АНАЛИЗ МНЕНИЙ И НАСТРОЕНИЙ В РЕГИОНАХ РФ

ОПИСАНИЕ КЕЙСА

Разработано Центром прикладного анализа больших данных

Национальный исследовательский Томский государственный университет

Megaputer Your Knowledge Partner™

Университетский консорциум исследователей больших данных

Центром прикладного анализа больших данных Национального исследовательского Томского государственного университета реализуется исследовательский проект, направленный на мониторинг мнений и настроений пользователей на основе данных социальной сети ВКонтакте. Данный проект подразумевает применение метода социологии 1 к 1 с охватом всех активных в социальных медиа жителей региона. Основной целью исследования является мониторинг региональных сообществ для выявления реального отношения жителей к теме, явлению или муниципальному объекту. После выявления отношения населения к различным сферам общественной жизни проводится идентификация основных центров производства и распространения информации (лидеров общественного мнения), анализ их активности, портретных характеристик и генерируемого ими контента.

Стоит отметить, что следует различать лидеров общественного мнения и сетевых активистов. Лидер общественного мнения: высказывает собственное мнение о тех или иных социальных, экономических или политических событиях; генерирует собственный контент на социальную, политическую или экономическую тематику, который составляют большую часть его публикаций; получает отклик на собственные публикации (лайки, репосты, комментарии) от других пользователей. В то время как сетевой активист: распространяет контент на социальную, политическую или экономическую тематику при помощи репостов; может генерировать собственные контент на вышеуказанные темы, но основной процент публикаций является репостами; может как получать, так и не получать отклика (лайки, репосты, комментарии) на собственные публикации.

Мониторинг региональных сообществ проводится на основе подхода к измерению субъективного качества жизни населения, разработанного Центром прикладного анализа больших данных Национального исследовательского Томского государственного университета, который подразумевает измерение субъективного благополучия населения регионов, зарегистрированного в социальных сетях, посредством анализа текстового контента региональных сообществ и цифрового следа пользователей,

проявленного в виде реакций (лайки, репосты, комментарии) на размещенные публикации. На данный момент подход постоянно совершенствуется и сейчас проводится совершенствование алгоритма классификации текстового контента и формулы расчета индекса субъективного благополучия. Разработанная методика подразумевает реализацию следующих этапов:

1

Отбор региональных сообществ по каждому субъекту Российской Федерации. Если ранее данные группы отбирались вручную, то на сегодняшний день это происходит автоматически.

