MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 29, NO. 4, 2022
journal homepage: www.mais-journal.ru

THEORY OF DATA

Detecting Mentions of Green Practices in Social Media Based on Text Classification

A. V. Glazkova[1], O. V. Zakharova[1], A. V. Zakharov[1], N. N. Moskvina[1], T. R. Enikeev[2], A. N. Hodyrev[1], V. K. Borovinskiy[1], I. N. Pupysheva[1]

DOI: 10.18255/1818-1015-2022-4-316-332

[1]University of Tyumen, 6 Volodarskogo str., Tyumen 625003, Russia.
[2]Novosibirsk State University, 1 Pirogova str., Novosibirsk 630090, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received October 6, 2022
After revision November 11, 2022
Accepted November 16, 2022

The paper is devoted to the task of searching for mentions of green practices in social media texts. The relevance of this task is dictated by the need to expand existing knowledge about the use of green practices in society and the spread of existing green practices. This paper uses a text corpus consisting of the texts published on the environmental communities of the VKontakte social network. The corpus is equipped with an expert markup of the mention of nine types of green practices. As part of this work, a semi-automatic approach is proposed to the collection of additional texts to reduce the class imbalance in the corpus. The approach includes the following steps: detecting the most frequent words for each practice type; automatic collecting texts in social media that contain the detected frequent words; expert verification and filtering of collected texts. The four machine learning models are compared to find the mentions of green practices on the two variants of the corpus: original and augmented using the proposed approach. Among the listed models, the highest averaged F1-score (81.32%) was achieved by Conversational RuBERT fine-tuned on the augmented corpus. Conversational RuBERT model was chosen for the implementation of the application prototype. The main function of the prototype is to detect the presence of the mention of nine types of green practices in the text. The prototype is implemented in the form of the Telegram chatbot.

Keywords: text classification; social network analysis; machine learning; BERT; green practices; natural language processing

INFORMATION ABOUT THE AUTHORS

Anna Valerevna Glazkova correspondence author	orcid.org/0000-0001-8409-6457. E-mail: a.v.glazkova@utmn.ru PhD, Associate Professor of the Department of Software.
Olga Vladimirovna Zakharova	orcid.org/0000-0002-1404-4915. E-mail: o.v.zakharova@utmn.ru PhD, Head of Project Office at Green Solutions Lab, Associate Professor, Department of State and Municipal Administration.
Anton Viktorovich Zakharov	orcid.org/0000-0002-0093-049X. E-mail: a.v.zakharov@utmn.ru PhD, Chief Scientific Officer, Department of the Ishim Pedagogical Institute.
Natalya Nikolayevna Moskvina	orcid.org/0000-0001-5198-276X. E-mail: n.n.moskvina@utmn.ru PhD, Associate Professor, Department of Physical Geography and Ecology.
Timur Ruslanovich Enikeev	orcid.org/0000-0001-8195-1278. E-mail: t.enikeev@g.nsu.ru Student.
Arseniy Nikolaevich Hodyrev	orcid.org/0000-0001-7151-9852. E-mail: stud0000247809@study.utmn.ru Student.
Vsevolod Konstantinovich Borovinskiy	orcid.org/0000-0001-6193-6548. E-mail: stud0000224807@study.utmn.ru Student.
Irina Nikolayevna Pupysheva	orcid.org/0000-0003-2870-4870. E-mail: i.n.pupysheva@utmn.ru PhD, Associate Professor of the Department of Philosophy.

Funding: The work was carried out during the Big Mathematical Workshop of the Mathematical Center in Akademgorodok.

For citation: A. V. Glazkova, O. V. Zakharova, A. V. Zakharov, N. N. Moskvina, T. R. Enikeev, A. N. Hodyrev, V. K. Borovinskiy, and I. N. Pupysheva, "Detecting Mentions of Green Practices in Social Media Based on Text Classification", *Modeling and analysis of information systems*, vol. 29, no. 4, pp. 316-332, 2022.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 29, № 4, 2022
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Поиск упоминаний экологических практик в социальных сетях с помощью методов классификации текстов

А. В. Глазкова[1], О. В. Захарова[1], А. В. Захаров[1], Н. Н. Москвина[1], Т. Р. Еникеев[2], А. Н. Ходырев[1], В. К. Боровинский[1], И. Н. Пупышева[1]

DOI: 10.18255/1818-1015-2022-4-316-332

[1]Тюменский государственный университет, ул. Володарского, д. 6, г. Тюмень, 625003 Россия.
[2]Новосибирский государственный университет, ул. Пирогова, д. 1, г. Новосибирск, 630090 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 6 октября 2022 г.
После доработки 11 ноября 2022 г.
Принята к публикации 16 ноября 2022 г.

Работа посвящена решению задачи поиска упоминаний экологических практик в текстах социальных сетей. Авторами составлен корпус текстов экологических сообществ социальной сети ВКонтакте, снабженный экспертной разметкой упоминаний девяти видов экологических практик. Предложен полуавтоматический подход к сбору дополнительных текстов для уменьшения несбалансированности видов экологических практик, представленных в корпусе. Подход включает в себя следующие этапы: определение наиболее частотных слов, характеризующих упоминания практик; автоматический сбор текстов, включающих в себя найденные частотные слова; экспертная проверка и фильтрация собранных текстов. Проведено сравнение четырех моделей машинного обучения для поиска упоминаний практик на двух вариантах корпуса: исходном и дополненном. Лучший усредненный показатель F-меры (81.32%) достигнут моделью Conversational RuBERT, дообученной на текстах дополненного корпуса. Данная модель выбрана в качестве основы для реализации прототипа приложения для поиска упоминаний экологических практик, реализованного в форме чат-бота Telegram.

Ключевые слова: классификация текстов; анализ социальных сетей; машинное обучение; BERT; экологические практики; обработка естественного языка

ИНФОРМАЦИЯ ОБ АВТОРАХ

Анна Валерьевна Глазкова автор для корреспонденции	orcid.org/0000-0001-8409-6457. E-mail: a.v.glazkova@utmn.ru канд. тех. наук, оцент кафедры программного обеспечения.
Ольга Владимировна Захарова	orcid.org/0000-0002-1404-4915. E-mail: o.v.zakharova@utmn.ru канд. филос. наук, руководитель проектного офиса Green Solutions Lab, доцент кафедры государственного и муниципального управления.
Антон Викторович Захаров	orcid.org/0000-0002-0093-049X. E-mail: a.v.zakharov@utmn.ru канд. пед. наук, нач. научного отд. Ишимского пед. института им. П. П. Ершова.
Наталья Николаевна Москвина	orcid.org/0000-0001-5198-276X. E-mail: n.n.moskvina@utmn.ru канд. геогр. наук, доцент кафедры физической географии и экологии.
Тимур Русланович Еникеев	orcid.org/0000-0001-8195-1278. E-mail: t.enikeev@g.nsu.ru студент.
Арсений Николаевич Ходырев	orcid.org/0000-0001-7151-9852. E-mail: stud0000247809@study.utmn.ru студент.
Всеволод Константинович Боровинский	orcid.org/0000-0001-6193-6548. E-mail: stud0000224807@study.utmn.ru студент.
Ирина Николаевна Пупышева	orcid.org/0000-0003-2870-4870. E-mail: i.n.pupysheva@utmn.ru канд. филос. наук, доцент кафедры философии.

Финансирование: Исследование выполнено в рамках работы на Большой математической мастерской, организованной Математическим центром в Академгородке в 2022 году.

