Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2022»

Москва, 15–18 июня 2022 г.

Classification community publications of the «VKontakte» for assessing the quality of life of the population

Polina Basina
TSU
Tomsk, Russia
basina@data.tsu.ru

Vyacheslav Goiko
TSU
Tomsk, Russia
goiko@data.tsu.ru

Evgeny Petrov
TSU
Tomsk, Russia
petrov@data.tsu.ru

Vyacheslav Bakulin
TSU
Tomsk, Russia
slava38710505@gmail.com

Abstract

Social networks are an everyday tool for users to express their opinions and preferences. User digital trace are a valuable source of data for understanding the problems of the population in various spheres of life. The focus of this work is aimed at developing an algorithm for automatic classification of text content «VKontakte» according to the selected categories of quality of life. This social network is one of the most popular platforms among users. The categories of quality of life are «education», «healthcare», «security», «social security», «work of authorities», «ecology» and «accessibility of goods and services». The paper uses static and contextualized models for creating vector representations and effective algorithms for classifying Russian-language content of social networks (LSTM, BiLSTM, GRU, RuBERT). We prefer the RuBERT -tiny model due to the best completeness indicators in most categories.

Keywords: quality of life, digital trace, VKontakte, natural language processing, text classification, RuBERT
DOI: 10.28995/2075-7182-2022-21-1001-1016

Классификация публикаций сообществ «ВКонтакте» для оценки качества жизни населения

Полина Басина
НИ ТГУ
Томск, Россия
basina@data.tsu.ru

Вячеслав Гойко
НИ ТГУ
Томск, Россия
goiko@data.tsu.ru

Евгений Петров
НИ ТГУ
Томск, Россия
petrov@data.tsu.ru

Вячеслав Бакулин
НИ ТГУ
Томск, Россия
slava38710505@gmail.com

Аннотация

Сегодня социальные сети — это повседневный инструмент пользователя для выражения своих мнений и предпочтений. Цифровые следы, создаваемые в сети, являются ценным источником данных для выделения проблем населения в различных сферах жизнедеятельности. Фокус данной работы сосредоточен на разработке алгоритма, позволяющего автоматически классифицировать текстовый контент социальной сети «ВКонтакте», являющейся одной из популярных платформ среди пользователей, по категориям качества жизни: «образование», «здравоохранение», «безопасность», «социальное обеспечение», «работа органов власти», «экология» и «доступность товаров и услуг». Для реализации поставленной задачи в рамках работы

использованы статичные и контекстуализированные модели создания векторных представлений и эффективные алгоритмы классификации русскоязычного контента социальных сетей (LSTM, BiLSTM, GRU, RuBERT). На сегодняшний день мы отдаем предпочтение модели RuBERT-tiny за счет лучших показателей полноты в большинстве категорий.

Ключевые слова: качество жизни, цифровые следы, «ВКонтакте», обработка естественного языка, классификация текстов, RuBERT

1 Введение

В современном мире социальные сети являются повседневным инструментом пользователей для выражения своих мнений и предпочтений. Согласно данным отчета «We Are Social» и «Kepios» в РФ за 2021 год количество пользователей[1] социальных сетей увеличилось на 7 млн и на начало 2022 года составляет 106 млн. Ежедневно среднестатистический пользователь проводит в социальных сетях 2 часа 27 минут; в качестве популярных причин использования выделяют — «поддержание связей», «заполнение свободного времени», «чтение новостей», «поиск контента», «обмен мнениями». Самую многочисленную ежемесячную аудиторию собирают платформы «WhatsApp», «ВКонтакте» и «Instagram»[2]. Согласно последней официальной информации «ВКонтакте»[3], социальная сеть фиксирует резкий рост активности аудитории и количества пользователей — «к примеру, на неделе с 21 по 27 февраля средняя ежедневная аудитория платформы в России выросла на 200 000 пользователей»[4].

Цифровые следы, создаваемые в социальных сетях, являются ценным источником для различных приложений — анализ мнений и настроений, обобщение и категоризация текстов, обнаружение фейковых новостей и другие [Abbas 2021]. Одним из популярных направлений выступает оценка качеств жизни населения. То, как люди оценивают различные области своей жизни (субъективное благополучие), имеет важное значение для управленческого сектора и научных исследований. В качестве традиционного подхода оценки качества жизни выступают опросы, являющиеся дорогостоящей и трудоемкой процедурой, которая имеет определенные ограничения. Однако, сегодня пользователи склоны открыто делиться своими настроениями и мнениями в виде постов и реакций в социальных сетях, представляя тем самым ценную информацию для оценки их благополучия с применением алгоритмов машинного обучения [Hao et al. 2014]. Последние несколько лет на факультете психологии Санкт-Петербургского государственного университета проводится проект «Стресс, здоровье и психологическое благополучие в социальных сетях: кросс-культурное исследование». Исследователи выявляют лексические паттерны психологического благополучия, анализируя поведение пользователей социальных сетей [Bogolyubova et al. 2018, Bogolyubova et al. 2017].

При этом важно отметить, что эффективность работы алгоритмов по обработке естественного языка, в частности для задач классификации, зависит от многих факторов, где одними из значимых являются язык и источник данных. Например, новостные статьи и посты в социальных сетях будут написаны разными стилями речи. В качестве особенностей текстов социальных сетей исследователи отмечают: использование жаргонизмов, неологизмов и диалектов; неполные предложения; речевые и орфографические ошибки; символы эмодзи, как средства придания сообщениям эмоциональной окраски [Moshkin et al. 2019]. М. Абрахам и П. Набенде провели эксперименты по классификации твитов, написанных на различных языках, для эпидемиологического надзора с использованием нейросетевых архитектур CNN, RNN, LSTM и BERT. Исследователи отметили разную производительность алгоритмов в зависимости от того языка, на котором написаны тексты [Abraham et al. 2021]. Е. В. Михалкова и др. для решения задачи определения интересов пользователей сравнили применимость алгоритмов классификации на данных русскоязычных текстов «ВКонтакте» и англоязычных постов «Twitter». Они использовали несколько алгоритмов машинного обучения — метод опорных векторов, наивный Байесовский классификатор, логистическая регрессия, деревья решений и k-ближайших соседей. В ходе экспериментов ис-

[1] Важно отметить, что под пользователями не следует понимать уникальных людей.
[2] Digital 2022: THE RUSSIAN FEDERATION https://datareportal.com/reports/digital-2022-russian-federation
[3] Актуальная информация на момент написания статьи.
[4] ВКонтакте фиксирует резкий всплеск аудитории и просмотра контента https://vk.com/press/users-activity

следователи сделали вывод, что выбор социальной сети является важным фактором для разработки модели, а языковые различия не влияют на результаты классификации при должной нормализации данных [Mikhalkova et al. 2018]. С. Ватерлоо и др. изучили нормы выражения эмоций в социальных сетях — «Facebook», «Twitter», «Instagram» и «WhatsApp». Авторы обнаружили различия в платформах с точки зрения проявляемых там реакций [Waterloo et al. 2018].

