P. 59–77.

3. Artificial Intelligence, Machine Learning, and Cardiovascular Disease / P. Mathur [et al.] // Clinical Medicine Insights: Cardiology. 2020. Vol. 14. DOI: 10.1177/1179546820927404.

4. Discovery of novel selective PI3Kγ inhibitors through combining machine learning-based virtual screening with multiple protein structures and bio-evaluation / J. Zhu [et al.] // Journal of Advanced Research. 2022. Vol. 36. P. 1–13.

5. Exploring the Potential of Artificial Intelligence and Machine Learning to Combat COVID-19 and Existing Opportunities for LMIC: A Scoping Review / M. Naseem [et al.] // Journal of Primary Care and Community Health. 2020. Vol. 11. Jan-Dec. DOI: 10.1177/2150132720963634.

6. Recognizing software names in biomedical literature using machine learning / Q. Wei [et al.] // Health Informatics Journal. 2020. Vol. 26 (1). P. 21–33. DOI: 10.1177/1460458219869490.

7. *Jain P.K., Pamula R., Srivastava G.* A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews // Computer science review. 2021. 100413. DOI: 10.1016/j.cosrev.2021.100413

8. Machine Learning for industrial applications: A comprehensive literature review / M.Bertolini [et al.] // Expert Systems With Applications. 2021. Vol. 175 (6). 114820. DOI: 10.1016/j.eswa.2021.114820

9. *MelekAkcay M., Etiz D., Celik O.* Prediction of Survival and Recurrence Patterns by Machine Learning in Gastric Cancer Cases Undergoing Radiation Therapy and Chemotherapy // Advances in Radiation Oncology. 2020. Vol. 5. P. 1179–1187.

10. Can machine learning be useful as a screening tool for depression in primary care / E.M. de Souza Filho [et al.] // Journal of Psychiatric Research. 2021. Vol. 132. P. 1–6.

11. *Derevitskii I.V., Kovalchuk S.V.* Machine Learning-Based Predictive Modeling of Complications of Chronic Diabetes // Procedia Computer Science. 2020. Vol. 178. P. 274–283.

12. *Balaji T.K., Annavarapu Ch.S.R., Bablani A.* Machine learning algorithms for social media analysis: A survey // Computer Science Review. 2021. May. Vol. 40. 100395. DOI: 10.1016/j.cosrev.2021.100395

13. Value co-creation for open innovation: An evidence-based study of the data driven paradigm of social media using machine learning / A. Adikari [et al.] // International Journal of Information Management Data Insights. 2021. Novmber. Vol. 1. Iss. 2. 100022.

14. Fine-grained assessment of greenspace satisfaction at regional scale using content analysis of social media and machine learning / Zh. Wang [et al.] // Science of The Total Environment. 2021. Vol. 776. Jul 1. 145908. DOI: 10.1016/j.scitotenv.2021.145908.

15. *Weimin Z.* From Generalization to Specialization: Reflection on the Application of Judicial Artificial Intelligence in China // Legal Forum. 2020. Vol. 35. Iss. 17. P. 20.

16. *Wang R.* Legal technology in contemporary USA and China // Computer law and security. 2020. Vol. 39. 105459. DOI: 10.1016/j.clsr.2020.105459

17. *Lusheng W.* Jurisprudence Conflict and Value Balance in the Application of Judicial Big Data: A Survey on the Article 33 of French Judicial Reform Act 2019 // The Journal of Comparative Law. 2020. Vol. 2. Iss. 133. P. 145.

18. *Sharma A., Shekhar H.* Intelligent Learning based Opinion Mining Model for Governmental Decision Making // Procedia Computer Science. 2020. Vol. 173. P. 216–224.

19. An approach for combining ethical principles with public opinion to guide public policy / E. Awad [et al.] // Artificial Intelligence. 2020. Vol. 287 (7710). 103349. DOI: 10.1016/j.artint.2020.103349

20. Comparing tweet sentiments in megacities using machine learning techniques: In the midst of COVID-19 / Zh. Yao [et al.] // Cities. September. 2021. Vol. 116. 103273.

21. Whether the weather will help us weather the COVID-19 pandemic: Using machine learning to measure twitter users' perceptions / Gupta M. [et al.] // Inter-

national Journal of Medical Informatics. 2021. Vol. 145. 104340. DOI: 10.1016/j.ijmedinf.2020.104340

22. Early Warning Scheme of COVID-19 related Internet Public Opinion based on RVM-L Model / R.Zhu [et al.] // Sustainable Cities and Society. 2021. Vol. 74. 103141. DOI: 10.1016/j.scs.2021.103141

23. Conceptualizing social protest and the significance of protest actions to large projects / Ph. Hanna [et al.] // The Extractive Industries and Society. 2016. Vol. 3. Iss. 1. P. 217–239.

