Использование машинного обучения для изучения качества жизни населения: методологические аспекты

УДК 303.7 DOI 10.26425/2658-347X-2022-5-1-87-97

Получено 17.12.2021 Доработано после рецензирования 18.01.2022 Принято 02.02.2022

Щекотин Евгений Викторович

Канд. филос. наук, доц., зав. лаб. ФГБОУ ВО «Новосибирский государственный университет экономики и управления «НИНХ», г. Новосибирск, Российская Федерация

ORCID: 0000-0001-7377-0645

E-mail: evgvik1978@mail.ru

Гойко Вячеслав Леонидович

Зав. лаб., ФГАОУ ВО «Национальный исследовательский Томский государственный университет», г. Томск, Российская Федерация

ORCID: 0000-0002-5985-3724

E-mail: goiko@data.tsu.ru

Басина Полина Александровна

Аналитик, ФГАОУ ВО «Национальный исследовательский Томский государственный университет», г. Томск, Российская Федерация

ORCID: 0000-0001-7904-7394

E-mail: polya.basina@yandex.ru

Бакулин Вячеслав Викторович

Аналитик, ФГАОУ ВО «Национальный исследовательский Томский государственный университет», г. Томск, Российская Федерация

ORCID: 0000-0003-2073-6341

E-mail: slava38710505@gmail.com

АННОТАЦИЯ

Оценка качества жизни населения является важной и актуальной задачей социологии. Машинное обучение, как инструмент классификации цифровых следов пользователей социальных сетей, позволяет сформировать базу для расчета индекса субъективного качества жизни. В статье последовательно рассмотрены все этапы применения алгоритмов машинного обучения для оценки качества жизни населения регионов Российской Федерации и вопросы повышения точности работы нейронной сети. Для обучения нейросети авторами был сформирован набор размеченных данных, извлеченных из региональных сообществ социальная сеть «ВКонтакте». Проанализированы различные подходы к векторизации текстов, общедоступные нейросетевые модели, предобученные на больших русскоязычных текстовых корпусах, а также метрики оценки результатов работы алгоритмов. Проведены вычислительные эксперименты с разными алгоритмами, по результатам которых был выбран алгоритм Rubert-tiny в связи с его высокой скоростью обучения и классификации. В ходе настройки параметров модели была достигнута точность *f*1-macro 0,545. Вычислительные эксперименты проводились с использованием скриптов на языке Python. Рассмотрены типичные ошибки, которые совершает нейронная сеть в процессе автоматической классификации контента. Результаты исследования можно использовать для расчета индекса онлайн-активности в социальной сети «ВКонтакте» пользователей из различных российских регионов, на основе которого в дальнейшем можно рассчитывать индекс субъективного качества жизни. Повышение точности работы нейронной сети позволит получить более надежные данные для оценки качества жизни в регионах на основе цифровых следов пользователей.

Ключевые слова

Качество жизни, благополучие, цифровые методы, нереактивные методы, цифровые следы, социальные сети, ВКонтакте, машинное обучение, классификации текстов

Для цитирования

Щекотин Е.В., Гойко В.Л., Басина П.А., Бакулин В.В. Использование машинного обучения для изучения качества жизни населения: методологические аспекты // Цифровая социология. 2022. Т. 5, № 1. С. 87–97.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-011-00391.

Using machine learning to study the population life quality: methodological aspects

Received 17.12.2021 Revised 18.01.2022 Accepted 02.02.2022

Evgeniy V. Shchekotin

Cand. Sci. (Philos.), Assoc. Prof., Head of the laboratory, Novosibirsk State University of Economics and Management, Novosibirsk, Russia

ORCID: 0000-0001-7377-0645

E-mail: evgvik1978@mail.ru

Vyacheslav L. Goiko

Head of the laboratory, National Research Tomsk State University, Tomsk, Russia

ORCID: 0000-0002-5985-3724

E-mail: goiko@data.tsu.ru

Polina A. Basina

Analyst, National Research Tomsk State University, Tomsk, Russia

ORCID: 0000-0001-7904-7394

E-mail: polya.basina@yandex.ru

Vyacheslav V. Bakulin

Analyst, National Research Tomsk State University, Tomsk, Russia

ORCID: 0000-0003-2073-6341

E-mail: slava38710505@gmail.com

ABSTRACT

Assessment of the population life quality is an important and relevant sociological task. Machine learning as a classification tool of social network users' digital traces makes it possible to create a base to calculate subjective life quality index. The article consistently reviews all stages of the machine learning algorithms application to assess the life quality of the population of the regions of the Russian Federation and the issues of improving neural network accuracy. To train the neural network the authors formed a set of marked-up data extracted from regional communities of the social network "VKontakte". Various approaches to text vectorisation, publicly available neural network models pre-trained on large Russian-language text corpora, as well as metrics for evaluating the algorithms results were analysed. Computational experiments with different algorithms were carried out, according to the results of which the Rubert-tiny algorithm was selected due to its high learning and classification rate. During the model parameters adjustment, the accuracy of f1-macro 0.545 was achieved. Computational experiments were carried out using Python scripts.Typical errors that a neural network makes in the process of automatic content classification were considered. The results of the study can be used to calculate the online activity index in the VKontakte social network of users from various Russian regions, on the basis of which the subjective life quality index will be calculated in the future. Improving the neural network accuracy will make it possible to obtain more reliable data for assessing the life quality in Russian regions based on users' digital traces.