Цифровые следы, создаваемые в социальной сети, с одной стороны, являются ценными источником данных для выделения проблем населения в различных сферах жизнедеятельности; с другой — представляют собой большие данные, изучение которых невозможно традиционными методами. Данные факторы обуславливают необходимость разработки автоматизированных решений. При этом учитывая разнообразие контента социальной сети, возникает необходимость его категоризации с применением экспертных мнений, что подразумевает под собой использование контролируемых методов машинного обучения. Результаты автоматической классификации в дальнейшем применяются для расчета индекса актуальности темы, который выражает то, насколько актуальна определенная тема (категория) в конкретном регионе в заданный временной промежуток. Индекс рассчитывается на основе цифровых следов анализируемого контента — лайки, комментарии, репосты. Фокус данной работы сосредоточен на разработке алгоритма, позволяющего автоматически классифицировать текстовый контент социальной сети «ВКонтакте», являющейся одной из популярных платформ среди пользователей, по категориям качества жизни. Статья состоит из 5 разделов: изучения практик применения алгоритмов машинного обучения для оценки благополучия пользователей социальных медиа, описания данных, описания экспериментов и методов оценки, результатов экспериментов и дальнейших путей развития.

2 Изучение практик применения алгоритмов машинного обучения для оценки благополучия пользователей социальных медиа

Е.В. Щекотин и др. условно выделяют три направления исследований, связанных с социальными медиа и благополучием: информационные технологии как инструмент изучения; социальные медиа как фактор влияния на благополучие; социальные сети как самодостаточный источник данных [Shchekotin, Myagkov et al. 2020]. Мы сосредоточимся на практиках оценки качества жизни на основе текстовых данных социальных сетей с применением алгоритмов машинного обучения.

В одной из работ предлагается единый подход к построению профиля субъективного благополучия на основе языка социальных сетей в обновлениях статуса «Facebook». Исследователи применяют анализ настроений для оценки аффективных характеристик пользователей («счастья») и обучают модель случайного леса для прогнозирования субъективного благополучия с использованием полученных оценок и других языковых функций обновлений статуса [Chen et al. 2017]. К. Джайдка и др. сравнили оценки благополучия на уровне округов США, основанные на данных «Twitter», с показателями индекса Гэллапа, рассчитанными на материалах телефонных опросов. Они обнаружили, что методы на уровне слов дали противоречивые измерения на уровне округа из-за региональных, культурных и социально-экономических различий в использовании языка. Однако, удаление всего лишь трех наиболее часто встречающихся слов привело к заметному улучшению результатов прогноза. Методы, основанные на данных, позволили получить надежные оценки, приближенные к индексу Гэллапа [Jaidka et al. 2020]. Другие авторы, используя данные социальных сетей 1785 пользователей с метками субъективного благополучия, обучают модели машинного обучения, которые способны «распознавать» индивидуальные оценки для пользователей [Hao et al. 2014]. М. Бхасин и др. анализируют аффективные и внутренние состояния пользователей. Они создали модель состояний счастья людей: G (длительное счастье), P (мерцание) и I (разочарование). Исследователи использовали XGBoost для классификации 54 066 пользователей «Twitter» на основе их твитов. Авторы утверждают, что, анализируя результаты классификации, могли бы повторно подтвердить характеристики, упомянутые в определении трех состояний (G, P, I), а также выявить дополнительные черты [Bhasin et al. 2021].

Многие исследователи акцентируют внимание на качестве жизни пользователей в период пандемии, когда социальные медиа позволяют получить уникальные данные. Ю. Хан и др. проанализировали с помощью алгоритмов классификации субъективное благополучие пользователей на основе сообщений в популярной в Китае социальной сети «Weibo» во время и после вспышки пандемии COVID-19. Результаты показывают тенденцию к снижению, а затем тенденцию к росту

уровня субъективного благополучия пользователей во время пандемии в целом [Han et al. 2022]. Ю. Ванг и др. изучили влияние изоляции на субъективное благополучие людей в Китае во время пандемии COVID-19 на материалах аналогичной социальной сети. Выборка состояла из двух групп: пользователи, проживающие в городах самоизоляции, и пользователи без ограничений на социальные контакты. Для каждой группы были рассчитаны показатели благополучия с помощью прогностических моделей машинного обучения в течение 2 недель до и после даты введения в действие блокировки жилых помещений, используя оригинальные сообщения пользователей в «Weibo» [Wang et al. 2020].

3 Описание данных

Рассматриваемый в рамках данной работы алгоритм обучен и применяется для контента социальной сети «ВКонтакте». Выбор сети обусловлен, с одной стороны, ее популярностью среди аудитории, что подтверждают статистические данные; с другой — возможностями самой платформы. Данные «ВКонтакте» обладают рядом преимуществ: публичный API; детализация контента во времени и по территориальным единицам; выражение собственного мнения пользователем (посты) и его открытое взаимодействие с контентом посредством различных реакций (лайки, комментарии, репосты); относительно низкие временные затраты. Среди недостатков отмечают смещение выборочной совокупности; технические трудности сбора данных; специфичность текстов социальной сети [Shchekotin, Kovarzh et al. 2020].

Для получения репрезентативных данных, позволяющих учесть территориальные особенности, которые могут проявляться в текстовом контенте как содержательно, так и с точки зрения языковых особенностей, в качестве источников были выбраны региональные сообщества. Важным критерием такого сообщества является территориальная принадлежность аудитории — не менее 50% подписчиков, указавших свое местоположение, должны быть из 1 региона, указанного пользователем как место проживания. Другие значимые характеристики, которые были использованы для отбора сообществ, представлены в работе [Shchekotin, Myagkov et al. 2020]. Полный список расположен в репозитории Github[5].