24. *El Feki Sh.* Sexual Politics in the Arab World // International Encyclopedia of the Social & Behavioral Sciences. 2nd ed. Elsevier, 2015. P. 791–796.

25. *Schuster J.* Intersectional expectations: Young feminists' perceived failure at dealing with differences and their retreat to individualism // Women's Studies International Forum. 2016. Vol. 58. P. 1–8.

26. Social media and farmer's resilience to drought as an environmental disaster: A moderation effect / S.S. Bathaiy [et al.] // International Journal of Disaster Risk Reduction. 2021. 1 June. Vol. 59. 102209.

27. Social media users' online subjective well-being and fatigue: A network heterogeneity perspective / P. Kaur [et al.] // Technological Forecasting and Social Change. 2021. November. Vol. 172. 121039.

28. *Wang J., Jia Y.* Social media's influence on air quality improvement: Evidence from China // Journal of Cleaner Production. 2021. 20 May. Vol. 298. 126769.

29. The role of social media-led and governmental information in China's urban disaster risk response: The case of Xiamen / I. Boas [et al.] // International Journal of Disaster Risk Reduction. 2020. December. Vol. 51. 101905.

30. *Zhao L.* The impact of social media use types and social media addiction on subjective well-being of college students: A comparative analysis of addicted and non-addicted students // Computers in Human Behavior Reports. 2021. Vol. 4. P. 100–122.

31. Social media, body satisfaction and well-being among adolescents: A mediation model of appearance-ideal internalization and comparison / H.K. Jarman [et al.] // Body Image. 2021. Vol. 36. P. 139–148.

32. *Aggarwal C C., Zhai C.X.* A survey of text classification algorithms // Mining text data. Springer. 2012. P. 163–222.

33. Neural Architectures for Named Entity Recognition / G. Lample [et al.] // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, 2016. P. 260–270.

34. Readings in information retrieval / Ed. by K.S. Jones, P. Willett. San Francisco: Morgan Kaufmann, 1997.

35. SQuAD: 100,000+ Questions for Machine Comprehension of Text / P. Rajpurkar [et al.] // Proceedings of the 2016 Conference on Empirical Methods

in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, 2016. P. 2383–2392.

36. *Кольцова О.Ю., Ефимова Т.Г.* Выявление социальных проблем и изменений через анализ больших массивов текстов в блогах и социальных сетях // Социальные коммуникации: универсум профессиональной деятельности. Материалы Всероссийского научно-практического симпозиума 9–10 ноября 2011 г. СПб.: Скифия-принт, 2011. С. 274–284.

37. *Kuratov Y., Arkhipov M.* Adaptation of deep bidirectional multilingual transformers for Russian language // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". Moscow. May 29–June 1. 2019. P. 333–339.

38. Attention is all you need / A. Vaswani [et al.] // Advances in neural information processing systems (NIPS 2017). Montreal: Curran Associates, 2017. P. 5998–6008.

39. *Nallapati R., Zhai F., Zhou B.* SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017. Vol. 31. No 1. Retrieved from: https://ojs.aaai.org/index.php/AAAI/article/view/10958

40. SMOTE: synthetic minority over-sampling technique / N.V. Chawla [et al.] // Journal of artificial intelligence research. 2002. Vol. 16. P. 321–335.

41. *Wei J., Zou K.* EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics. Hong Kong, China, 2019. P. 6382–6388.

42. *Николенко С., Кадурин А., Архангельская Е.* Глубокое обучение. СПб.: Питер, 2017.

43. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin [et al.] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. P. 4171–4186.