Keywords

Life quality, well-being, digital methods, non-reactive methods, digital traces, social networks, VKontakte, machine learning, text classifications

For citation

Shchekotin E.V., Goiko V.L., Basina P.A., Bakulin V.V. (2022) Using machine learning to study the population life quality: methodological aspects. *Digital sociology*, vol. 5, no 1, pp. 87–97. DOI: 10.26425/2658-347X-2022-5-1-87-97

Acknowledgements

The reported study was funded by the Russian Foundation for Basic Research as a part of scientific project No. 20-011-00391.

ВВЕДЕНИЕ / INTRODUCTION

Стремительное развитие цифровой социологии в последние два-три года стимулирует применение широкого спектра цифровых методов для решения все новых исследовательских задач. В данной статье представлен методологический подход к использованию алгоритмов машинного обучения для изучения качества жизни населения регионов Российской Федерации. В нашем исследовании применяется метод машинного обучения с учителем. Одной из задач, которые решаются с помощью машинного обучения, является классификация, в том числе автоматическая классификация текстов из Интернета [Dawson, 2019]. В дальнейшем изложении будет рассмотрено как можно применять алгоритмы машинного обучения для реализации такой социологической задачи как оценка качества жизни населения, где в качестве источника данных используются цифровые следы пользователей в социальной сети «ВКонтакте».

Оценка качества жизни населения является важной и актуальной задачей для социологии, так как повышение качества жизни является одной из основополагающих функций государственного управления [Чичканов, Васильева, 2014]. Уже в 1960–70-е гг. сформировались два основных подхода к решению этой задачи, которые различаются в том, какие источники социологической информации используются для решения этой задачи. Это объективный и субъективный подходы [McGillivray & Clarke (eds), 2006]. В первом случае источников информации служат статистические данные, которые собираются различными организациями, во втором измерение осуществляется на основе субъективных оценок людей, которые фиксируются при помощи традиционных социологических методов (опрос, фокус-группы, интервью и т.п.).

Вместе с тем, по мере усиления цифровизации и увеличение доступности Интернета для широких масс населения, у социологов возникла идея использовать в качестве источника данных цифровые следы, которые пользователи оставляют в онлайн-пространстве [Щекотин, 2021]. Чаще всего в качестве источника информации о настроениях и мнения людей используются социальные сети (Facebook, «ВКонтакте») и мессенджеры (Twitter), но также в роли такого источника могут выступать поисковые запросы пользователей и оцифрованные литературные материалы. Изучение цифровых следов пользователей относится к числу так называемых «нереактивных» или «незаметных» методов социологического исследования, которые появились еще в 1960-х гг. [Николаенко, Федорова, 2017]. Суть этих методов заключается в том, чтобы в процессе исследования избегать взаимодействия между исследователем и респондентом. О. В. Крыштановская [2018] называет такие методы «бесконтактной социологией». Ее особенность в том, что «можно не выполнять традиционные манипуляции, если люди в социальных сетях и блогах и так высказываются относительно всех политических событий. Можно не спрашивать их об этом, если они сами пишут свое мнение, не ожидая анкет социологов».

Для изучения цифровых следов пользователей в Интернете социологами необходимо решить, по меньшей мере, две непростые в техническом отношении задачи – это сбор и анализ данных, что в свою очередь требует обращения к методам компьютерных наук. Использование компьютерных технологий для решения таких стандартных для социологии задач как сбор и анализ данных (нужно подчеркнуть, что речь идет об интеллектуальном анализе данных, а не о традиционном статистическом анализе) вызывает вполне закономерную настороженность представителей социологии, так как представители компьютерных наук опираются на иные стандарты достоверности научного знания, отличные от «золотого стандарты», сложившегося в социологии в 1930–40-е гг. [Schober et al, 2016]. Еще в 2015 г. Ю.Н. Толстова справедливо замечала, что социологи «социологи игнорируют новые IT» [Толстова, 2015], сегодня ситуация меняется, все больше исследователей обращаются к цифровой социологии.

Новые возможности, которые открываются перед социологами в связи с применением компьютерных технологии, и их недостатки в сравнении с традиционными социологическими методами достаточно хорошо отрефлексированы в научной литературе, поэтому не будем акцентировать внимание на этом аспекте. Можно отослать к статье М. Б. Богданова и И. Б. Смирнова, в которой детально рассмотрены возможности и ограничения изучения цифровых следов и машинного обучения в социологическом исследовании [Богданов, Смирнов, 2021]. Мы в данной статье сосредоточимся на описании конкретного исследовательского кейса, связанного с применением алгоритмов машинного обучения в целях оценки качества жизни.

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ / RESEARCH METHODOLOGY

Важнейшим источником цифровых следов являются социальные сети. Согласно данным ежегодного отчета Global Digital 2021, социальными

сетями пользуется 67,8 % населения России (99 млн чел.). В России среди социальных сетей наиболее популярной является сеть «ВКонтакте». Так, по данным Левада-центра (внесен в реестр организаций, выполняющих функции иностранного агента) три четверти россиян являются пользователями социальных сетей, из них 43 % посещают «ВКонтакте»[1]. Согласно исследованию ВЦИОМа 61 % опрошенных пользуются «ВКонтакте» не менее одного раза в полгода, при этом 49 % посещают эту социальную сеть ежедневно или несколько раз в неделю[2]. Поэтому вполне логично, что мы обратились именно к этой социальной сети как источнику цифровых следов пользователей.