Каждый объект базы данных (далее — БД) «ВКонтакте» имеет числовой идентификатор, позволяющий с помощью API получить о нем информацию и связанные объекты. Например, при помощи идентификаторов сообществ могут быть выгружены их публикации, комментарии к ним (с указанием ID автора комментария), списки пользователей, которым понравилась публикации. Поскольку при создании нового объекта «ВКонтакте» ему присваивается идентификатор, являющийся результатом инкрементации идентификатора ранее созданного объекта, можно сгенерировать необходимый список идентификаторов для выгрузки без обращения к «ВКонтакте». Программное обеспечение для сбора данных реализовано на скриптовом языке Python, имеет ряд модулей, в частности, для работы с API «ВКонтакте», записи результатов в хранилище и обеспечения параллелизма при выгрузке. Для хранения выгрузок используется СУБД PostgreSQL.

Для обучения и оценки алгоритма классификации был сформирован набор размеченных данных — 84 000 постов «ВКонтакте». Были использованы случайные посты, опубликованные в региональных сообществах в период с января по июль 2021 года. Авторами публикаций могли выступать как участники сообществ, так и сами сообщества; при отборе постов не учитывались социолингвистические параметры авторов. В выборку могли попасть любые сообщения вне зависимости от количества их цифровых следов (лайки, комментарии, репосты, просмотры).

Учитывая специфику контента социальных сетей, необходимо было очистить данные от неинформативных сообщений, к которым относятся: развлекательный контент, спортивные события, рекламные и коммерческие сообщения, заметки фан-клубов и др. Отфильтрованные сообщения были размечены согласно выделенным категориям качества жизни: «образование», «здравоохранение», «безопасность», «социальное обеспечение», «работа органов власти», «экология» и «доступность товаров и услуг». Каждое сообщение могло быть отнесено только к 1 категории. На предыдущем этапе исследования было выделено 19 категорий [Shchekotin, Myagkov et al. 2020].

[5] Methodology of formation of the register of regional communities of the Vkontakte social network https://github.com/datacentr/Methodology-of-formation-of-the-register-of-regional-communities-of-the-Vkontakte-social-network

Показатели субъективного благополучия сформированы на основе анализа существующих подходов и моделей оценки. В данной статье выделенные раннее категории укрупнены, что обусловлено следующими факторами: пересечение категорий; некоторые из выделенных раннее категорий собирали мало сообщений; упрощение процедуры разметки.

Категория	**Описание категории**	**Пример сообщения**
Образование	К данной категории относятся посты на следующие темы: дошкольное, общее, профессиональное и послевузовское образование, курсы повышения квалификации, дополнительное образование детей и взрослых.	*«Хабаровские школы частично перейдут на дистанционное обучение. Ученики среднего и старшего звена после каникул в учебные заведения не вернутся. Заниматься они будут удалённо, из дома»*
Здравоохранение	К данной категории относятся сообщения, связанные с процедурами лечения, процессом оказания медицинских услуг, материальным оснащением медицинских учреждений.	*«Массовая вспышка коронавирусной инфекции зафиксирована в учреждении социальной защиты Тотемского района»*
Безопасность	К данной категории относятся сообщения, связанные с ситуациями нарушения, предотвращения и обеспечения безопасности жителей.	*«Труп мужчины нашли в Тюмени в Антипино 20 апреля. Тело было обнаружено в районе ул. Изумрудная. Сейчас следователям предстоит выяснить обстоятельства смерти человека»*
Социальное обеспечение	К данной категории относятся сообщения, связанные с оказанием помощи и поддержки социально-незащищенным слоям населения государством.	*«Семьи из Карелии получили выплаты на строительство жилья»*
Политика	К данной категории относятся сообщения о свободе СМИ, протестном потенциале, свободе выборов, отношению к власти, политические решения, внутренняя политика.	*«Власти Москвы отказали местному отделению КПРФ в праве провести митинг 23 февраля, сославшись на ограничения из-за коронавируса.»*
Экология	Эта категория представляет информацию о взаимодействии субъектов с окружающей средой (природные ресурсы, животный мир). Мы акцентируем внимание как на процессы разрушающего и неконтролируемого влияния человека (несанкционированные свалки, выбросы заводов), так и на осознанные практики проявления экологического сознания/культуры (раздельный сбор мусора, субботники, переработка мусора, зоозащитники).	*«На берегу Верхнего пруда и горожанами, и журналистами были обнаружены алые следы, похожие на кровь. А местные жители рассказали, что до этого здесь были замечены браконьеры. Общественники тогда направили письменные жалобы в различные инстанции»*

Категория	**Описание категории**	**Пример сообщения**
Доступность товаров и услуг	К данной категории относятся сообщения, связанные с вопросами формирования и изменения ценовой политики, физической и ценовой доступностью товаров и услуг в соответствии с доходами населения.	*«В Ульяновской области внедрят методики сдерживания цен в области здравоохранения, дополнительного образования и продовольствия. Ситуацию 26 марта обсудили на совещании по финансово-экономическим вопросам.*
ЖКХ и инфраструктура	Данная категория представляет собой работу органов власти в сфере оказания жилищно-коммунальных услуг и организации доступной среды (инфраструктуры).	*«И, снова всем доброго субботнего утра от жителей многострадального микрорайона Осипенко. Сегодня в очередной раз потёк котёл на котельной МУП ДКР. Батареи стали остывать, а горячая вода по температуре упала до 40 градусов.»*

Таблица 1: Описание и примеры сообщений для каждой категории[6]

На первом этапе разметки сообщений происходила оценка понимания разметчиками инструментария исследования; они делали пробную разметку датасета, который состоял из индивидуальных сообщений (в дальнейшем проходили экспертную проверку) и ханипотов (общие для всех разметчиков сообщения, которые заранее размечены экспертами и позволяют контролировать качество разметки). Данный шаг позволил нам выявить общие и индивидуальные проблемные места для их дальнейшей проработки. После этого каждый разметчик выполнил контрольную разметку, которая прошла экспертную проверку. На основе описанных выше этапов происходит квалификация высокоэффективных разметчиков (разметчиков, допустивших минимальное количество ошибок), удаление низкоэффективных и постоянная обратная связь с участника для улучшения ими понимания задачи. Для дальнейшей разметки были использованы следующие подходы, позволяющие минимизировать случайные факторы (усталость, ограниченность времени) и смещение выборки:

1. Итеративность («порционная выдача датасетов» на некоторый временный период, определенный эмпирическим путем);
2. Рандомное представление сообщений (по регионам и сообществам) в датасетах каждого разметчика;
3. Ханипоты (проверочные сообщения).