Для цитирования: A. V. Glazkova, O. V. Zakharova, A. V. Zakharov, N. N. Moskvina, T. R. Enikeev, A. N. Hodyrev, V. K. Borovinskiy, and I. N. Pupysheva, “Detecting Mentions of Green Practices in Social Media Based on Text Classification”, *Modeling and analysis of information systems*, vol. 29, no. 4, pp. 316-332, 2022.

Введение

В условиях ухудшающейся экологической ситуации и дефицита ресурсов необходимо активно привлекать общество к социальным *экологическим (зеленым) практикам*, то есть к повседневным действиям, направленным на гармонизацию отношений человека и его окружающей среды [1]. Согласно данным Всероссийского центра изучения общественного мнения, некоторые из этих действий, способные повлиять на экологическую ситуацию, не только мало распространены, но и остаются практически незамеченными [2, 3]. Привлечение общества к экологическим практикам возможно, например, путем масштабирования уже имеющихся на сегодняшний день экологических практик, особенно тех, которые направлены на сокращение потребления, а значит, на сокращение используемых ресурсов и производимых загрязнений. В зависимости от того, как экологические практики влияют на потребление, их можно разделить на два типа: адаптационные и трансформационные. *Адаптационными* практиками будем называть практики, которые являются реакцией общества на ухудшающуюся экологическую обстановку, но не предполагают сокращения потребления, а *трансформационными* – практики, которые рассчитаны на сокращение производства товаров и услуг и потребления обществом вещества и энергии.

Чтобы эффективно внедрять экологические практики, необходимо обладать определённой информационной базой о том, какие экологические практики уже существуют в обществе, насколько распространены среди них те, которые ведут к сокращению потребления, кто является инициатором подобных практик, кто их поддерживает и так далее. Однако на сегодняшний день знаний о распространенности экологических практик очень мало, поскольку сбор требуемого объема информации традиционными социологическими методами (анкетирование, интервью) является очень трудоемким и занимает много времени [4]. Тем не менее, в социальных сетях в настоящее время сформирован значительный объем неструктурированной текстовой информации, связанной с экологической тематикой. Автоматический анализ текстов экологических сообществ в социальных сетях позволил бы собрать и структурировать большое количество текстовых данных в рассматриваемой предметной области, ускорить их обработку и сделать выводы о распространенности тех или иных видов практик. В связи с этим возникает необходимость разработки методов автоматического получения информации об экологических практиках в социальных сетях. В данной работе описывается подход к автоматическому поиску упоминаний экологических практик в текстах социальных сетей с помощью методов классификации текстов. Авторами представлен корпус текстов социальных сетей, снабженный экспертной разметкой девяти видов экологических практик. Предлагается подход к полуавтоматическому дополнению исходного корпуса текстов для уменьшения дисбаланса количества различных видов практик. Приводятся и обсуждаются результаты сравнения нескольких моделей машинного обучения для автоматического поиска упоминаний экологических практик. Одним из результатов работы является прототип приложения для поиска упоминаний экологических практик в текстах социальных сетей, реализованный в форме чат-бота Telegram.

Работа структурирована следующим образом. Раздел 1 содержит обзор текущего состояния области классификации текстов применительно к анализу текстов социальных сетей. В разделе 2 приводится постановка задачи. Раздел 3 содержит описание используемого корпуса текстов. В разделе 4 приводится перечень использованных моделей машинного обучения. Описание полученных результатов содержится в разделе 5.

1. Обзор смежных работ

Данная работа связана с анализом текстов, размещенных в социальных сетях, и, в частности, их классификацией с применением методов машинного обучения, обработки естественного языка и компьютерной лингвистики. Открытость и разнообразие текстовых данных, размещенных в социальных сетях, предоставляет широкие возможности для изучения общественного мнения

с помощью инструментов компьютерного анализа и позволяет анализировать пути распространения социально значимой информации в онлайн-источниках [5]. Таким образом, социальные сети служат своеобразным индикатором общественных взглядов и тенденций. Изучение контента социальных сетей является важной задачей для научного, политического и коммерческого сообществ [6], что делает автоматическую обработку и анализ постов актуальной тематикой исследований в области компьютерных наук.

Эффективность классификации текстов в области социальных сетей росла параллельно развитию методологии обработки естественного языка (natural language processing). Для решения данной задачи привлекались различные методы, начиная от классификаторов, основанных на применении правил, и заканчивая современными моделями, базирующимися на использовании глубоких нейронных сетей [7]. На сегодняшний день большинство задач классификации текстов решаются с использованием методов машинного обучения и, в частности, глубокого обучения.

Среди традиционных методов машинного обучения наивный байесовский классификатор и логистическая регрессия применяются для анализа тональности (sentiment analysis) постов в Twitter [8, 9] и «риторики вражды» (hate speech detection) [10, 11]. Также для классификации текстов социальных сетей широко используются метод ближайших соседей, метод опорных векторов и случайный лес (например, в работах [12—14]). Среди нейросетевых моделей распространено использование сетей долгой краткосрочной памяти (Long Short-Term Memory, LSTM) [15], которые применялись для анализа тональности [16—18], поиска оскорбительного контента [19, 20] и недостоверной («фейковой») информации [21, 22], а также сверточных нейронных сетей (Convolutional Neural Networks, CNN), использование которых для задачи классификации текстов было впервые предложено в статье [23]. В частности, сверточные нейронные сети применялись для классификации постов социальных сетей в работах [24, 25]. В работах [26, 27] были предложены подходы к гибридному использованию LSTM и CNN. На сегодняшний день, наиболее высокое качество во многих задачах классификации текстов показывают нейросетевые модели, основанные на архитектуре Transformer [28] и, в частности, на использовании лингвистических моделей Bidirectional Encoder Representations from Transformers (BERT) [29], RoBERTa [30] и их модификаций. Так, в ряде соревнований по машинному обучению, связанных с тематикой классификации текстов социальных сетей и проводимых в рамках крупнейших конференций, лучшие результаты были получены с помощью BERT и ее вариаций (например, [31—33]).

В течение последних лет было опубликовано большое количество исследований, связанных с классификацией текстов социальных сетей русскоязычного сегмента Интернета. Авторами данных исследований представлено множество разнообразных подходов. Так, в работах [34—36] рассматриваются подходы к автоматической классификации постов в Twitter с помощью словарей. В статьях [37—39] описываются результаты применения методов машинного обучения к задаче классификации текстов социальных сетей. В работах [40, 41] оценивается эффективность использования признаков различной природы для классификации постов. Подходы, основанные на применении глубоких нейронных сетей, представлены, в частности, в статьях [42—45].

Помимо классификации постов социальных сетей, данная работа также связана с анализом текстов экологической тематики. Ряд работ в данной области посвящен библиометрическому анализу научных текстов для выявления основных трендов экологической повестки [46]. В частности, в работе [47] представлены результаты обширного частотного анализа текстов экологических журналов для выявления динамики словоупотребления тех или иных терминов. В статье [48] проведена кластеризация аннотаций статей экологической тематики. В работах [49—51] представлены результаты частотного анализа публикаций в более узких предметных областях. Некоторые исследования посвящены применению методов машинного обучения в области анализа экологических текстов. Так, в работе [52] предложен подход к автоматическому подбору статей по тематике

биоразнообразия с помощью логистической регрессии и сверточных нейронных сетей. Авторы получили достаточно высокое качество бинарной классификации (ROC AUC ≥ 98%) на датасете, содержащем заголовки и аннотации статей. При этом эксперименты, проведенные в указанной работе, показали, что на датасетах сравнительно небольшой размерности логистическая регрессия и нейронная сеть продемонстрировали близкие показатели качества. В работе [53] предлагается модель для автоматического извлечения таксономических категорий на основе нейронных сетей с Transformer-архитектурой. Предложенная модель продемонстрировала лучшее показатели F-меры в сравнении с другими моделями для распознавания таксономических категорий на двух корпусах текстов (> 75% для COPIOUS [54] и > 88% для Bacteria Biotope [55]). Однако качество на тестовой выборке, содержащей тексты биомедицинской тематики, оказалось существенно ниже, что позволило сделать вывод о важности разработки предметно-ориентированных моделей для анализа биомедицинских и экологических текстов.