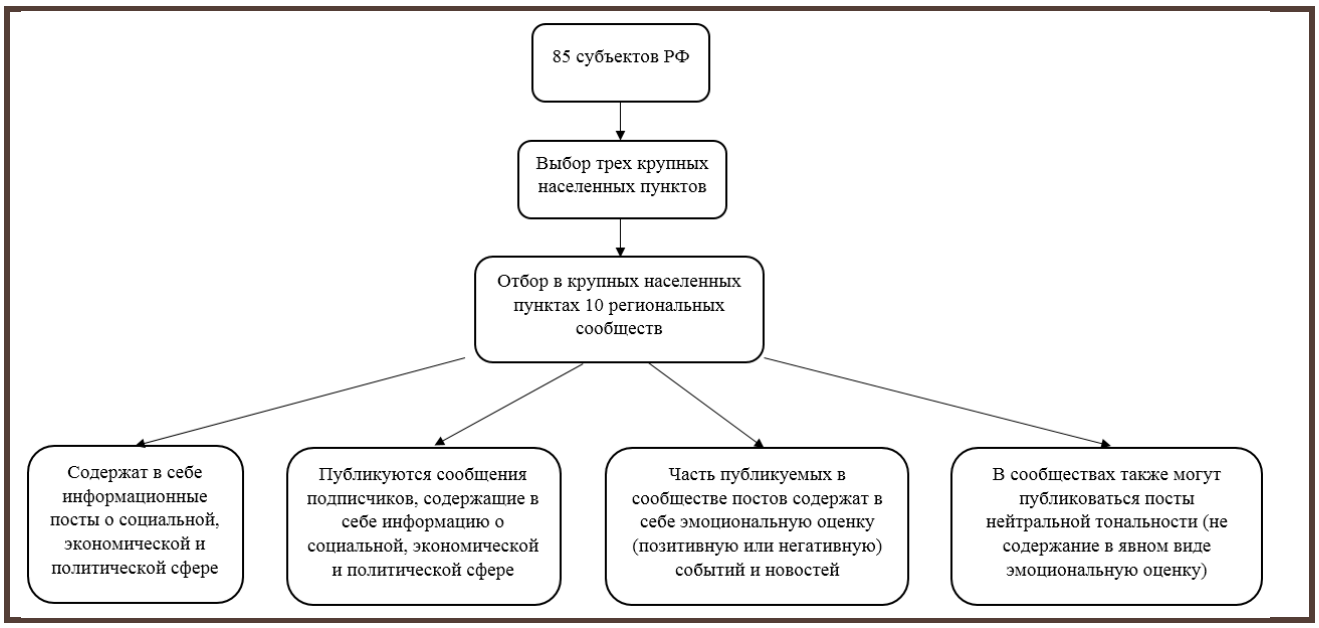

Рисунок 1 – Процесс ручного отбора региональных сообществ

Рисунок 2 – Процесс автоматического отбора региональных сообществ

2

Из отобранных региональных сообществ за необходимый временной период (например, 2019 год) выгружается текстовый контент с указанием даты размещения и количества лайков, репостов, комментариев и просмотров. Выгрузка данных проводилась с помощью платформы по сбору и анализу данных социальных медиа Университетского консорциума исследователей больших данных (www.opendata.university), разработанной командой Центра прикладного анализа больших данных на основе API (интерфейс прикладного программирования).