Для получения репрезентативных данных, позволяющих учесть специфику регионов, в качестве источников были выбраны так называемые «региональные сообщества», которые обладают следующими характеристиками:

– не менее 50 % подписчиков сообщества должны быть из одного региона (регион определяется посредством указанного пользователем места проживания);

– содержат в себе посты о социальном, экономическом и политическом положении в регионе;

– публикуются сообщения подписчиков, содержащие информацию о социальной, экономической и политической сфере;

– часть публикуемых в сообществе постов содержат в себе эмоциональную оценку (позитивную или негативную) событий и новостей. Более подробно механизм сбора данных и некоторые результаты исследования описан в наших прошлых публикациях [Щекотин и др., 2020].

Например, в рамках представленной выше концепции к региональным относятся сообщества: «Выбирай в Серпухове» (https://vk.com/gc_vibiray, Московская область), «[Белоярский Online]» (https://vk.com/beloyarsky_online, Свердловская область), «Балашов онлайн» (https://vk.com/balashov_0nline, Саратовская область), «Вестник Челябинска» (https://vk.com/novostichel, Челябинская область), «Елецкий мкрн. г. Липецка» (https://vk.com/eletskiy48, Липецкая область).

Как было отмечено выше, цифровые следы социальной сети представляют собой большие данные, исследовать которые традиционными методами социологии невозможно. Таким образом, возникает необходимость разработки алгоритма, позволяющего автоматически классифицировать тестовые публикации «ВКонтакте». Учитывая специфику постов, следует решить 2 задачи: определение релевантных сообщений (очистка от «мусора») и категоризация полученных текстов (в нашем случае, это выделение индикаторов качества жизни). В рамках данного исследования были выделены следующие показатели качества жизни: «образование», «здравоохранение», «безопасность», «социальное обеспечение», «работа органов власти», «экология» и «доступность товаров и услуг». В качестве единиц анализа выступали сообщения (посты), публикуемые в региональных сообществах. Для каждого показателя были составлены списки маркерных слов и тем, наличие которые позволяет однозначно отнести сообщение к конкретному показателю качества жизни или же к «мусору».

Процедура классификации текстов состоит из нескольких этапов – предварительная обработка данных, векторизация обработанных текстов (извлечение признаков), построение классификатора и оценка полученных результатов [Двойникова, Карпов, 2020] (рис. 1). Для построения классификационных моделей был использован скриптовый язык программирования Python.

Для обучения и оценки алгоритма классификации был сформирован набор размеченных данных – 84 000 постов из социальной сети «ВКонтакте». В наборе присутствует сильный дисбаланс классов (рис. 2). Большинство сообщений являются нерепрезентативными («мусором»); наиболее крупные категории – «безопасность», «ЖКХ и инфраструктура» и «здравоохранение»; наименее – «доступность товаров и услуг», «социальное обеспечение» и «экология».

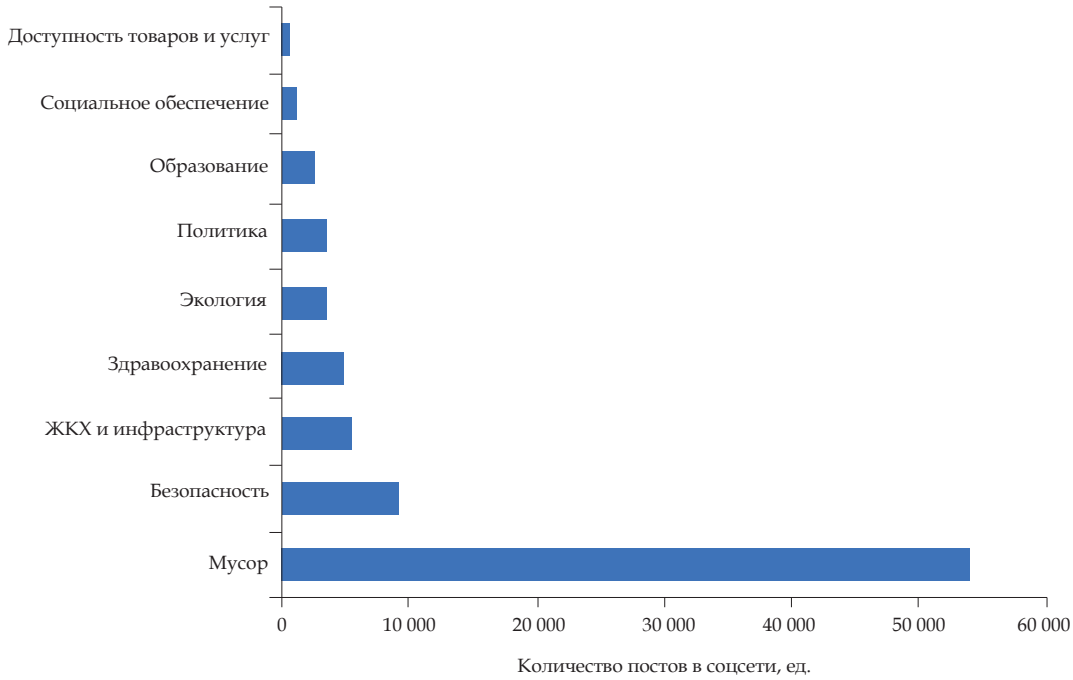
Источник / *Source: [Двойникова, Карпов, 2020]*

Рис. 1. **Процедура классификации текстов**

Figure 1. Text classification procedure

[1] *Левада-Центр* (2021). Социальные сети в России. Режим доступа: https://www.levada.ru/2021/02/23/sotsialnye-seti-v-rossii/ (дата обращения: 09.12.2021).

[2] *ВЦИОМ* (2021). Медиапотребление и активность в интернете. Режим доступа: https://wciom.ru/analytical-reviews/analiticheskii-obzor/mediapotreblenie-i-aktivnost-v-internete (дата обращения: 09.12.2021).