Обзор смежных работ показывает, что недавние достижения в области обработки естественного языка демонстрируют высокое качество автоматического анализа текстов для решения множества практических задач, в том числе задачи классификации текстов социальных сетей. Технологии искусственного интеллекта используются в различных областях научного знания. Тем не менее, автоматический анализ естественного языка в области экологических исследований, представлен в настоящее время преимущественно в форме частотного анализа и лишь точечно в виде моделей машинного обучения. Развитие методологии анализа текстов экологической тематики, основанной на применении машинного обучения, представляется перспективной задачей, решение которой служит преодолению существующего пробела в области автоматического анализа экологических текстов.

2. Постановка задачи

Целью данной работы является разработка подхода к автоматическому поиску упоминаний экологических практик в текстах социальных сетей. При наличии нескольких видов экологических практик и размеченного корпуса текстов задача автоматического поиска упоминаний экологических практик в тексте на естественном языке может быть решена как задача классификации. В таком случае задача может быть представлена в формализованном виде следующим образом. Дано множество текстов $T = \{t_1, t_2, ..., t_n\}$ и множество экологических практик $P = \{p_1, p_2, ..., p_m\}$. Требуется найти решающую функцию, приближающую неизвестную целевую зависимость $F : T \longrightarrow 2^P$, значения которой известны только на обучающей выборке. При этом каждому тексту $t_i \in T$, $1 \leq i \leq n$, соответствует некое подмножество экологических практик $P_i \subseteq P$, то есть один текст может содержать упоминания нескольких практик или не содержать упоминаний практик.

В данной работе использован подход, основанный на разбиении задачи классификации с несколькими метками (multi-label classification) на бинарные по схеме «один против остальных» (one-vs-rest). В таком случае задача с несколькими метками преобразуется в m задач бинарной классификации (m – количество рассматриваемых практик), целью каждой из которых является определение класса, к которому относится текст $t_i \in T$. В бинарной постановке каждый классификатор определяет наличие в тексте упоминания одной из экологических практик. Преимуществами подхода «один против остальных» являются его вычислительная эффективность за счет использования бинарных классификаторов и интерпретируемость, то есть возможность получить информацию о каждом классе в отдельности, использовав соответствующий классификатор.

3. Корпус текстов

Для решения задачи поиска упоминаний экологических практик в текстах социальных сетей был использован корпус постов экологических сообществ социальной сети ВКонтакте. Корпус включает в себя посты шести крупнейших экологических сообществ Тюменской области, собранные

в период с января по июнь 2021 года с помощью инструмента VK API[1]. Корпус не включает в себя посты, не содержащие текстовой информации, а также дублирующиеся посты. Общий объем используемого текстового корпуса – 1987 текстов. Корпус также снабжен информацией о количестве комментариев и лайков для каждого поста.

3.1. Разметка корпуса

Используемый в работе текстовый корпус снабжен экспертной разметкой упоминаний экологических практик, включающей в себя практики следующих видов:

- адаптационные практики:
  - сортировать отходы (P1);
  - изучать маркировку товаров (P2);
  - перерабатывать отходы (P3);
  - подписывать петиции (P4);
- трансформационные практики:
  - отказываться от покупок (P5);
  - обменивать (P6);
  - совместно использовать (P7);
  - продвигать ответственное потребление (P8);
  - ремонтировать (P9).

Разметка представляет собой выделение фрагмента, содержащего упоминание практики, открывающим и закрывающим тэгами в соответствии со следующей схемой:

<номер практики>фрагмент, содержащий упоминание практики</номер практики>.

Например, в тексте «На улице холодает, чтобы <3>ускорить свою сдачу сырья</3>, не забывайте заранее <1>максимально его сортировать и подготавливать к сдаче</1>» выделены фрагменты с упоминаниями практик с номерами 3 и 1 («перерабатывать отходы» и «сортировать отходы» соответственно).

Разметка корпуса проводилась двумя экспертами из Тюменского государственного университета, имеющими опыт в области изучения экологических практик. На первом этапе разметки от экспертов требовалось независимо друг от друга выделить в постах максимально короткие, но семантически полные упоминания экологических практик. На втором этапе проводилась проверка и корректировка разметки. Консенсус в случае расхождения между двумя выполненными разметками достигался в ходе дискуссии обоими экспертами.

3.2. Количественный анализ корпуса

На рисунке 1 представлено распределение количества постов, содержащих упоминания экологических практик, в исходном корпусе текстов. Наиболее упоминаемыми практиками в корпусе являются сортировка отходов (37% от общего количества текстов, содержащих упоминания экологических практик), переработка отходов (27.3%) и продвижение ответственного потребления (17.2%). При этом для ряда практик корпус содержит лишь единичные случаи упоминания. В частности, миноритарными являются практики совместного использования (3 текста, содержащих упоминания, или 0.2% от общего количества текстов, содержащих упоминания экологических практик) и ремонта (11 текстов, 0.6%). Поскольку количество текстов, упоминающих разные практики, значительно различается, разработка классификаторов требует предварительного принятия мер по балансировке размеров классов.

Таблица 1 содержит основные количественные характеристики корпуса текстов. Как видно из данных, представленных в таблице, около трети постов экологических сообществ, входящих в состав корпуса, не содержит упоминаний экологических практик (34.17%), и примерно такое же

[1] https://dev.vk.com/reference

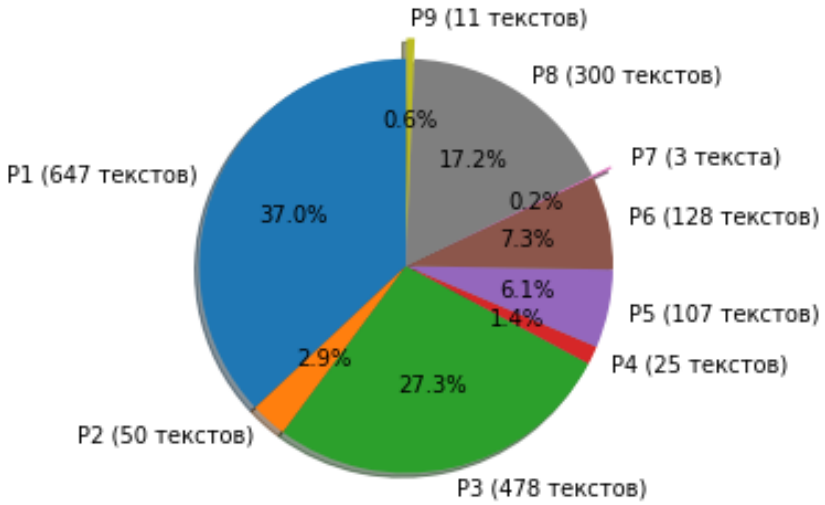
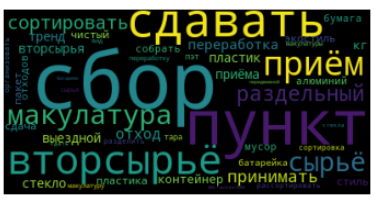
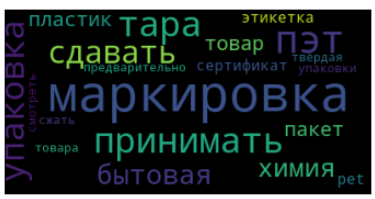