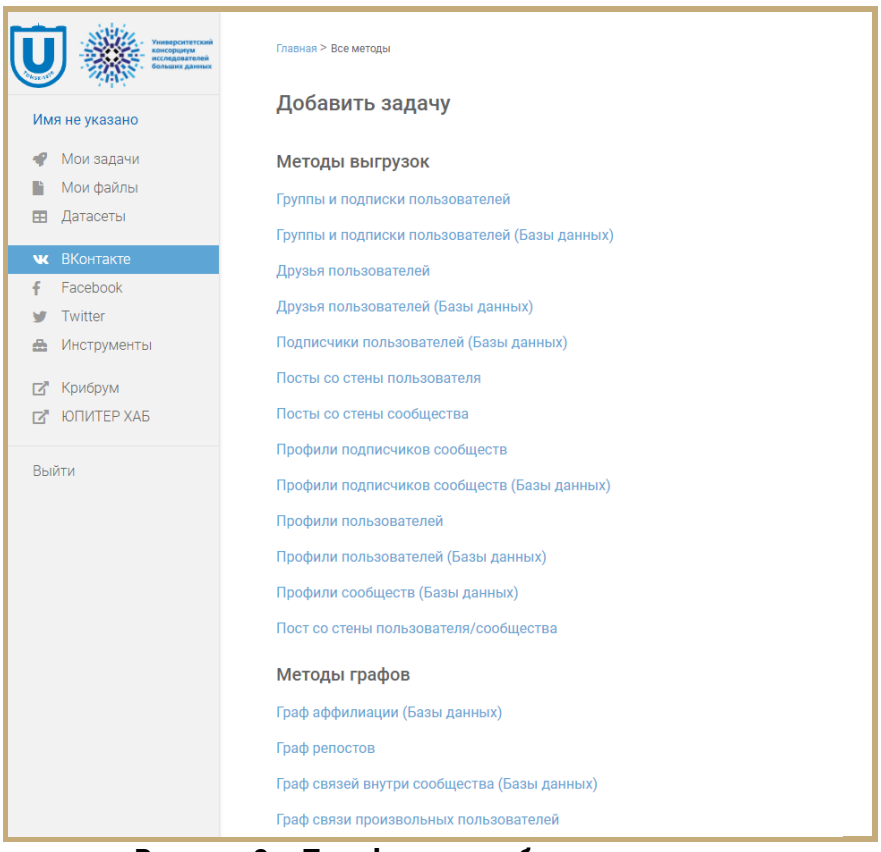

Рисунок 3 – Платформа по сбору данных

В целом, за время проведения исследования анализу было подвергнуто:

Рисунок 4 – Объем проанализированных данных

3

Выгруженный текстовый контент проходит этап предварительной обработки с использованием встроенных инструментов узлов текстового анализа платформы PolyAnalyst: индексирование текста (разметка всего текста по лексическому составу), удаление

дубликатов и похожих между собой текстов, очистка текста от знаков пунктуации, удаление html тегов и гиперссылок, эмодзи и нерелевантных слов; перевод слов в нижний регистр; удаление стоп-слов (с помощью дополненных пользовательских словарей), к которым относятся предлоги, причастия, междометия, цифры, частицы а также часто употребляемые слова, не несущие смысловой нагрузки, исправление орфографических ошибок и опечаток;

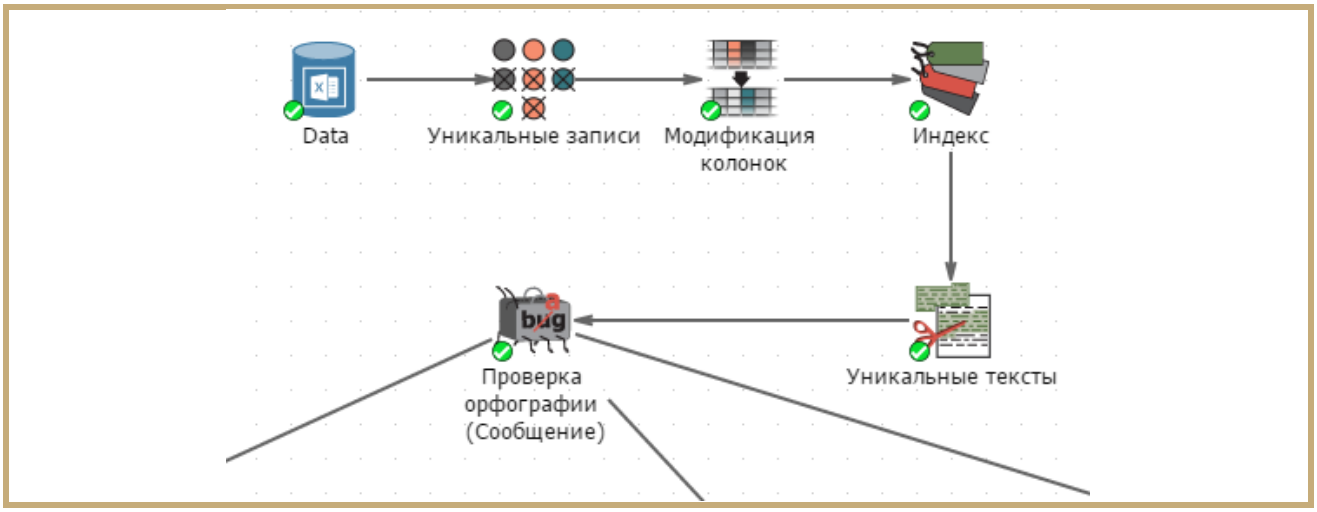

Рисунок 5 – Предобработка данных в PolyAnalyst

4

Автоматическая разметка публикаций региональных сообществ по критерию мусор/не мусор, категориям и типам тональности;

Мусор – это, прежде всего, сообщения рекламного характера, а также сообщения затрагивающие темы, выходящие за рамки данного исследования (содержащие информацию о вакансиях; спортивных и культурно-массовых мероприятиях; обмене и бесплатной передаче товаров; конкурсах и акциях; рецептах, доставках и питании; астрологических прогнозах; продаже вещей; предложениях знакомства романтического и эротического характера; обсуждения личной жизни участков сообщества и т.п.