[6] Примеры сообщений, представленные в таблицы, были взяты из региональных сообществ

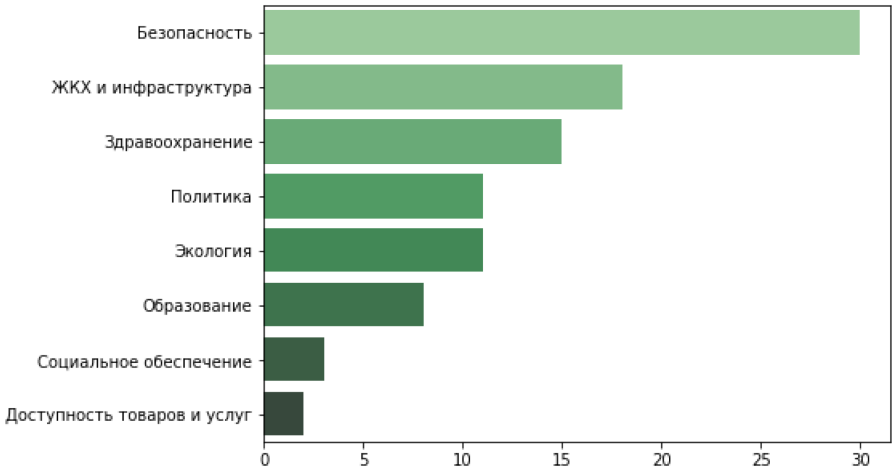

Рисунок 1: Распределение сообщений по категориям в %

В получившемся наборе присутствует сильный дисбаланс классов (рисунок 1). 64% сообщений от общего объема являются нерелевантным («мусором»); наиболее крупные категории — «безопасность», «ЖКХ и инфраструктура» и «здравоохранение»; наименее — «доступность товаров и услуг», «социальное обеспечение». Нерелевантные сообщения были сохранены в выборке, так при применении алгоритма классификации в реальных условиях всегда сохраняется необходимость фильтрации сообщений и только потом дальнейшая их категоризация.

Для эффективного анализа характеристик сообщений, с точки зрения их деления на репрезентативные и нет, была проведена предварительная обработка (удаление ссылок, хэштегов, упоминаний, используемых в социальной сети (через @), нерелевантных символов («\n», лишние пробелы), стоп-слов (встроенные стоп-слова на русском и английском языке библиотеки nltk); перевод в нижний регистр; сохранение токенов, состоящих только из буквенных символов русского и английского алфавитов; лемматизация). Основной язык постов — русский; однако, встречаются специфичные термины на английском языке. Представленные на рисунках 2-4 характеристики позволяют понять сообщения с точки зрения их информативных характеристик. Один символ может быть представлен буквой из русского или английского алфавитов; слова — леммы (приведенные словоформы к нормальной (словарной) форме).

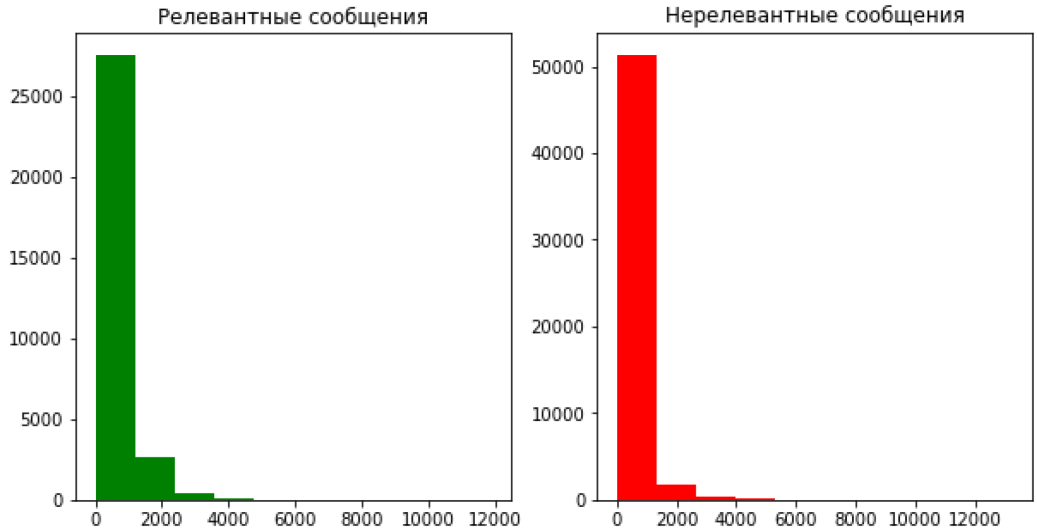

Рисунок 2: Количество символов в очищенных текстах

В 75% случаев количество символов в релевантных сообщениях не превышает 672, в случае нерелевантных — 358 (рисунок 2).