Artemova Ekaterina,
HSE University, Moscow, Russia, elartemova@hse.ru

Maksimenko Aleksandr,
HSE University, Moscow, Russia, maximenko.al@gmail.com

Ohrimenko Dmitriy,
HSE University, Moscow, Russia, ohrimenko@hse.ru

Application of machine learning methods in the classification of corruption-related content in Russian-speaking and English-speaking Internet media

The paper attempts to classify the corruption-related media content of Russian-language and English-language Internet media using machine learning methods. The methodological approach proposed in the article is very relevant and promising, since, according to our earlier data, corruption monitoring mechanisms used in foreign publications based on the use of advanced information technologies have rather limited potential effectiveness and are not always adequately interpreted. The study shows the principles and grounds for identifying identification parameters, and also describes in detail the layout scheme of the collected news array. In the course of automatic text processing, which took place in 2 stages (vectorization of the text and the use of a learning model), it was possible to solve the main 4 tasks: highlighting a significant quote from a news article to identify a text on corruption topics, predicting the type of news message, predicting a relevant article of the Criminal Code of the Russian Federation, which is used to determine responsibility for the described corruption offense, as well as predicting the type of relationship in corruption offenses. The results obtained showed that modern methods of automatic text processing successfully cope with the tasks of identification and classification of corruption-related content in both Russian and English.
Keywords: corruption-related content; corruption offenses; artificial intelligence; Internet media; clustering algorithms; Russian-language media; English-language media

References

1. Krylova D.V., Maksimenko A.A. Using artificial intelligence in corruption discernment and counteraction: international experience review (in Russian), *Public Administration. E-journal.* 2021. № 84. P. 245–255.
2. Cruz J.A., Wishart D.S. Applications of machine learning in cancer prediction and prognosis, *Cancer informatics*. 2007. Vol. 2. P. 59–77.
3. Mathur P. et al. Artificial Intelligence, Machine Learning, and Cardiovascular Disease, *Clinical Medicine Insights: Cardiology*. 2020. Vol. 14. DOI: 10.1177/1179546820927404.
4. Zhu J. et al. Discovery of novel selective PI3Kγ inhibitors through combining machine learning-based virtual screening with multiple protein structures and bio-evaluation, *Journal of Advanced Research.* 2022. Vol. 36. P. 1–13.
5. Naseem M. et al. Exploring the Potential of Artificial Intelligence and Machine Learning to Combat COVID-19 and Existing Opportunities for LMIC: A Scoping Review, *Journal of Primary Care and Community Health.* 2020. Vol. 11. Jan-Dec. DOI: 10.1177/2150132720963634.
6. Wei Q. et al. Recognizing software names in biomedical literature using machine learning, *Health Informatics Journal.* 2020. Vol. 26 (1). P. 21–33. DOI: 10.1177/1460458219869490.
7. Jain P.K., Pamula R., Srivastava G. A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews, *Computer science review.* 2021. 100413. DOI: 10.1016/j.cosrev.2021.100413
8. Bertolini M. et al. Machine Learning for industrial applications: A comprehensive literature review, *Expert Systems with Applications.* 2021. Vol. 175 (6). 114820. DOI: 10.1016/j.eswa.2021.114820
9. MelekAkcay M., Etiz D., Celik O. Prediction of Survival and Recurrence Patterns by Machine Learning in Gastric Cancer Cases Undergoing Radiation Therapy and Chemotherapy, *Advances in Radiation Oncology.* 2020. Vol. 5. P. 1179–1187.
10. de Souza Filho E.M. et al. Can machine learning be useful as a screening tool for depression in primary care, *Journal of Psychiatric Research.* 2021. Vol. 132. P. 1–6.
11. Derevitskii I.V., Kovalchuk S.V. Machine Learning-Based Predictive Modeling of Complications of Chronic Diabetes, *Procedia Computer Science.* 2020. Vol. 178. P. 274–283.

12. Balaji T.K., Annavarapu Ch.S.R., Bablani A. Machine learning algorithms for social media analysis: A survey, *Computer Science Review.* 2021. May. Vol. 40. 100395. DOI: 10.1016/j.cosrev.2021.100395
13. Adikari A. et al. Value co-creation for open innovation: An evidence-based study of the data driven paradigm of social media using machine learning, *International Journal of Information Management Data Insights.* 2021. Novmber. Vol. 1. Iss. 2. 100022.
14. Wang Zh. et al. Fine-grained assessment of greenspace satisfaction at regional scale using content analysis of social media and machine learning, *Science of The Total Environment.* 2021. Vol. 776. Jul 1. 145908. DOI: 10.1016/j.scitotenv.2021.145908.
15. Weimin Z. From Generalization to Specialization: Reflection on the Application of Judicial Artificial Intelligence in China, *Legal Forum.* 2020. Vol. 35. Iss. 17. P. 20.
16. Wang R. Legal technology in contemporary USA and China, *Computer law and security.* 2020. Vol. 39. 105459. DOI: 10.1016/j.clsr.2020.105459
17. Lusheng W. Jurisprudence Conflict and Value Balance in the Application of Judicial Big Data: A Survey on the Article 33 of French Judicial Reform Act 2019, *The Journal of Comparative Law.* 2020. Vol. 2. Iss. 133. P. 145.
18. Sharma A., Shekhar H. Intelligent Learning based Opinion Mining Model for Governmental Decision Making, *Procedia Computer Science.* 2020. Vol. 173. P. 216–224.
19. Awad E. et al. An approach for combining ethical principles with public opinion to guide public policy, *Artificial Intelligence.* 2020. Vol. 287 (7710). 103349. DOI: 10.1016/j.artint.2020.103349
20. Yao Zh. et al. Comparing tweet sentiments in megacities using machine learning techniques: In the midst of COVID-19, *Cities*. September. 2021. Vol. 116. 103273.
21. Gupta M. et al. Whether the weather will help us weather the COVID-19 pandemic: Using machine learning to measure twitter users' perceptions, *International Journal of Medical Informatics.* 2021. Vol. 145. 104340. DOI: 10.1016/j.ijmedinf.2020.104340
22. Zhu R. et al. Early Warning Scheme of COVID-19 related Internet Public Opinion based on RVM-L Model, *Sustainable Cities and Society.* 2021. Vol. 74. 103141. DOI: 10.1016/j.scs.2021.103141