Fig. 1. The distribution of the number of texts containing green practice mentions in the corpus

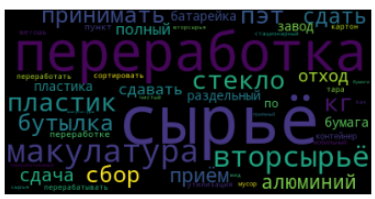
Рис. 1. Распределение количества текстов, содержащих упоминания экологических практик, в корпусе



P1)

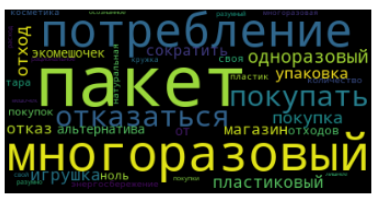


P2)

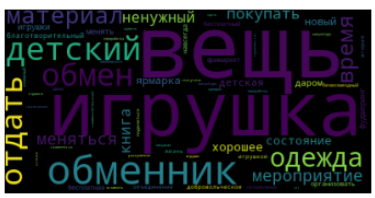


P3)

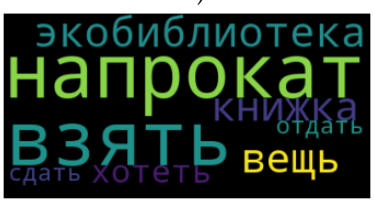
P4)



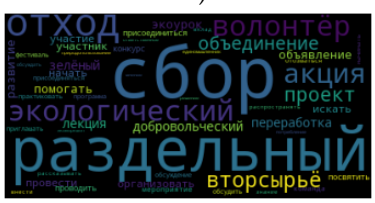
P5)



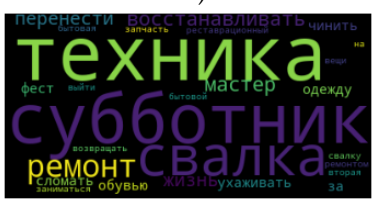
P6)



P7)



P8)



P9)

Fig. 2. The visual representation of the word frequency in the fragments of posts containing the mentions of green practices

Рис. 2. Визуальное представление частотности слов во фрагментах постов, содержащих упоминания экологических практик

количество постов (31%) содержит более одного упоминания. На рисунке 2 в виде «облаков слов» показано визуальное представление частотности слов, входящих в фрагменты постов, содержащих упоминания экологических практик. Эмпирический анализ данных визуальных представлений позволяет сделать выводы о том, что высокочастотные слова для некоторых практик в значительной мере совпадают (в частности, для практик сортировки отходов и переработки отходов).

Table 1. The quantitative characteristics of the corpus

Таблица 1. Количественные характеристики корпуса

Характеристика	Значение
Среднее количество символов в посте	761.67
Среднее количество слов в посте	113.74
Количество постов	1987
Количество постов, не содержащих упоминаний экологических практик	679
Количество постов, содержащих более одного упоминания экологических практик	616

3.3. Дополнение корпуса

Поскольку исходный корпус текстов являлся несбалансированным и упоминания некоторых практик встречались лишь в виде единичных примеров, было решено провести сбор дополнительных текстов для расширения исходного датасета и укрупнения миноритарных категорий постов. В качестве источника дополнительных текстов использовалась социальная сеть ВКонтакте. Так как посты, содержащие упоминания искомых миноритарных экологических практик, являются достаточно редкими, и частотные слова, описывающие различные практики, в некоторой мере пересекаются, не представляется возможным полностью автоматизировать процесс поиска дополнительных текстов. Исходя из этого, в данной работе был использован следующий подход к полуавтоматическому дополнению исходного корпуса текстов.

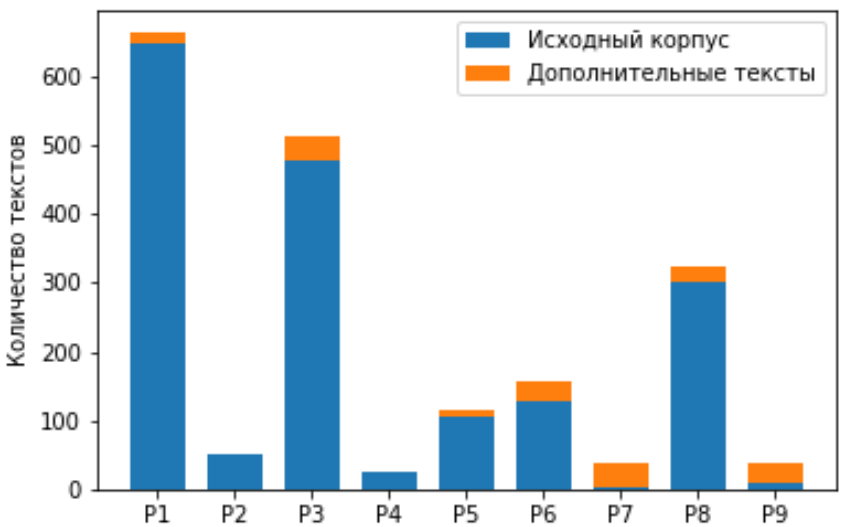


Fig. 3. The number of texts per green practice in the original and augmented corpora

Рис. 3. Количество текстов для каждой экологической практики в исходном и дополненном корпусах

1. На первом этапе для каждой экологической практики были составлены списки слов, наиболее частотных внутри фрагментов, содержащих упоминания этой практики.
2. Далее, на основании выявленных частотных слов с помощью сервиса автоматического поиска целевой аудитории в социальных сетях был получен набор постов, содержащих сочетания частотных слов, присутствующих в упоминаниях миноритарных экологических практик. В данной работе использовался сервис TargetHunter[2].
3. На третьем этапе было проведено экспертное оценивание автоматически собранных текстов с целью уточнения наличия в них упоминаний экологических практик. При этом эксперты проверяли наличие не только миноритарных, но и других видов практик.

На рисунке 3 представлено соотношение количества текстов, содержащих упоминания различных экологических практик, в исходном и дополненном датасетах. Поскольку в ходе экспертного оценивания большая часть автоматически собранных текстов была отмечена как не содержащая упоминаний миноритарных практик, дисбаланс классов в дополненном корпусе сохраняется. Несмотря на это, количество текстов, включающих в себя упоминания миноритарных практик было увеличено в несколько раз.