Составлено авторами по материалам исследования / *Compiled by the authors based on the research materials*

***Рис. 2.* Распределение категорий в размеченном наборе данных по количеству сообщений**

Figure 2. Distribution of categories in the marked-up data set by the number of messages

Предварительная обработка позволяет определить релевантную информацию (убрать «шум»); при ее отсутствии вероятность неверных результатов в модели машинного обучения значительно возрастает. В данном исследовании были выполнены следующие шаги: очистка текстов от знаков препинания, смайликов, нерелевантных слов с использованием регулярных выражений; перевод всех слов в нижний регистр; удаление стоп-слов (предлоги, междометия, числа, частицы и т.д.), что позволяет очистить данные от часто встречающихся слов, которые не имеют специфического смысла.

Следующий этап заключается в преобразовании документов к виду, пригодному для алгоритмов машинного обучения – отображение текстов в виде векторов. В текстовых данных в качестве признаков могут выступать – слова и словосочетания, символы и их последовательности, предложения и т.д. На сегодняшний день сложилось множество различных подходов, которые обладают своими преимуществами и недостатками [Двойникова, Карпов, 2020]. В рамках данной работы рассмотрим несколько наиболее распространенных методов. Прежде всего следует отметить классические методы, среди которых выделяют:

1) «Мешок слов» (Bag of Words) – данный метод позволяет представлять текст в виде набора, присутствующих в нем слов. Признаковое пространство состоит из всех слов, которые встретились в данных, а в качестве значений признаков используется оценка встречаемости слова в конкретном текстовом документе [Soumya & Joseph, 2014];

2) Метод TF-IDF представляет собой статистическую меру, которая используется для определения важности слова относительно одного документа и коллекции документов в целом. Вес каждого слова пропорционален количеству раз, когда это слово используется в документе, и обратно пропорционален частоте использования этого слова в других документах коллекции [Jones, 2004];

3) One-hot encoding (прямое кодирование) – метод, который преобразует слова в бинарные векторы; размер каждого вектора слова равен объему всех слов в тексте и состоит из 0 и 1, где 1 соответствует кодируемому слову. Для представления фразы необходимо объединить векторы слов, из которых она состоит [Potdar et al, 2017].

В качестве недостатков рассмотренных методов следует отметить отсутствие информации о порядке слов и их связях, невозможности представления неизвестных слов (отсутствующих в обучающем корпусе), а также разреженность полученных векторов. Эмбеддинги слов позволяют снять указанные ограничения – вектора имеют фиксированную размерность, которая не зависит от объема используемых в данных слов, и учитывают контекст использования слов. Эмбеддинг – векторное представление слова, которое получается в процессе длительного обучения нейросетевой модели на больших корпусах текстов.

В машинном обучение встречается практика, когда нейронная сеть со скрытым слоем нейронов учится решать некоторую задачу; однако, впоследствии нейросетевая модель не используется для этой задачи. В рассматриваемом нами случае нейросети обучаются на «фальшивых» задачах (например, угадывание следующего слова в предложении), в процессе чего происходит заполнение и корректировка матрицы весов нейронной сети; на основе полученных весов в последствии строятся эмбеддинги слов. Полученные представления слов (вектора) могут использоваться для решения задач классификации алгоритмами машинного обучения; однако, современные нейросетевые модели могут не только создавать эмбеддинги, но и сразу строить классификаторы на их основе. В рамках данной работы рассмотрим общедоступные модели, предобученные на больших русскоязычных текстовых корпусах [Kutuzov & Kuzmenko, 2017].

1. Word2vec – нейросеть, в которой существует два основных алгоритма обучения: CBOW и Skipgram. «Фальшивая» задача архитектуры CBOW – предсказать текущее слово, исходя из окружающего его контекста; архитектура Skipgram использует текущее слово для того, чтобы определить окружающие его слова. После обучения модели используются семантические вектора слов из скрытого слоя нейросети. В отличии от классических способов представления текстов, Word2vec учитывает встречаемость слов в одинаковых контекстах, а не частоту их появления; слова, используемые в одном контексте, будут иметь похожие векторы. Однако, каждое слово будет иметь только один вектор [Mikolov et al, 2013a; Mikolov et al, 2013b].

2. Glove – основная идея метода заключается в способе подсчета частоты появления слов в текстовом корпусе. Фактически он состоит из двух основных этапов: на первом происходит построение матрицы смежности из обучающего корпуса, а на втором – факторизация матрицы для получения векторов. Созданные представления демонстрируют интересные линейные подструктуры векторного пространства слов [Pennington et al, 2014].

3. FastText является развитием модели Word2vec; в качестве отличий следует отметить, что теперь используются не только векторы слов, но и векторы *n*-грам. Например, при $n = 3$, вектор для слова «вода» будет представлен суммой векторов, состоящих из следующих триграмм: «<во», «вод», «ода», «да>» («<» и «>» – начало и конец слова). Данный алгоритм позволяет получать вектора для слов, которые отсутствовали в словаре при обучении или содержат ошибки и опечатки [Joulin et al, 2016].

4. BERT – нейронная сеть, основанная на архитектуре трансформер. В отличие от других моделей, BERT обучает контекстно-зависимые представления слов. Например, в случае обучения модели Word2vec для многозначных слов будет получен один эмбеддинг слова. Модель BERT учитывает окружающий контекст предложения и генерирует различные эмбеддинги для многозначных слов. Идея векторизации в BERT заключается в том, что каждому слову из текста присваивается число, обозначающее порядковый номер слова в словаре, далее это число преобразуется в вектор из 512 символов. Словарь, который использует данная нейросеть, построен таким образом, что слова, близкие по смыслу, располагаются рядом. Тем самым нейронная сеть BERT векторизует текст, учитывая близость слов [Devlin et al, 2019].