23. Hanna Ph. et al. Conceptualizing social protest and the significance of protest actions to large projects, *The Extractive Industries and Society.* 2016. Vol. 3. Iss. 1. P. 217–239.
24. El Feki Sh. Sexual Politics in the Arab World, International Encyclopedia of the Social & Behavioral Sciences. 2nd ed. Elsevier, 2015. P. 791–796.
25. Schuster J. Intersectional expectations: Young feminists' perceived failure at dealing with differences and their retreat to individualism, *Women's Studies International Forum.* 2016. Vol. 58. P. 1–8.
26. Bathaiy S. S. et al. Social media and farmer's resilience to drought as an environmental disaster: A moderation effect, *International Journal of Disaster Risk Reduction.* 2021. 1 June. Vol. 59. 102209.
27. Kaur P. et al. Social media users' online subjective well-being and fatigue: A network heterogeneity perspective, *Technological Forecasting and Social Change.* 2021. November. Vol. 172. 121039.
28. Wang J., Jia Y. Social media's influence on air quality improvement: Evidence from China, *Journal of Cleaner Production.* 2021. 20 May. Vol. 298. 126769.
29. Boas I. et al. The role of social media-led and governmental information in China's urban disaster risk response: The case of Xiamen, *International Journal of Disaster Risk Reduction.* 2020. December. Vol. 51. 101905.
30. Zhao L. The impact of social media use types and social media addiction on subjective well-being of college students: A comparative analysis of addicted and non-addicted students, *Computers in Human Behavior Reports.* 2021. Vol. 4. P. 100–122.
31. Jarman H.K. et al. Social media, body satisfaction and well-being among adolescents: A mediation model of appearance-ideal internalization and comparison, *Body Image.* 2021. Vol. 36. P. 139–148.
32. Aggarwal C.C., Zhai C.X. A survey of text classification algorithms, *Mining text data.* Springer. 2012. P. 163–222.
33. Lample G. et al. Neural Architectures for Named Entity Recognition, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* San Diego, California, 2016. P. 260–270.
34. Jones K.S., Willett P. (eds.) *Readings in information retrieval.* San Francisco: Morgan Kaufmann, 1997.

35. Rajpurkar P. et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2016. P. 2383–2392.
36. Koltsova O.Y., Efimova T.G. Identification of social problems and changes through the analysis of large arrays of texts in blogs and social networks (in Russian), *Social communications: the universe of professional activity.* Proceedings of the Russian scientific and practical symposium, November 9–10, 2011. SPb., 2011. P. 274–284.
37. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*. Moscow. May 29–June 1. 2019. P. 333–339.
38. Vaswani A. et al. Attention is all you need, *Advances in neural information processing systems (NIPS 2017)*. Montreal: Curran Associates, 2017. P. 5998–6008.
39. Nallapati R., Zhai F., Zhou B. SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.* 2017. Vol. 31. No 1. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10958 (date of access: 20.12.2021)
40. Chawla N. V. et al. SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research.* 2002. Vol. 16. P. 321–335.
41. Wei J., Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.* Association for Computational Linguistics. Hong Kong, China, 2019. P. 6382–6388.
42. Nikolenko S., Kadurin A., Arkhangelskaya E. *Deep learning* (in Russian). St. Petersburg: Peter, 2017.
43. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. P. 4171–4186.