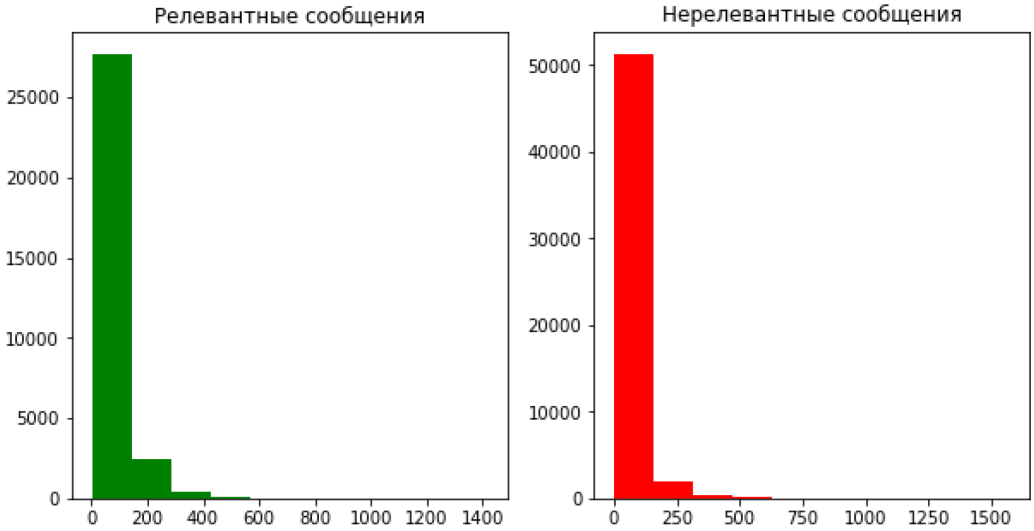

Рисунок 3: Количество слов в очищенных текстах

Количество слов в каждом релевантном очищенном тексте не превышает в 75% случаев 79 слов; в случае нерелевантных — 45 слов (рисунок 3).

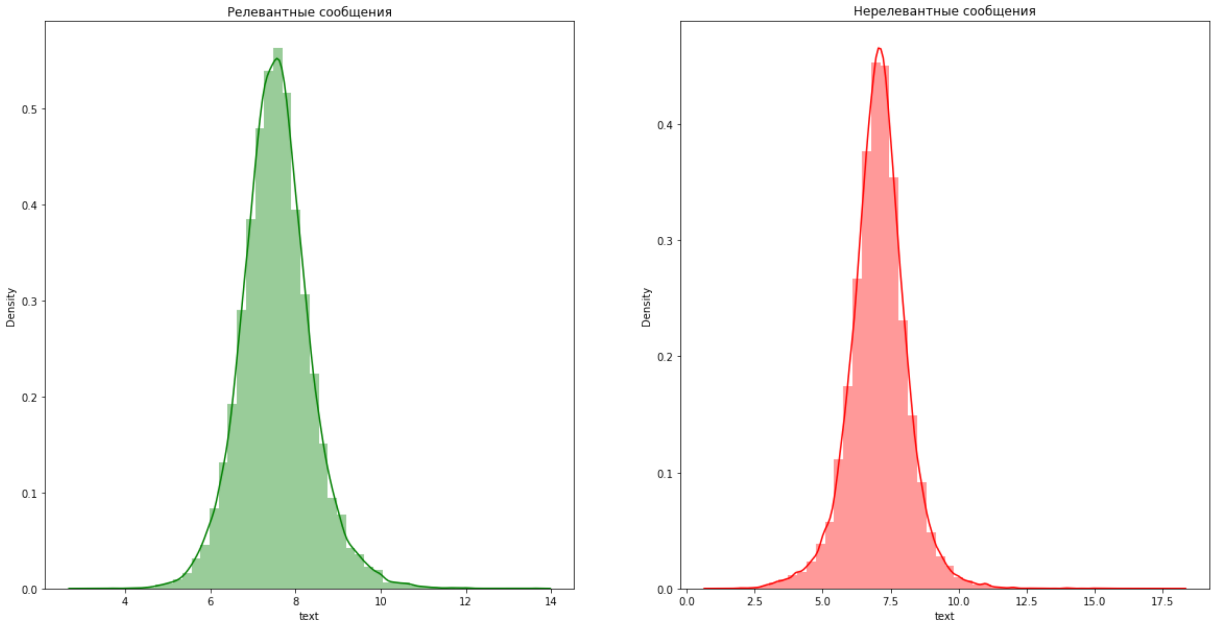

Рисунок 4: Средняя длина слов в очищенных текстах

Средняя длина слова во всех текстах не превышает в 75% случаев 8 символов (рисунок 4).

4 Описание экспериментов и методов оценки

Процедура классификации текстов состоит из нескольких этапов — предварительная обработка (очистка от «шума»), векторизация (извлечение признаков), построение классификатора и оценка полученных результатов [Dvoinikova et al., 2020]. Данные были разделены в соотношение: 80% обучающие и 20% тестовые. Некоторые шаги в предварительной обработке зависели от способа векторизации текстов. Например, в случае предварительно обученных моделей данные необходимо

было подготовить в соответствии с выбранной моделью. А. Голубев и другие исследователи обращают внимание, что в случае моделей BERT предварительная обработка не оказывается значимого влияния (изменение около 0,01%) [Golubev et al. 2020]. В таблице 2 представлены способы предварительной обработки и объем корпуса в словоупотреблениях (токенах).

Используемая модель	**Способ предварительной обработки**	**Объем корпуса (количество токенов)**
Исходные данные	Отсутствует	5 109 864
RuBERT	Токенизация, удаление знаков пунктуации, приведение в нижний регистр, сохранение буквенных символов, фильтрация по количеству символов.	3 227 809
Статичные модели создания векторных представлений, обученные на данных	Токенизация; перевод в нижний регистр; удаление ссылок, хэштегов, упоминаний, используемых в социальной сети (через @), нерелевантных символов («\n», лишние пробелы), стоп-слов (встроенные стоп-слова на русском и английском языке библиотеки nltk); сохранение токенов, состоящих только из буквенных символов русского и английского алфавитов; лемматизация.	3 585 784
Предварительно обученные статичные модели векторных представлений	Аналогичные процедура, которая описана выше, а также POS-тегирование	3 541 104

Таблица 2: Способы предварительной обработки и объем корпуса (количество токенов)

Следующий этап заключался в отображении текстов в виде векторов. В рамках работы использованы статичные и контекстуализированные модели создания векторных представлений. Статичные модели построены на принципах дистрибутивной семантики, основная идея которых заключается в том, что значение определяется употреблением, а семантика может быть получена из контекстов, в которых употребляется данное слово. Кутузов и др. отмечают преимущество алгоритмов, построенных на принципах дистрибутивной семантики, так как они основаны на данных и не требуют трудоемкого и субъективного процесса построения онтологии. Также в своей работы авторы отметили, что для русского языка модель Word2vec с архитектурой skip-gram показала результаты хуже, чем CBOW; в предыдущих исследованиях для английского языка была отмечена обратная тенденция [Kutuzov et al. 2015]. Задача архитектуры CBOW — предсказать текущее слово, исходя из окружающего его контекста; skip-gram использует текущее слово для определения окружающих его слов [Mikolov et al. 2013]. FastText является развитием модели Word2vec; однако, теперь формируются не только векторы слов, но и векторы n-грам, что позволяет эффективно работать со словами отсутствующими в словаре или содержащими ошибки и опечатки [Joulin et al. 2017].