4. Модели

Данный раздел содержит описание моделей машинного обучения для поиска упоминаний экологических практик в текстах социальных сетей с помощью методов машинного обучения. Для экспериментов по проведению бинарной классификации текстов были использованы несколько моделей машинного обучения. Во-первых, были применены три широко используемых алгоритма машинного обучения: *логистическая регрессия*, *случайный лес* и *метод опорных векторов*. Данные классификаторы были реализованы с помощью библиотеки Scikit-learn [56] и языка программирования Python. При реализации случайного леса было использовано количество деревьев, равное 50. Для остальных гиперпараметров были выбраны настройки по умолчанию. Начальное число, используемое генератором случайных чисел (random_state), было зафиксировано в значении 0. В качестве модели для представления текстов была выбрана модель Bag-of-words («мешок слов») с максимальным размером словаря (max_features), равным 5000. Во-вторых, в ходе экспериментов была использована русскоязычная лингвистическая модель *Conversational RuBERT*[3] [57] (далее – RuBERT), которая представляет собой мультиязычную BERT [29], дополнительно обученную на текстах русскоязычной Википедии, новостных русскоязычных текстах, а также корпусах OpenSubtitles [58], Dirty, Pikabu и Taiga [59]. Для каждой экологической практики проводилось дообучение (fine-tuning) модели RuBERT в течение трех эпох с использованием следующих гиперпараметров: максимальная длина входной последовательности – 256 токенов, размер батча – 8, скорость обучения – 4e-5. Для реализации использовалась библиотека Simple Transformers[4] для языка программирования Python.

5. Результаты

Оценка качества моделей проводилась отдельно на исходном и дополненном датасетах. При этом была использована стратифицированная кросс-валидация с разбиением данных на три части. Для представления результатов использовалась F-мера с усреднением по обоим классам (macro averaging). Результаты моделей представлены в таблице 2, лучший результат для каждой практики выделен полужирным шрифтом. Рисунок 4 иллюстрирует различия в результатах, полученных до и после пополнения корпуса.

Результаты, полученные на исходном корпусе для видов практик, меньше всего представленных в датасете, являются ожидаемо низкими. При этом стоит отметить, что для ряда практик на

[2] https://targethunter.ru/
[3] https://huggingface.co/DeepPavlov/rubert-base-cased-conversational
[4] https://simpletransformers.ai/

исходных данных традиционные методы машинного обучения показали лучшие результаты в сравнении с RuBERT (в частности, практики отказа от покупок и обмена). В целом, достаточно высокое качество (более 75%) на исходном корпусе было достигнуто только для трех наиболее широко представленных экологических практик (сортировка отходов, переработка отходов и продвижение ответственного потребления) и для практики изучения маркировки товаров (вероятно, в связи с наличием специфических слов, входящих в фрагменты, упоминающие данную практику).

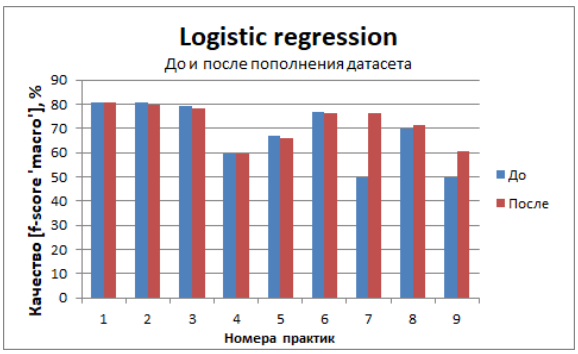
Несмотря на сохраняющуюся несбалансированность корпуса в дополненном датасете, результаты, полученные классификаторами в большинстве случаев значительно улучшились. При этом для большей части практик лучший показатель качества достигнут с помощью модели RuBERT. Так, наиболее высокое значение F-меры получено для практик переработки отходов (82.57% на дополненном датасете, прирост 4.82% по сравнению с исходными данными) и ремонта (82.29% на дополненном датасете, прирост 32.42%). Поскольку при обучении каждого бинарного классификатора используются тексты, относящиеся к разным видам практик, добавление дополнительных текстов в корпус позволило улучшить качество классификации не только при поиске тех практик, примеры упоминаний которых были добавлены при дополнении корпуса, но и при поиске практик, количество упоминаний которых не увеличилось или увеличилось незначительно при пополнении датасета (например, практики подписания петиций и отказа от покупок). В целом, добавление дополнительных текстов в корпус значительно повысило качество классификаторов. При этом по сравнению с результатами на исходном корпусе в случае модели RuBERT показатели качества выросли для восьми экологических практик из девяти рассмотренных, в случае метода опорных векторов – для семи, в случае логической регрессии – для пяти и в случае случайного леса – для двух. Средний прирост качества составил 18.73% для RuBERT, 5.28% для метода опорных векторов, 3.95% для логистической регрессии и 1.15% для случайного леса. Таким образом, эксперименты, проведенные на данном текстовом корпусе, показали, что RuBERT демонстрирует достаточно высокое качество классификации даже при наличии небольшого количества примеров в классах.

Table 2. The results for comparing models (F-score, %): LR – Logistic Regression, RF – Random Forest, SVM – Support Vector Machines, BERT – Conversational RuBERT

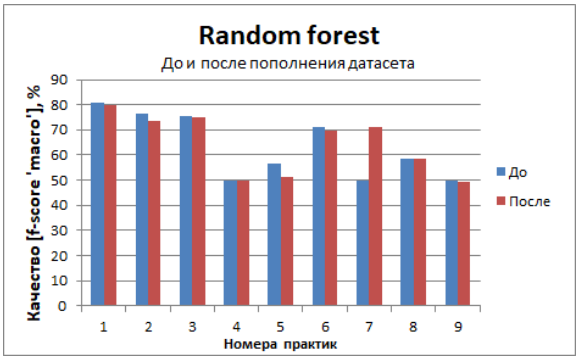
Таблица 2. Результаты сравнения моделей (F-мера, %): LR – логистическая регрессия, RF – случайный лес, SVM – метод опорных векторов, BERT – Conversational RuBERT

Практика	Исходный корпус					Дополненный корпус				
	Количество постов	LR	RF	SVM	BERT	Количество постов	LR	RF	SVM	BERT
P1	647	80.48	80.59	77.13	**83.2**	663	80.68	80.06	78.33	81.15
P2	50	80.56	76.41	**82.9**	78.18	51	79.77	73.53	81.86	80.63
P3	478	79.2	75.37	75.02	77.75	511	78.12	75.2	76	**82.57**
P4	25	59.66	49.62	74.75	52.22	25	59.68	49.64	80.57	**81.46**
P5	107	66.87	56.42	65.9	48.65	116	65.99	51.14	68.35	**79.57**
P6	128	76.68	70.91	79.15	48.35	158	76.22	69.83	78.98	**81.26**
P7	3	49.96	49.96	49.96	49.95	40	76.38	71.1	79.13	**81.92**
P8	300	69.89	58.78	70.69	75.14	324	71.31	58.35	70.77	**80.99**
P9	11	49.83	49.83	49.79	49.87	40	60.54	49.43	58.83	**82.29**

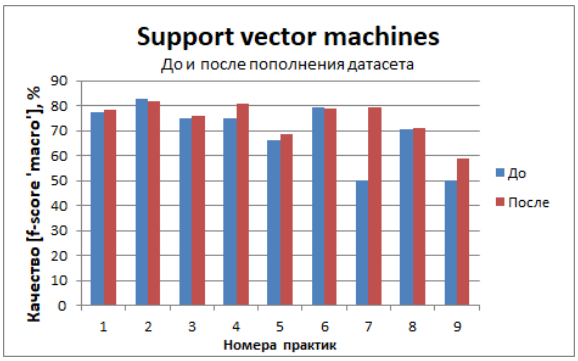
Модель RuBERT, дообученная (fine-tuned) на текстах дополненного корпуса, показала наилучшее среднее качество классификации текстов, содержащих упоминания практик, с помощью подхода «один против остальных» (81.32%). Исходя из этого, данная модель была выбрана для реализации прототипа приложения для поиска упоминаний экологических практик в текстах социальных