5. ELMo – модель представления контекстно-зависимых слов. Основная идея заключается в том, чтобы сначала построить для каждого слова в тексте посимвольный эмбеддинг, после чего применить для них рекуррентную нейросеть LSTM; таким образом, полученные эмбеддинги учитывают контекст, в котором встретилось слово [Двойникова, Карпов, 2020; Peters et al, 2018].

Прежде чем перейти к алгоритмам классификации рассмотрим метрики оценки их результатов. Один из информативных способов оценки – матрица ошибок (матрица несоответствий). Размерность матрицы определяется количеством целевых меток (категорий); строки – фактические классы (правильные категорий), столбцы – спрогнозированные алгоритмом значения. Число в каждой ячейке демонстрирует количество или процентное соотношение примеров, когда спрогнозированный класс совпадает или отличен от фактического класса. Элементы главной диагонали в таблице 1, закрашенные серым цветом, соответствуют правильным результатам классификации; остальные значения, выделенные оранжевым цветов, показывают ошибочно классифицированные примеры [Мюллер, Гвидо, 2016].

***Таблица 1.* Пример матрицы ошибок**

Table 1. Example of a confusion matrix

Предсказанный класс / Истинный класс	Категория 1	Категория 2	Категория 3
Категория 1			
Категория 2			
Категория 3			

Источник / *Source: [Мюллер, Гвидо, 2016]*

В качестве обобщающих метрик информации, которые содержится в матрице неточностей, выступают правильность (англ. accuracy), точность (англ. precision) и полнота (англ. recall). Правильность – это количество верно классифицированных примеров среди всех примеров. Точность (англ. precision) показывает, сколько из спрогнозированных положительных примеров оказались действительно положительными; полнота (англ. recall) – сколько от общего числа фактических положительных примеров было предсказано как положительный класс. Однако, рассматриваемые метрики не позволяют получить полной картины, поэтому в качестве обобщающей оценки данных показателей используется *f*1-score, которая представляет собой гармоническое среднее точности и полноты. *F*1-мера используется при оценке несбалансированных наборов данных, при этом является трудно интерпретируемой метрикой. В случае задач мультиклассовой классификации *f*1-score рассчитывается для каждого класса и полученные значения усредняются: macro – все классы имеют одинаковый вес; weighted – рассчитывается вес каждого класса в зависимости от их представленности в выборке; micro усреднение позволяет присвоить одинаковый вес каждому примеру [Мюллер, Гвидо, 2016]. Учитывая присутствующий в данных дисбаланс и одинаковую важность каждого класса, для оценки и сравнения работы моделей нами была выбрана метрика *F*1-macro, которая позволяет обобщить метрики точности и полноты.

РЕЗУЛЬТАТЫ / RESULTS

Нами был учтен предыдущий опыт построения классификационной модели в рамках данного проекта, поэтому в качестве базового подхода была использована следующая связка – TF-IDF-представление текстов (векторизация) и градиентный бустинг (классификатор). Однако, обращаем внимание, что полученные результаты не могут быть сравнимы, так как классификационные модели отличаются количеством категорий и процедурой разметки данных. Градиентный бустинг представляет собой ансамблевый метод; его суть заключается в последовательном построении нескольких моделей машинного обучения (ансамбля), где каждая последующая модель стремится восполнить недостатки предыдущей [3]. Основная идея ансамблевых методов заключается в том, что на основе сочетания слабых моделей можно создать мощную модель [Chen & Guestrin, 2016]. В результате обучения была получена точность *f*1-macro – 0,47; данное значения является базовым, по сравнению с которым мы можем сравнивать другие подходы к решению нашей задачи.

С 2018 г. лидирующие позиции в обработке естественного языка занимает архитектура трансформер. Она основана на механизме внимания (attention), что позволяет модели обращать внимание на разные части текста и лучше понимать закономерности, необходимые для решения задачи. Результаты модели BERT показали значительный прирост по сравнению с предыдущими SOTA-решениями [Devlin et al, 2019]. В связи с этим, дальнейшие эксперименты проводились с реализацией BERT (табл. 2).

В данной работе мы сфокусировали внимание на улучшении точности алгоритма Rubert-tiny tuned, так как данная модель позволяет достигнуть баланса между точностью *f*1-macro и размером модели (количество слоев нейронной сети, которые влияют на скорость работы модели). Нам удалось достичь точности *f*1-macro – 0,545 (общая прогностическая способность модели) за счет экспериментов с оптимизатором (определяет оптимальный набор параметров модели (вес и смещение)), количеством эпох (эпоха представляет собой одну итерацию в процессе обучения модели) и параметром learning rate (коэффициент скорости обучения, который позволяет корректировать веса на каждой итерации.).

[3] *CatBoost* (2022). CatBoost is a high-performance open source library for gradient boosting on decision trees. Режим доступа: https://catboost.ai/ (дата обращения: 09.12.2021).