В качестве контекстуализированных представлений использованы векторы русскоязычной версии модели BERT. В случае моделей Word2vec и FastText для многозначных слов будет получен один эмбеддинг слова. BERT учитывает окружающий контекст предложения и генерирует различные эмбеддинги для многозначных слов, а также векторизует текст, учитывая близость слов [Devlin et al. 2019].

Мы проанализировали популярные и эффективные алгоритмы классификации русскоязычного контента социальных сетей — рекуррентные нейронные сети (LSTM, BiLSTM, GRU). Однако, на сегодняшний день предобученная на большом корпусе русскоязычных данных модель RuBERT демонстрирует лучшие показатели. В таблице 3 представлены обобщенные результаты [Oliseenko et al. 2021, Narynov et al. 2020, Shulginov et al. 2021, Smetanin 2020, Golubev et al. 2020, Konstantinov et al. 2021, Kuratov et al. 2019, Arkhipenko et al. 2016].

Задача	**Источник данных**	**Лучший алгоритм**
Психологические особенности пользователей	«Вконтакте»	LSTM, BiLSTM
Определение токсичных, агрессивных и оскорбительных комментариев	«Вконтакте», «Пикабу», «2ch»	RuBERT
Сентимент-анализ	«ВКонтакте», «Twitter»	RuBERT, GRU, LSTM

Таблица 3: Популярные алгоритмы классификации русскоязычного контента социальных сетей

Один из информативных способов оценки — матрица ошибок. Учитывая присутствующий в данных дисбаланс и одинаковую важность каждого класса, для оценки и сравнения работы моделей была выбрана метрика f1-macro, которая позволяет обобщить метрики точности и полноты [Muller et al. 2016]. Для построения моделей использован скриптовый язык программирования Python.

5 Результаты экспериментов

В качестве статичных векторных представлений были как обучены модели Word2vec и FastText (размер вектора — 300, размер окна — 2, архитектура CBOW), так и использованы предобученные на основе корпуса «Araneum Russicum Maximum»[7]. В таблице 4 представлены результаты экспериментов.

Алгоритм	**f1-macro**	**Алгоритм**	**f1-macro**
Word2Vec + LSTM	0,41	Word2Vec Araneum + LSTM	0,52
Word2Vec + BiLSTM	0,44	Word2Vec Araneum + BiLSTM	0,49
Word2Vec + GRU	0,48	Word2Vec Araneum + GRU	0,50
Word2Vec + BiGRU	0,47	Word2Vec Araneum + BiGRU	0,52
Word2Vec + BiLSTMBiGRU	0,51	Word2Vec Araneum + BiLSTMBiGRU	0,52
FastText + LSTM	0,45	FastText Araneum + LSTM	0,51
FastText + BiLSTM	0,46	FastText Araneum + BiLSTM	0,54
FastText + GRU	0,49	FastText Araneum + GRU	0,50
FastText + BiGRU	0,49	FastText Araneum + BiGRU	0,536
FastText + BiLSTMBiGRU	0,53	FastText Araneum + BiLSTMBiGRU	0,552
multilingual BERT	0,52	RuBERT DeepPavlov	0,531
		RuBERT-tiny	0,527
		RuBERT-tiny tuned	0,56

Таблица 4: Оценка алгоритмов автоматической классификации

[7]RusVectores: модели https://rusvectores.org/ru/models/

Рассмотрим случаи, где в качестве векторных представлений использованы статичные встраивания, обученные на исследовательской выборке. Лучшие результаты демонстрирует модель FastText с использованием нейросетевой архитектуры GRU (однонаправленная и двунаправленная) и комбинированных решений (последовательное использование слоев BiGRU и BiLSTM в рамках одной архитектуры). Предобученные векторные вложения показывают лучшие результаты по сравнению с моделями, обученными только на имеющихся данных — FastText Araneum и модель BiLSTMBiGRU, где точность f1-macro равна 0.552. Среди всех экспериментов лучшие результаты достигнуты моделью RuBERT-tiny tuned, где f1-macro равно 0.56. Для понимания экспериментов сравним матрицы несоответствий моделей, показавших наибольшие результаты. На рисунках 5-6 приведена матрица ошибок (матрица несоответствий). Значения выражены в процентах от количества экземпляров истинной категории; по диагонали матрицы — значения полноты по каждому классу.

Рисунок 5: Матрица несоответствий RuBERT

Рисунок 6: Матрица несоответствий FastText Araneum + BiLSTMBiGRU

Сравнивая между собой показатели матриц несоответствий, мы наблюдаем следующую картину. Модель RuBERT-tiny с точки зрения полноты лучше определяет категории — «образование» (58% против 48%), «здравоохранение» (79% против 73%), «политика» (41% против 37%), «доступность товаров и услуг» (39% против 36%) и «ЖКХ и инфраструктура» (70% против 64%). Модель, основанная на предобученных встраиваниях FastText, лучше определяет категории «безопасность» (79% против 75%), «социальное обеспечение» (65% против 54%), «экология» (60% против 58%).

Наиболее часто модели ошибаются при определении релевантных сообщений — сообщения, относящиеся к определенной категории, модели относят к «мусору». В случае других общих ошибок моделей следует отметить: сообщения категории «социальное обеспечение» модели относят к «политике» (8% ошибок алгоритма RuBERT и 5% BiLSTMBiGRU), «экология» к «безопасности» (7% ошибок алгоритма RuBERT и 9% BiLSTMBiGRU), «доступность товаров и услуг» к «ЖКХ и инфраструктуре» (8% ошибок алгоритма RuBERT и 6% BiLSTMBiGRU).

Значение метрик, представленные на рисунках 6–7, позволяют нам оценить качество каждой категории с помощью агрегирующих метрик: точность (precision), полнота (recall) и f1-score.