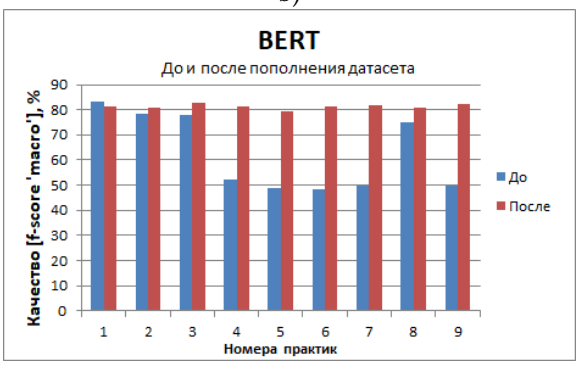
a)



b)



c)



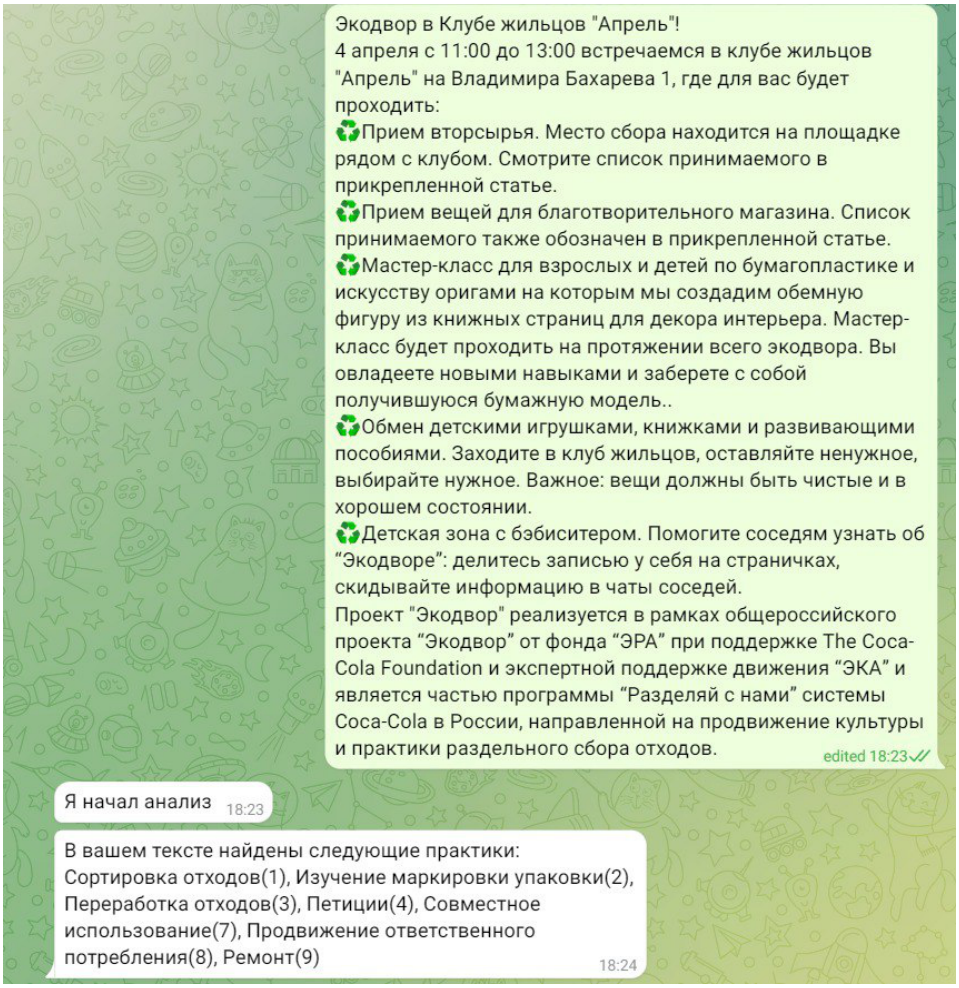
d)

Fig. 4. The difference of the results obtained for the original and augmented corpora: a) Logistic Regression, b) Random Forest, c) Support Vector Machines, d) Conversational RuBERT

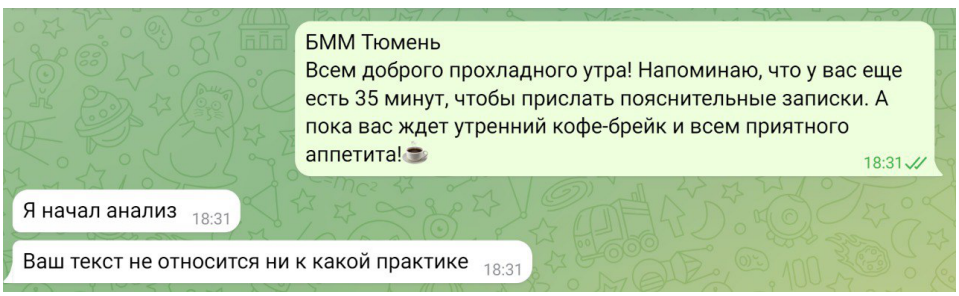
Рис. 4. Различие в результатах, полученных на исходном и дополненном корпусах: a) логистическая регрессия, b) случайный лес, c) метод опорных векторов, d) Conversational RuBERT

сетей. Прототип приложения[5], реализованный в форме чат-бота Telegram, определяет наличие в тексте упоминаний рассмотренных в данной работе девяти типов экологических практик. Когда текст отправляется боту, пользователь получает сообщение о том, что начался анализ текста, в свою очередь бот поочередно запускает девять обученных классификаторов, что занимает в сумме около 63 секунд. В дальнейшем планируется проведение мероприятий по оптимизации работы сервиса для уменьшения временных затрат на получение прогнозов классификаторов. По окончании анализа выводится информация о том, какие практики были найдены в тексте (рисунок 5). Помимо основного функционала, чат-бот содержит кнопки «О проекте», «Команда проекта». Из каждого состояния «беседы» с ботом можно вернуться в главное меню или на шаг назад. Для реализации чат-бота был выбран язык программирования Python и библиотека pyTelegramBotAPI[6].

[5] https://t.me/dlya_proekta_bot
[6] https://github.com/eternnoir/pyTelegramBotAPI



a)



b)

Fig. 5. The output of the prototype: a) the mentions of seven green practices have been found, b) no practice mentions have been found

Рис. 5. Результаты работы прототипа приложения: а) в тексте найдены упоминания семи экологических практик, b) в тексте не найдены упоминания практик

Заключение

Данная работа посвящена решению задачи поиска упоминаний экологических практик в текстах социальных сетей. Автоматизация поиска упоминаний экологических практик позволит, анализируя большие объемы контента социальных сетей, делать выводы о текущей распространенности различных видов экологических практик, эффективности существующих способов их внедрения

и, как следствие, возможных путях масштабирования уже имеющихся практик. Решение данной задачи является значимым с социальной точки зрения, поскольку экологические практики выполняют функцию важнейшего инструмента защиты окружающей среды, сокращения потребления и переработки отходов.