***Таблица 2.* Оценка алгоритмов автоматической классификации**

Table 2. Evaluation of automatic classification algorithms

Алгоритм	Точность *F1*-macro
TF-IDF vectors + Catboost	0,47
multilingual BERT	0,52
1. Deep Pavlov Bert	0,531
2. Rubert-tiny	0,527
3. Rubert-tiny tuned	0,545

Составлено авторами по материалам исследования / *Compiled by the authors based on the research materials*

В таблице 3 приведена матрица ошибок (матрица несоответствий). Значения выражены в процентах от количества экземпляров истинной категории; по диагонали матрицы – значения полноты по каждому классу. Наиболее часто модель ошибается при определении релевантных сообщений – сообщения, относящиеся к определенной категории, модель относит к «мусору». Рассмотрим детально каждую категорию. По 2 % сообщений категории «образования» модель определяет к «здравоохранению» и «безопасности» и по 1 % – «политика» и «ЖКХ и инфраструктура», 2 % и 1 % сообщений категории «здравоохранения» – прогнозируются к категориям «безопасность» и «образование», соответственно.

***Таблица 3.* Матрица ошибок, % от количества экземпляров истинной категории**

Table 3. Confusion matrix, percentage of the number of instances of the true category

Предсказанный класс / Истинный класс	Мусор	Образование	Здравоохранение	Безопасность	Соц. Обеспечение	Политика	Экология	Доступность товаров и услуг	ЖКХ и инфраструктура
Мусор	87	1	1	3	0	2	1	0	2
Образование	28	63	2	2	0	1	0	0	1
Здравоохранение	14	1	79	2	0	0	0	0	0
Безопасность	17	0	1	76	0	1	1	0	2
Соц. Обеспечение	24	2	3	1	54	10	0	0	2
Политика	33	1	3	4	3	46	1	0	4
Экология	26	0	0	6	0	0	55	0	8
Доступность товаров и услуг	41	0	3	0	3	0	0	40	6
ЖКХ и инфраструктура	22	1	0	2	0	1	2	0	68

Составлено авторами по материалам исследования / *Compiled by the authors based on the research materials*

В случае сообщений, относящихся к «безопасности», модель ошибается следующим образом: 2 % – «ЖКХ и инфраструктура» и по 1 % наблюдений отнесены к категориям «здравоохранение», «политика», «экология». В категории «социальное обеспечение» наблюдается большой процент ошибок – 10 % наблюдений, определяются моделью как политическая категория; 3 % – «здравоохранение», по 2 % – «образование» и «ЖКХ и инфраструктура» и 1 % – «безопасность». В случае категории «политика» модель совершает ошибки: по 4 % сообщений – «безопасность» и «ЖКХ и инфраструктура», 3 % – «здравоохранение» и «социальное обеспечение» и 1 % – «образование» и «экология».

Модель ошибается при определении категории «экология» два раза, относя 8 % сообщений «экологии» – к категории «ЖКХ и инфраструктура» и 6 % к «безопасности». Категория «доступность товаров и услуг» имеет наименьший объем сообщений в нашем наборе данных и алгоритм определяет 6 % сообщений – «ЖКХ и инфраструктура» и по 3 % сообщений к категориям «здравоохранение», «социальное обеспечение». В случае сообщений, относящихся к категории «ЖКХ и инфраструктура» – по 2 % сообщений прогнозируются к категориям «безопасность» и «экология» и по 1 % – «образование» и «политика». Таким образом, исходя из матрицы ошибок, мы наблюдаем основные ошибки модели, связанные с отнесением релевантных сообщений к «мусору», а также определение категории «социальное обеспечение» к «политике» и «экологии» к «ЖКХ и инфраструктура».

***Таблица 4.* Матрица значения точности, полноты и *f*1-метрики по каждой категории, %**

Table 4. Matrix of accuracy, completeness and *f*1 metrics for each category, %

Показатель	Мусор	Образование	Здравоохранение	Безопасность	Соц. Обеспечение	Политика	Экология	Доступность товаров и услуг	ЖКХ и инфраструктура
Точность	88	47	76	70	58	51	41	64	67
Полнота	88	62	78	75	53	46	58	38	65
*f*1	88	53	77	73	55	48	48	48	66

Составлено авторами по материалам исследования / *Compiled by the authors based on the research materials*

Значения метрик, представленные в таблице 4, позволяют нам оценить качество каждой категории с помощью агрегирующих метрик: точность (precision), полноту (recall) и f1-score. Наименьшее значение точности (процент правильно классифицированных примеров среди спрогнозированных) получены для категорий – «экология» (41 %), «образование» (47 %) и «политика» (51 %). С точки зрения полноты (процент правильно спрогнозированных примеров среди фактических) – «доступность товаров и услуг» (38 %), «политика» (46 %)

и «социальное обеспечение» (53 %). Наименьшие значение метрики *f1*-score, объединяющей в себе точность и полноту, были получены для категорий – «политика», «экология» и «доступность товаров и услуг» – 48 % для каждой.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ / DISCUSSION

Рассмотрим подробнее результаты обучения. Наиболее часто модель ошибается при определении релевантных сообщений («мусор»); алгоритм, как относит «мусорные» сообщениям к тематическим категориям, так и определяет тематические сообщения к «мусору». Проиллюстрируем тезис примерами сообщений, на которых ошибается нейронная сеть.

1. «Российские следователи ищут насильников и убийц школьницы».

Согласно исследовательской логике и разметке, данное сообщение является нерелевантным («мусор») – на его основе трудно сделать какие-либо информативные выводы. Нейронная сеть относит сообщение к категории «безопасность».

2. «Более 300 жителей Карелии сделали тест на ВИЧ. Завершилась акция «Тест на ВИЧ: Экспедиция 2020». В течение недели по пяти городам республики проехал мобильный пункт тестирования. Последней точкой маршрута стал Петрозаводск» (правильная категория – «здравоохранение»; предсказанная – «мусор»).