После чего, нерелевантные сообщения (“мусор”) очищались и дальнейшая работа проводилась только с релевантными (“не мусор”).

Релевантные сообщения классифицировались по следующим 19 категориям:

Социальная сфера

Категории	Индикаторы/Темы
Образование	Дошкольное, среднее, среднее профессиональное, высшее, дополнительное
ЖКХ	Оказание услуг ЖКХ населению Управление
Здравоохранение	Обслуживание, качество лечения, работа с отдельными группами населения
Инфраструктура	Дороги, топливо, доступность (между населенными пунктами, до домов и т.д.), загруженность дорог, уборка снег, ливневые системы, доступность жилья и т. д.
Безопасность	Работа служб (Полиция, МЧС, ГИБДД (ДПС), Росгвардия) Обстановка (преступления и правонарушения)
Экология	Вырубка лесов, сокращение биоразнообразия, загрязнение воды/воздуха, перенаселение, деградация земель, отходы жизнедеятельности человека, влияние экономической и политической деятельности на экологическую ситуацию, последствия человеческого влияние на экологию, меры предотвращения экологических катастроф и т.д.
Отношения между людьми	Доброжелательность/недоброжелательность прохожих, соседей, коллег, случайных людей (попутчики, пассажиры и т.д.)
Общее эмоциональное состояние	Выражение чувств (счастлив/несчастлив, доволен/недоволен, раздражен/воодушевлен, печален/рад и т.д.)

Экономическая сфера

Категории	Индикаторы/Темы
Работа	Уровень безработицы, зарплата, условия труда, официально/неофициальное трудоустройство, полная/неполная ставка
Товары	Цены и влияние инфляции Конкуренция между производителями
Налоги	Бюрократия, уровень налоговой ставки, налоговая нагрузка населения, распределение собранных налогов

Кредитование и предпринимательство	Кредиты и ипотека Предпринимательство (барьеры, бюрократия и т.д.)
Социальная поддержка от государства	Субсидии, пенсии, льготы, пенсионный возраст

Политическая сфера

Категории	Индикаторы/Темы
Свобода СМИ	Цензура/свобода слова, достоверность информации, объективность и т.д.
Протестный потенциал *(недовольство населения)*	Политический протест (протест к представителям власти) Социальный протест (протест против социального неравенства, проблем, существующих в обществе) Культурный протест (протест в культурной жизни и повлёкший возмущение населения)
Свобода выборов	Честность выборов, давление на избирателей, прозрачность, вбросы, явка, конкуренция на выборах
Отношение к власти	Отношение населения к отдельным политическим персонам и их деятельности
Политические решения	Оценка проблемы в регионе Управленческие и кадровые решения властей Законодательство (новые законы и изменения старых) Нарушения прав граждан Отношение к населению Бюджет (формирование и распределение) Оценка населением принятых решений
Внутренняя политика	Оценка функций, возложенных на региональные власти: - способность организовывать хозяйственно-экономическую жизнь в регионе - способность поддерживать стабильность - соблюдение социальной справедливости при распределении благ - безопасное использование ресурсов страны, - поддержание законности и порядка

Для автоматической классификации использовался узел “Классификации текстов” совместно с модулем предобработки данных извлечением ключевых слов”.