В рамках данного исследования впервые было проведено сравнение эффективности нескольких моделей машинного обучения для задачи поиска упоминаний экологических практик. Задача поиска экологических практик была сформулирована как задача бинарной классификации текстов с помощью подхода «один против остальных». Работа выполнена на корпусе текстов социальной сети ВКонтакте, снабженном экспертной разметкой упоминаний экологических практик различных типов. Для уменьшения влияния несбалансированности исходных данных был предложен подход к полуавтоматическому дополнению корпуса текстов, заключающийся в выделении наиболее частотных слов для каждой экологической практики, сборе дополнительных текстов, содержащих выделенные частотные слова, с помощью существующих сервисов для анализа социальных сетей и экспертной проверке собранных текстов. В результате сравнения моделей машинного обучения лучшее качество классификации было получено с помощью лингвистической модели Conversational RuBERT, дообученной с помощью дополненного корпуса текстов. Результат работы представлен в виде прототипа приложения – чат-бота Telegram для поиска упоминаний экологических практик в тексте на естественном языке.

В рамках дальнейшей работы будут протестированы подходы к генерации дополнительных текстов для миноритарных видов экологических практик (в частности, с помощью генерации текстов с использованием RuGPT-3[7] и перефразирования [60]) и извлечению точных фрагментов (span detection), содержащих упоминания практик, а также опробованы различные подходы к дообучению моделей (в частности, сопоставительное дообучение по аналогии с работой [61]). Кроме того, дальнейшие планы включают в себя завершение разработки приложения и дополнение его функционала.

References

[1] O. Zakharova, I. Pupysheva, T. Payusova, A. Zakharov, and S. L., “Green Values in Crowdfunding Projects”, *Glocalism*, no. 1, p. 6, 2021. DOI: 10.12893/gjcpi.2021.1.6.

[2] *VCIOM. Jekologicheskaja povestka: za desjat’ mesjacev do vyborov v Gosdumu (analiticheskij doklad). 2020-12-30*, http://www.wciom.ru, Accessed: 2021-03-18.

[3] Y. V. Ermolaeva and M. V. Rybakova, “Civil social practices of waste recycling in Russia (Moscow and Kazan)”, *IIOAB Journal*, vol. 10, no. S1, pp. 153–156, 2019.

[4] O. Zakharova, T. Payusova, I. Akhmedova, and L. Suvorova, “Green Practices: Ways to Investigation”, *Sotsiologicheskie issledovaniya*, no. 4, pp. 25–36, 2021. DOI: 10.31857/S013216250012084-5.

[5] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, “Detection and resolution of rumours in social media: A survey”, *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–36, 2018. DOI: 10.1145/3161603.

[6] D. Rogers, A. Preece, M. Innes, and I. Spasić, “Real-time text classification of user-generated content on social media: Systematic review”, *IEEE Transactions on Computational Social Systems*, 2021. DOI: 10.1109/TCSS.2021.3120138.

[7] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, “A Survey on Text Classification: From Traditional to Deep Learning”, *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1–41, 2022. DOI: 10.1145/3495162.

[7] https://github.com/ai-forever/ru-gpts

[8] F. C. Permana, Y. Rosmansyah, and A. S. Abdullah, "Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter Social Media", in *Journal of Physics: Conference Series*, IOP Publishing, vol. 893, 2017, p. 012 051. DOI: 10.1088/1742-6596/893/1/012051.

[9] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naive Bayes, decision tree, and random forest algorithm", *Procedia Computer Science*, vol. 161, pp. 765–772, 2019. DOI: 10.1016/j.procs.2019.11.181.

[10] N. R. Fatahillah, P. Suryati, and C. Haryawan, "Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech", in *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, IEEE, 2017, pp. 128–131. DOI: 10.1109/SIET.2017.8304122.

[11] K. K. Kiilu, G. Okeyo, R. Rimiru, and K. Ogada, "Using Naive Bayes algorithm in detection of hate tweets", *International Journal of Scientific and Research Publications*, vol. 8, no. 3, pp. 99–107, 2018. DOI: 10.29322/IJSRP.8.3.2018.p7517.

[12] Z. Peng, Q. Hu, and J. Dang, "Multi-kernel SVM based depression recognition using social media data", *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 43–57, 2019. DOI: 10.1007/s13042-017-0697-1.

[13] P. Karthika, R. Murugeswari, and R. Manoranjithem, "Sentiment analysis of social media network using random forest algorithm", in *2019 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)*, IEEE, 2019, pp. 1–5. DOI: 10.1109/INCOS45849.2019.8951367.

[14] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM", in *2015 International Conference on Data and Software Engineering (ICoDSE)*, IEEE, 2015, pp. 170–174. DOI: 10.1109/ICODSE.2015.7436992.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain, "Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis", *Cognitive Computation*, vol. 10, no. 4, pp. 639–650, 2018. DOI: 10.1007/s12559-018-9549-x.

[17] M. Tripathi, "Sentiment analysis of Nepali COVID19 tweets using NB SVM and LSTM", *Journal of Artificial Intelligence*, vol. 3, no. 03, pp. 151–168, 2021. DOI: 0.36548/jaicn.2021.3.001.

[18] R. Monika, S. Deivalakshmi, and B. Janet, "Sentiment analysis of US airlines tweets using LSTM/RNN", in *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, IEEE, 2019, pp. 92–95. DOI: 10.1109/IACC48062.2019.8971592.

[19] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets", in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760. DOI: 10.1145/3041021.3054223.

[20] A. Bisht, A. Singh, H. Bhadauria, J. Virmani, *et al.*, "Detection of hate speech and offensive language in Twitter data using LSTM model", in *Recent trends in image and signal processing in computer vision*, Springer, 2020, pp. 243–264. DOI: 10.1007/978-981-15-2740-1_17.

[21] V. Rupapara, F. Rustam, A. Amaar, P. B. Washington, E. Lee, and I. Ashraf, "Deepfake tweets classification using stacked Bi-LSTM and words embedding", *PeerJ Computer Science*, vol. 7, e745, 2021. DOI: 10.7717/peerj-cs.745.

[22] A. Wani, I. Joshi, S. Khandve, V. Wagh, and R. Joshi, "Evaluating deep learning approaches for COVID19 fake news detection", in *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, Springer, 2021, pp. 153–163. DOI: 10.1007/978-3-030-73696-5_15.

[23] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification", in *Twenty-ninth AAAI conference on artificial intelligence*, 2015. DOI: 10.5555/2886521.2886636.

[24] S. Bansal, "A Mutli-Task Mutlimodal Framework for Tweet Classification Based on CNN (Grand Challenge)", in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, IEEE, 2020, pp. 456–460. DOI: 10.1109/BigMM50055.2020.00075.

[25] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis", *Future Generation Computer Systems*, vol. 115, pp. 279–294, 2021. DOI: 10.1016/j.future.2020.08.005.

[26] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Dimensional sentiment analysis using a regional CNN-LSTM model", in *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2016, pp. 225–230. DOI: 10.18653/v1/P16-2037.

[27] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "A combined CNN and LSTM model for Arabic sentiment analysis", in *International cross-domain conference for machine learning and knowledge extraction*, Springer, 2018, pp. 179–191. DOI: 10.1007/978-3-319-99740-7_12.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach", *arXiv preprint arXiv:1907.11692*, 2019. DOI: 10.48550/arXiv.1907.11692.

[31] A. El Mahdaouy, A. El Mekki, K. Essefar, A. Skiredj, and I. Berrada, "CS-UM6P at SemEval-2022 Task 6: Transformer-based Models for Intended Sarcasm Detection in English and Arabic", in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022, pp. 844–850. DOI: 10.18653/v1/2022.semeval-1.117.

[32] M. Du, S. D. Gollapalli, and S.-K. Ng, "NUS-IDS at CheckThat! 2022: Identifying Check-worthiness of Tweets using CheckthaT5", *Working Notes of CLEF*, 2022.