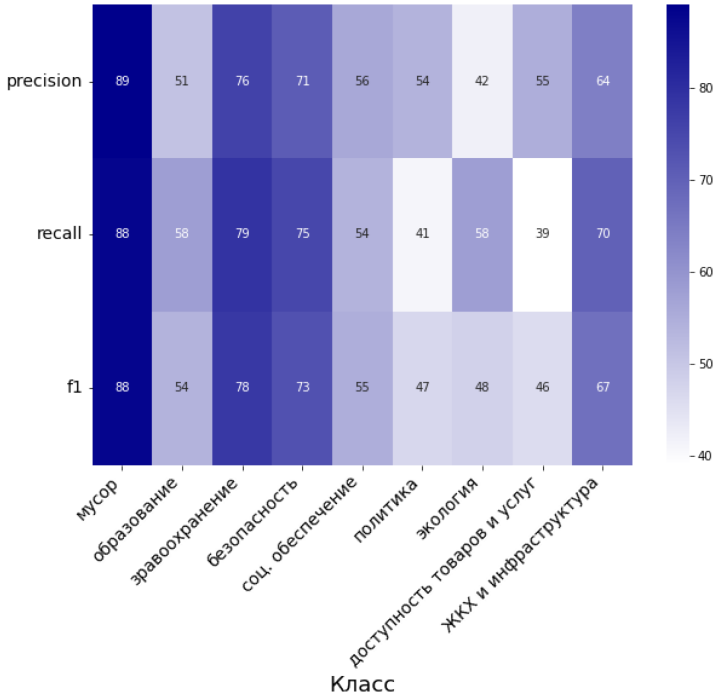

Рисунок 7: Матрица значений точности, полноты и f1-метрики по каждой категории — RuBERT

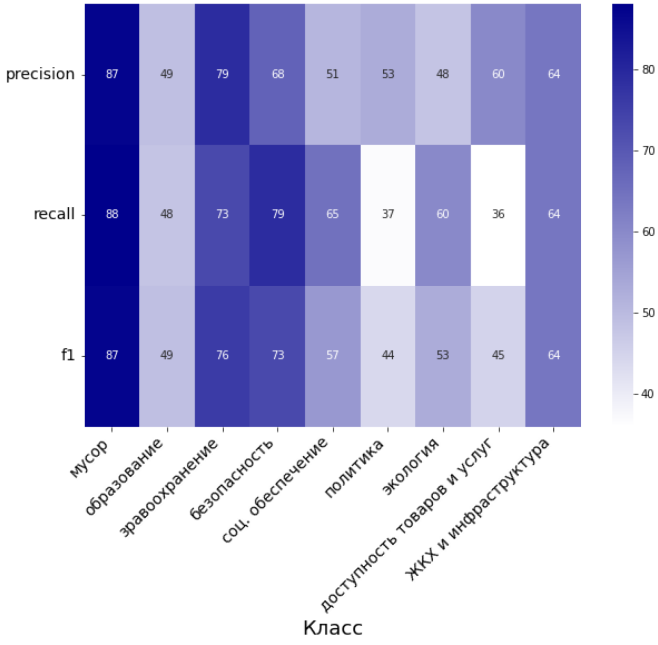

Рисунок 8: Матрица значений точности, полноты и f1-метрики по каждой категории — FastText Araneum + BiLSTMBiGRU TMbiGRU

В таблице 5 представлены примеры сообщения, в которых совершают ошибки рассматриваемые в рамках работы модели.

Сообщение[8]	**Правильная категория**	**RuBERT**	**BiLSTMBiGRU**
«Когда очень сильно не везёт! В рязанском селе погиб пешеход. Его не заметила пьяная женщина-водитель»	Безопасность	Нерелевантное сообщение	Безопасность
«Накануне в Кировском районе жители обнаружили среди мусора тела двух младенцев»	Безопасность	Экология	ЖКХ и инфраструктура
«*За 8 месяцев с начала года в 21 пилотном регионе, получившем федеральное финансирование, заключено более 60 тыс. социальных контрактов. Это 88,2% от запланированного на 2020 год объема социальных контрактов. В 2020 году востребованным стало заключение соцконтрактов на преодоление сложной жизненной ситуации — 33,4 тыс. семей (54,1%) и на помощь в поиске работы — 22,6 тыс. семей (36,5%)»*	Социальное обеспечение	Нерелевантное сообщение	ЖКХ и инфраструктура
«Уголовное дело по статье «мошенничество, совершенное группой лиц по предварительному сговору, в крупном размере» суд рассмотрит в отношении 41-летнего председателя участковой избирательной комиссии Екатеринбурга и его 32-летнего сообщника, сообщает пресс-служба прокуратуры РБ».	Безопасность	Безопасность	Политика

Таблица 5: Примеры сообщений, в которых ошибаются рассматриваемые модели

Таким образом, детальный анализ матриц несоответствий позволил нам понять преимущества и недостатки каждой модели. На сегодняшний день мы отдаем предпочтение модели Ruber-tiny за счет лучших показателей полноты в большинстве категорий. Модель FastText, как мы предполагали, показывает лучшие результаты по сравнению с векторными представлениями Word2vec. Мы считаем, что это объясняется тем, что FastText может векторизовать слова, не встречающиеся в обучающей выборке.

6. Дальнейшие пути развития

В дальнейшем мы видим два направления развития. Во-первых, необходимо продолжить эксперименты со способами предварительной обработки и векторизации текстов, в частности, использовать как модели предобученные на других корпусах, которые будут релевантными обучающей выборке, так и модели Glove и ELMo. Во-вторых, разделить процедуру категоризации контента на два этапа: сначала разработать алгоритм бинарной классификации, который позволит проводить фильтрацию сообщений; далее — категоризировать уже отфильтрованные сообщения. В текущей версии модели большинство ошибок связаны именно с категорией «нерепрезентативные сообщения».

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

[8] Примеры сообщений, представленные в таблицы, были взяты из региональных сообществ

References

[1] Abbas Ash Mohammad Social network analysis using deep learning: applications and schemes // Social Network Analysis and Mining. — 2021. — Vol. 11

[2] Abraham Mark, Nabende Peter Evaluation of different machine learning approaches and input text representations for multilingual classification of tweets for disease surveillance in the social web // Journal of Big Data. — 2021. — Vol. 8.