Рассмотрим в качестве примеров ошибки нейросети в сообщениях по тематическим категориям:

1) «Новосибирцы 25 апреля пожаловались на зловоние, которое распространяется по нескольким районам» (правильная категория – «экология»; предсказанная – «здравоохранение»);

2) «Совсем недоброе утро случилось у жителя Владивостока. Бетонная плита упала на его машину с крыши девятиэтажки. Мужчина спокойно чистил авто от снега, когда услышал странные звуки и поднял взгляд вверх. Увернуться от огромной плиты удалось в последний момент. Машине повезло гораздо меньше» (правильная категория – «безопасность»; предсказанная – «ЖКХ и инфраструктура»);

3) «В академическом лицее Петрозаводска зафиксирован случай COVID-19. Заболевший ученик не является младшеклассником. По рекомендации Роспотребнадзора, класс отправлен на дистанционное обучение. Карантин в школе не введен» (правильная категория – «здравоохранение»; предсказанная – «образование»).

ЗАКЛЮЧЕНИЕ / CONCLUSIONS

Таким образом, спроектированная и обученная модель Rubert-tiny позволяет с точностью *f1*-macro – 0,545 классифицировать посты социальной сети «ВКонтакте» согласно категориям оценки качества жизни населения. Результаты классификации в дальнейшем будут применены для расчета индекса онлайн-активности пользователей регионов и калькуляции индекса субъективного благополучия. Таким образом, повышение точности работы алгоритма позволит получить более достоверные результаты.

СПИСОК ЛИТЕРАТУРЫ

Богданов М.Б., Смирнов И.Б. (2021). Возможности и ограничения цифровых следов и методов машинного обучения в социологии // Мониторинг общественного мнения: экономические и социальные перемены. № 1. С. 304–328. https://doi.org/10.14515/monitoring.2021.1.1760

Двойникова А.А., Карпов А.А. (2020). Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных // Информационно-управляющие системы. № 4 (107). С. 20–30. https://doi.org/10.31799/1684-8853-2020-4-20-30

Крыштановская О.В. (2018). Бесконтактная социология: новые формы исследований в цифровую эпоху // Цифровая социология. № 1. С. 4-9. https://doi.org/10.26425/2658-347X-2018-1-4-8

Мюллер А., Гвидо С. (2016). Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными / пер. с анг. и ред. А.В. Груздева. М.: Вильямс. 393 с.

Николаенко Г.А., Федорова А.А. (2017). Нереактивная стратегия: применимость незаметных методов сбора социологической информации в условиях Web 2.0 на примере цифровой этнографии и BigData // Социология власти. Т. 29, № 4. С. 36–54. https://doi.org/10.22394/2074-0492-2017-4-36-54

Толстова Ю.Н. (2015). Социология и компьютерные технологии // Социологические исследования. № 8 (376). С. 3–13.

Чичканов В.П., Васильева Е.В. (2014). Управление качеством жизни в регионе: оценка эффективности и механизм // Государственное управление. Электронный вестник. № 47. С. 163–182.

Щекотин Е.В. (2021). Цифровые следы как новый источник данных о качестве жизни и благополучии: обзор современных тенденций // Вестник Томского государственного университета. № 467. С. 170–181. https://doi.org/10.17223/15617793/467/21

Щекотин Е.В., Мягков М.Г., Гойко В.Л., Кашпур В.В., Коварж Г.Ю. (2020). Субъективная оценка (не)благополучия населения регионов РФ на основе данных социальных сетей // Мониторинг общественного мнения: экономические и социальные перемены. № 1 (155). С. 78–116. https://doi.org/10.14515/monitoring.2020.1.05

Chen T., Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System // KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pp. 785–794. https://doi.org/10.1145/2939672.293978515

Dawson C. (2019). A–Z of digital research methods. New York: Routledge. 424 p.

Devlin J., Chang M., Lee K., Toutanova K. (2019). 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 (NAACL-HLT), V. 1. Pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

Jones K.S. (2004). A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. V. 60, No. 5. Pp. 493–502. https://doi.org/10.1108/00220410410560573

Joulin A., Grave E., Bojanowski P., Mikolov T. (2016). Bag of tricks for efficient text classification // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. V. 2. Valencia. Spain: Association for Computational Linguistics. Pp. 427–431. https://doi.org/10.18653/V1/E17-2068

Kutuzov A., Kuzmenko E. (2017). WebVectors: A toolkit for building web interfaces for vector semantic models // Communications in Computer and Information Science. V. 661. Pp. 155–161. https://doi.org/10.1007/978-3-319-52920-2_15

McGillivray M., Clarke M. [Eds]. (2006.) Understanding human well-being. Tokyo, New York, Paris: United Nations University Press. 380 p.

Mikolov T., Chen K., Corrado G., Dean J. (2013a). Efficient estimation of word representations in vector space // Proceedings of Workshop at ICLR. Scottsdale. May 2–4. Pp. 1–11.

Mikolov T., Yih W.-T., Zweig G. (2013b). Linguistic regularities in continuous space word representations // Proceedings of NAACL HLT. Atlanta. Georgia. June 9–14. Pp. 746–751.