[33] A. Glazkova, M. Glazkov, and T. Trifonov, "g2tmn at constraint@ aaai2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection", in *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, Springer, 2021, pp. 116–127. DOI: 10.1007/978-3-030-73696-5_12.

[34] Y. Rubtsova, "Constructing a corpus for sentiment classification training", *Software & Systems*, no. 1 (109), pp. 72–78, 2015. DOI: 10.15827/0236-235X.109.072-078.

[35] I. Bolshakova and K. Lagutina, "Avtomaticheskaja klassifikacija tekstov na russkom jazyke s pomoshh'ju tonal'nogo slovarja", no. 14, pp. 6–13, 2022.

[36] A. Kotelnikova, D. Paschenko, and E. Razova, "Lexicon-based methods and BERT model for sentiment analysis of Russian text corpora", in *CEUR Workshop Proceedings*, 2021, pp. 73–81.

[37] N. Loukachevitch and Y. Rubtsova, "SentiRuEval-2016: overcoming time gap and data sparsity in tweet sentiment analysis", in *Computational Linguistics and Intellectual Technologies*, 2016, pp. 416–426.

[38] A. Chernyaev, A. Spryiskov, A. Ivashko, and Y. Bidulya, "A rumor detection in Russian tweets", in *International Conference on Speech and Computer*, Springer, 2020, pp. 108–118. DOI: 10.1007/978-3-030-60276-5_11.

[39] E. Mikhalkova, Y. Karyakin, and I. Glukhikh, "Large Scale Retrieval of Social Network Pages by Interests of Their Followers", in *Computational Science – ICCS 2018*, Cham: Springer International Publishing, 2018, pp. 234–246. DOI: 10.1007/978-3-319-93698-7_18.

[40] E. Pronoza, P. Panicheva, O. Koltsova, and P. Rosso, "Detecting ethnicity-targeted hate speech in Russian social media texts", *Information Processing & Management*, vol. 58, no. 6, p. 102 674, 2021, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2021.102674.

[41] K. V. Lagutina, N. S. Lagutina, and E. I. Boychuk, "Text classification by genre based on rhythm features", *Modeling and analysis of information systems*, pp. 280–291, 2021. DOI: 10.18255/1818-1015-2021-3-280-291.

[42] K. Svetlov and K. Platonov, "Sentiment analysis of posts and comments in the accounts of Russian politicians on the social network", in *2019 25th Conference of Open Innovations Association (FRUCT)*, IEEE, 2019, pp. 299–305. DOI: 10.23919/FRUCT48121.2019.8981501.

[43] I. Kozitsin, A. Chkhartishvili, A. Marchenko, D. Norkin, S. Osipov, I. Uteshev, V. Goiko, R. Palkin, and M. Myagkov, "Modeling political preferences of Russian users exemplified by the social network Vkontakte", *Mathematical Models and Computer Simulations*, vol. 12, no. 2, pp. 185–194, 2020. DOI: 10.1134/S2070048220020088.

[44] P. Basina, V. Goiko, E. Petrov, and V. Bakulin, "Classification community publications of the "VKontakte" for assessing the quality of life of the population", *Computational Linguistics and Intellectual Technologies*, p. 18, 2022. DOI: 10.28995/2075-7182-2022-21-1001-1016.

[45] A. Sboev, I. Moloshnikov, A. Naumov, A. Levochkina, and R. Rybka, "The Russian Language Corpus and a Neural Network to Analyse Internet Tweet Reports About COVID-19", *PoS*, vol. DLCP2021, p. 017, 2021. DOI: 10.22323/1.410.0017.

[46] M. J. Farrell, L. Brierley, A. Willoughby, A. Yates, and N. Mideo, "Past and future uses of text mining in ecology and evolution", *Proceedings of the Royal Society B*, vol. 289, no. 1975, p. 20 212 721, 2022. DOI: 10.1098/rspb.2021.2721.

[47] S. C. Anderson, P. R. Elsen, B. B. Hughes, R. K. Tonietto, M. C. Bletz, D. A. Gill, M. A. Holgerson, S. E. Kuebbing, C. McDonough MacKenzie, M. H. Meek, *et al.*, "Trends in ecology and conservation over eight decades", *Frontiers in Ecology and the Environment*, vol. 19, no. 5, pp. 274–282, 2021. DOI: 10.1002/fee.2320.

[48] J. Knott, E. LaRue, S. Ward, E. McCallen, K. Ordonez, F. Wagner, I. Jo, J. Elliott, and S. Fei, "A roadmap for exploring the thematic content of ecology journals", *Ecosphere*, vol. 10, no. 8, e02801, 2019. DOI: 10.1002/ecs2.2801.

[49] F. R. Dayeen, A. S. Sharma, and S. Derrible, "A text mining analysis of the climate change literature in industrial ecology", *Journal of Industrial Ecology*, vol. 24, no. 2, pp. 276–284, 2020. DOI: 10.1111/jiec.12998.

[50] F. Romero-Perdomo, J. D. Carvajalino-Umaña, J. L. Moreno-Gallego, N. Ardila, and M. Á. González-Curbelo, "Research Trends on Climate Change and Circular Economy from a Knowledge Mapping Perspective", *Sustainability*, vol. 14, no. 1, p. 521, 2022. DOI: 10.3390/su14010521.

[51] O. J. Luiz, J. D. Olden, M. J. Kennard, D. A. Crook, M. M. Douglas, T. M. Saunders, and A. J. King, "Trait-based ecology of fishes: A quantitative assessment of literature trends and knowledge gaps using topic modelling", *Fish and Fisheries*, vol. 20, no. 6, pp. 1100–1110, 2019. DOI: 10.1111/faf.12399.

[52] R. Cornford, S. Deinet, A. De Palma, S. L. Hill, L. McRae, B. Pettit, V. Marconi, A. Purvis, and R. Freeman, "Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets", *Global Ecology and Biogeography*, vol. 30, no. 1, pp. 339–347, 2021. DOI: 10.1111/geb.13219.

[53] N. Le Guillarme and W. Thuiller, "TaxoNERD: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature", *Methods in Ecology and Evolution*, vol. 13, no. 3, pp. 625–641, 2022. DOI: 10.1111/2041-210X.13778.

[54] N. T. Nguyen, R. S. Gabud, and S. Ananiadou, "COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature", *Biodiversity data journal*, no. 7, 2019. DOI: 10.3897/BDJ.7.e29626.

[55] R. Bossy, L. Deléger, E. Chaix, M. Ba, and C. Nédellec, "Bacteria biotope at BioNLP open shared tasks 2019", in *Proceedings of the 5th workshop on BioNLP open shared tasks*, 2019, pp. 121–131. DOI: 10.18653/v1/D19-5719.

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in Python", *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[57] Y. Kuratov and M. Arkhipov, "Adaptation of deep bidirectional multilingual transformers for Russian language", in *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 2019, pp. 333–339.

[58] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles", 2016.

[59] T. Shavrina and O. Shapovalova, "To the methodology of corpus construction for machine learning: "Taiga" syntax tree corpus and parser", *Proceedings of the "Corpora"*, pp. 78–84, 2017.

[60] A. Fenogenova, "Russian paraphrasers: Paraphrase with transformers", in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, 2021, pp. 11–19.

[61] I. Bondarenko, "Contrastive fine-tuning to improve generalization in deep NER", 2022. DOI: 10.28995/2075-7182-2022-21-70-80.