[3] Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D. Comparison of Neural Network Architectures for Sentiment Analysis of Russian Tweets //Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2016». — 2016.

[4] Bogolyubova Olga, Panicheva Polina, Tikhonov Roman, Ivanov Viktor, Ledovaya Yanina Dark Personalities on Facebook: Harmful Online Behaviors and Language // Computers in Human Behavior. — 2018. — Vol. 78 — P.151–159.

[5] Bogolyubova Olga, Tikhonov Roman, Ivanov Viktor, Panicheva Polina, Ledovaya Yanina Violence Exposure, Traumatic Stress and Well-being in a Sample of Russian Adults: a Facebook-based Study // Journal of Interpersonal Violence, First Published — 2017. — Vol. 5-6. — P.1476-1491.

[6] Bhasin Mayank, Harshit, Goyal Pawan Which acts model happiness?: an exploratory analysis on Twitter and Goodreads // ASONAM '21: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. — 2021. — P. 577-584.

[7] Chen Lushi, Kosinski Michal, Stillwell David, Davidson Robert Building a profile of subjective well-being for social media users. — PLOS ONE. — 2017 — Vol. 12

[8] Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina 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). — 2019. — Vol. 1 — P. 4171–4186.

[9] Dvoinikova Anastasia, Karpov Alexey Analytical review of approaches to the recognition of the tonality of Russian-language text data // Information and control systems. — 2020. — № 4 (107). — P. 20–30.

[10] Golubev Anton, Loukachevitch Natalia Improving Results on Russian Sentiment Datasets — 2020. — Vol. arXiv:2007/14310v1.

[11] Han Yingying, Pan Wenhao, Li Jinjin, Zhang Ting, Zhang Qiang, Zhang Emily Developmental Trend of Subjective Well-Being of Weibo Users During COVID-19: Online Text Analysis Based on Machine Learning Method // Frontiers in Psychology. — 2022. — Vol.12.

[12] Hao Bibo, Li Lin, Gao Rui, Li Ang, Zhu Tingshao Sensing Subjective Well-Being from Social Media // Lecture Notes in Computer Scienceю. — 2014. — P. 324-325.

[13] Jaidka Kokil, Giorgi Salvatore, Schwartz H., Kern Margaret, Ungar Lyle, Eichstaedt, Johannes Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods // Proceedings of the National Academy of Sciences. - 2020. - Vol 117. - P. 10165–10171

[14] Joulin Armand, Grave Edouard, Bojanowski Piotr, Mikolov Tomas Bag of Tricks for Efficient Text Classification. — 2017. — Vol. arXiv:1607.01759.

[15] Konstantinov Andrey, Moshkin Vadim, Yarushkina Nadejda Approach to the use of language models BERT and Word2vec in sentiment analysis of social network texts // Springer Nature Switzerland. — 2021. — Vol. 337. — P. 462-437.

[16] Kuratov Yuri, Arkhipov Mikhail Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language — 2019. — Vol. arXiv:1905.07213v1

[17] Kutuzov Andrei, Andreev Igor Texts in, meaning out: neural language models in semantic similarity task for Russian. — 2015. — Vol. arXiv:1504.08183.

[18] Mikhalkova Elena, Karyakin, Yuri, Ganzherli, Nadezhda, Grigoriev Dmitry Machine Learning Classification of User Interests Across Languages and Social Networks // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2018». — 2018.

[19] Mikolov Tomas, Chen Kai, Corrado, G., Dean Jeffrey Efficient Estimation of Word Representations in Vector Space — 2013a. — Vol. arXiv:1301.3781.

[20] Mikolov Tomas, Yih Wen-tau, Zweig G. Linguistic Regularities in Continuous Space Word Representations // In Proceedings of NAACL HLT. — 2013b. — P. 746-751.

[21] Moshkin Vadim, Yarushkina Nadejda, Andreev Ilya The Sentiment Analysis of Unstructured Social Network Data Using the Extended Ontology SentiWordNet // Proceedings — international conference on developments in esystems engineering, dese 12th International Conference on the Developments in eSystems Engineering, DeSE 2019. — 2019. — P. 576-580.

[22] Muller Andreas, Guido Sara Introduction to Machine Learning using Python. A guide for data specialists. — 2016. — Moscow: Williams. 393 p.

[23] Narynov Sergazy, Mukhtarkhanuly Daniyar, Omarov Batyrkhan Dataset of depressive posts in Russian language collected from social media // Data in Brief. — 2020. — Vol. 29

[24] Oliseenko Valerii, Tulupyeva Tatiana Neural Network Approach in the Task of Multi-label Classification of User Posts in Online Social Networks //2021 XXIV International Conference on Soft Computing and Measurements (SCM). — 2021. — P. 46-48
[25] Shchekotin E.V., Kovarzh G.Yu., Goiko V.L., Petrov E.Yu., Bakulin V.V. Assessment of the quality of life of the population of the regions of the Russian Federation based on digital data: methodological aspects // Vectors of well-being: economy and society. — 2020. — № 3 (38). — P. 138-156.
[26] Shchekotin E. V., Myagkov M. G., Goiko V. L., Kashpur V. V., Kovarzh G. Yu. Subjective assessment of the (non) well-being of the population of the regions of the Russian Federation based on social network data // Monitoring of public opinion: Economic and social changes. — 2020. — №1 (155). — P. 78–116.
[27] Shulginov V., Mustafin, R., Tillabaeva A.A. Automatic Detection of Implicit Aggression in Russian Social Media Comments // Conference: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021"– 2021.
[28] Smetanin, Sergey Toxic Comments Detection in Russian // Conference: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2020» — 2020.
[29] Wang Yilin, Wu Peijing, Liu Xiaoqian, Li Sijia, Zhu Tingshao, Zhao Nan Subjective Well-Being of Chinese Sina Weibo Users in Residential Lockdown During the COVID-19 Pandemic: Machine Learning Analysis. // Journal of Medical Internet Research. — 2020 — Vol.22.
[30] Waterloo Sophie, Baumgartner Susanne, Peter Jochen, Valkenburg Patti Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram and WhatsApp // New Media & Society. — 2017. — Vol. 20 — P. 1813–1831.