Pennington J., Socher R., Manning C.D. (2014). GloVe: Global vectors for word representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics. Pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. (2018). Deep contextualized word representations // Proceedings of NAACL-HLT. V. 1. June 1–6. New Orleans. Louisiana: Association for Computational Linguistics. Pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202

Potdar K., Pardawala T.S., Pai C.D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers // International Journal of Computer Applications. V. 175, No. 4. Pp. 7–9. https://doi.org/10.5120/IJCA2017915495

Schober M.F., Pasek J., Guggenheim L., Lampe C., Conrad F.G. (2016). Research synthesis: Social media analyses for social measurement // Public Opinion Quarterly. V. 80, No. 1. Pp. 180–211. https://doi.org/10.1093/poq/nfv048

Soumya G.K., Joseph S. (2014). Text classification by augmenting bag of words (BOW) representation with co-occurrence feature // IOSR Journal of Computer Engineering. V. 16, No. 1. Pp. 34–38. https://doi.org/10.9790/0661-16153438

REFERENCES

Bogdanov M.B. and Smirnov I.B. (2021), "Opportunities and limitations of digital footprints and machine learning methods in Sociology", *Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsial'nye peremeny*, no. 1, pp. 304–328. (In Russian). https://doi.org/10.14515/monitoring.2021.1.1760

Chen T. and Guestrin C. (2016), "XGBoost: A Scalable Tree Boosting System", *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. https://doi.org/10.1145/2939672.293978515

Chichkanov V.P. and Vasilyeva E.V. (2014), "Management of regional life quality: effectiveness evaluation and mechanism", *Gosudarstvennoe upravlenie. Elektronnyi vestnik*, no. 47, pp. 163–182. (In Russian).

Dawson C. (2019), *A–Z of digital research methods*, Routledge, New York, USA.

Devlin J., Chang M., Lee K. and Toutanova K. (2019), "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 (NAACL-HLT)*, vol. 1, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dvoynikova A.A. and Karpov A.A. (2020), "Analytical review of approaches to Russian text sentiment recognition", *Information and control systems*, no. 4 (107), pp. 20–30. (In Russian). https://doi.org/10.31799/1684-8853-2020-4-20-30

Jones K.S. (2004), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, vol. 60, no. 5, pp. 493–502. https://doi.org/10.1108/00220410410560573

Joulin A., Grave E., Bojanowski P. and Mikolov T. (2016), "Bag of tricks for efficient text classification", *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2. Valencia, Spain: Association for Computational Linguistics, pp. 427–431. https://doi.org/10.18653/V1/E17-2068

Kryshtanovskaya O.V. (2018), "Contactless sociology: new forms of research in a digital age", *Digital Sociology*, no. 1, pp. 4-9. (In Russian). https://doi.org/10.26425/2658-347X-2018-1-4-8

Kutuzov A. and Kuzmenko E. (2017), "WebVectors: A toolkit for building web interfaces for vector semantic models", *Communications in Computer and Information Science*, vol. 661, pp. 155–161. https://doi.org/10.1007/978-3-319-52920-2_15

McGillivray M., Clarke M. [Eds], (2006.) *Understanding human well-being*, United Nations University Press, Tokyo, Japan; New York, USA; Paris, France.

Mikolov T., Chen K., Corrado G. and Dean J. (2013a), "Efficient estimation of word representations in vector space", *Proceedings of Workshop at ICLR*, Scottsdale, May 2–4, pp. 1–11.

Mikolov T., Yih W.-T. and Zweig G. (2013b), "Linguistic regularities in continuous space word representations", *Proceedings of NAACL HLT*, Atlanta, Georgia, June 9–14, pp. 746–751.

Müller A. and Guido S. (2016), *Introduction to machine learning with Python*, trans. from Eng. and ed. A.V. Gruzdeva, Williams, Moscow, Russia. (In Russian).

Nikolaenko G.A. and Fedorova A.A. (2017), "Non-reactive strategy: unobtrusive methods of gathering sociological information in web 2.0 age – evidence from digital ethnography and big data", *Sociology of power*, vol. 29, no. 4, pp. 36–54. (In Russian). https://doi.org/10.22394/2074-0492-2017-4-36-54

Pennington J., Socher R. and Manning C.D. (2014), "GloVe: Global vectors for word representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018), "Deep contextualized word representations", *Proceedings of NAACL-HLT*, vol. 1, June 1–6, New Orleans, Louisiana, Association for Computational Linguistics, pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202

Potdar K., Pardawala T.S. and Pai C.D. (2017), "A comparative study of categorical variable encoding techniques for neural network classifiers", *International Journal of Computer Applications*, vol. 175, no. 4, pp. 7–9. https://doi.org/10.5120/IJCA2017915495

Shchekotin E.V. (2021), "Digital footprints as a new source of data on quality of life and well-being: an overview of current trends", *Tomsk State University journal*, no. 467, pp. 170-181. (In Russian). https://doi.org/10.17223/15617793/467/21

Shchekotin E.V., Myagkov M.G., Goiko V.L., Kashpur V.V. and Kovarzh G.Yu. (2020), "Subjective measurement of population ill-being/well-being in the Russian regions based on social media data", *Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsial'nye peremeny*, no. 1 (155), pp. 78–116. (In Russian). https://doi.org/10.14515/monitoring.2020.1.05

Schober M.F., Pasek J., Guggenheim L., Lampe C. and Conrad F.G. (2016), "Research synthesis: Social media analyses for social measurement", *Public Opinion Quarterly*, vol. 80, no. 1, pp. 180–211. https://doi.org/10.1093/poq/nfv048

Soumya G.K. and Joseph S. (2014), "Text classification by augmenting bag of words (BOW) representation with co-occurrence feature", *IOSR Journal of Computer Engineering*, vol. 16, no. 1, pp. 34–38. https://doi.org/10.9790/0661-16153438

Tolstova Yu.N. (2015), "Sociology and computer technologies", *Sotsiologicheskie issledovaniya*, no. 8 (376), pp. 3–13. (In Russian).