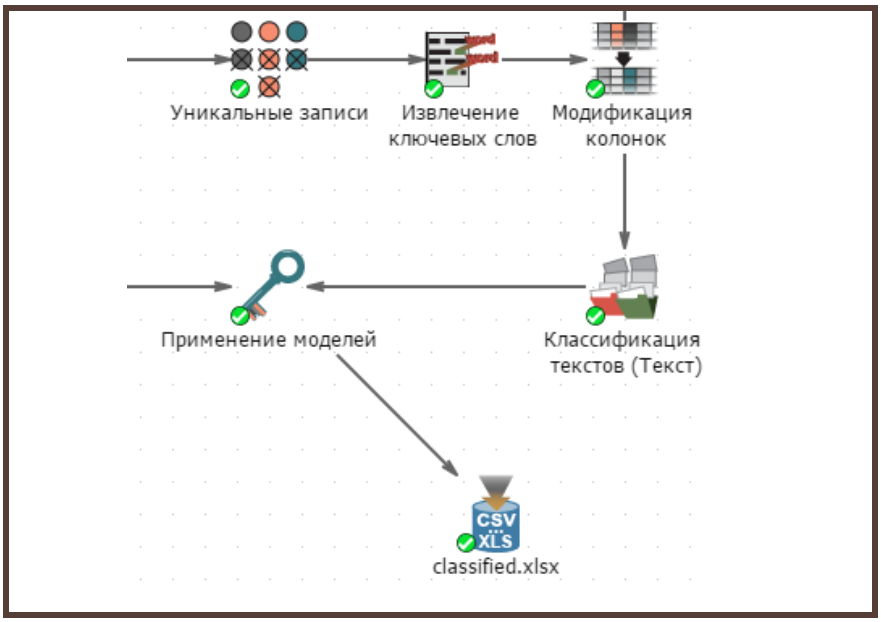

Рисунок 6 – Классификация текстового контента по категориям с помощью PolyAnalyst

После чего проводился анализ тональности сообщений по 3-м категориям: положительная, нейтральная и негативная.

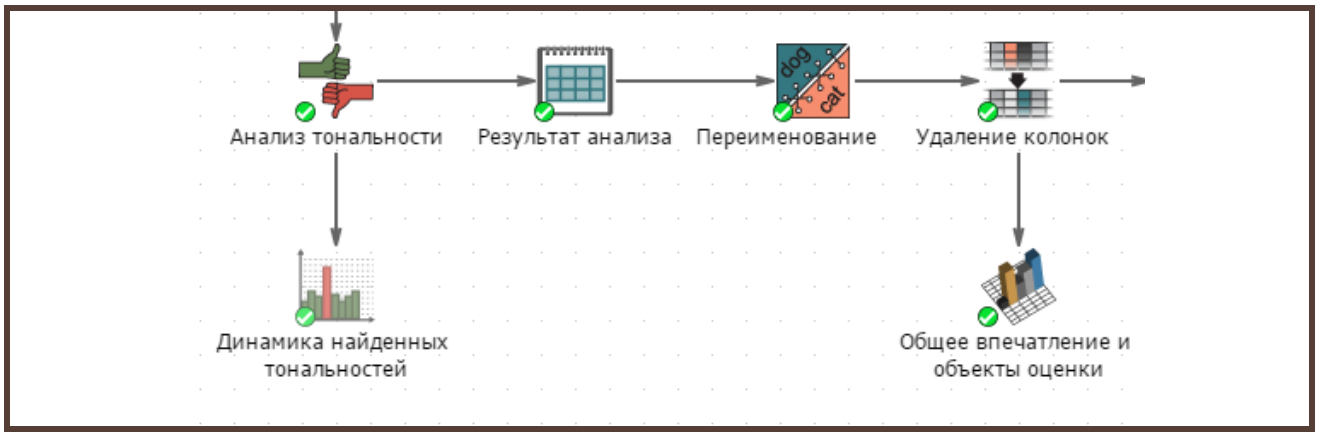

Рисунок 7 – Классификация текстового контента по типам тональности с помощью PolyAnalyst

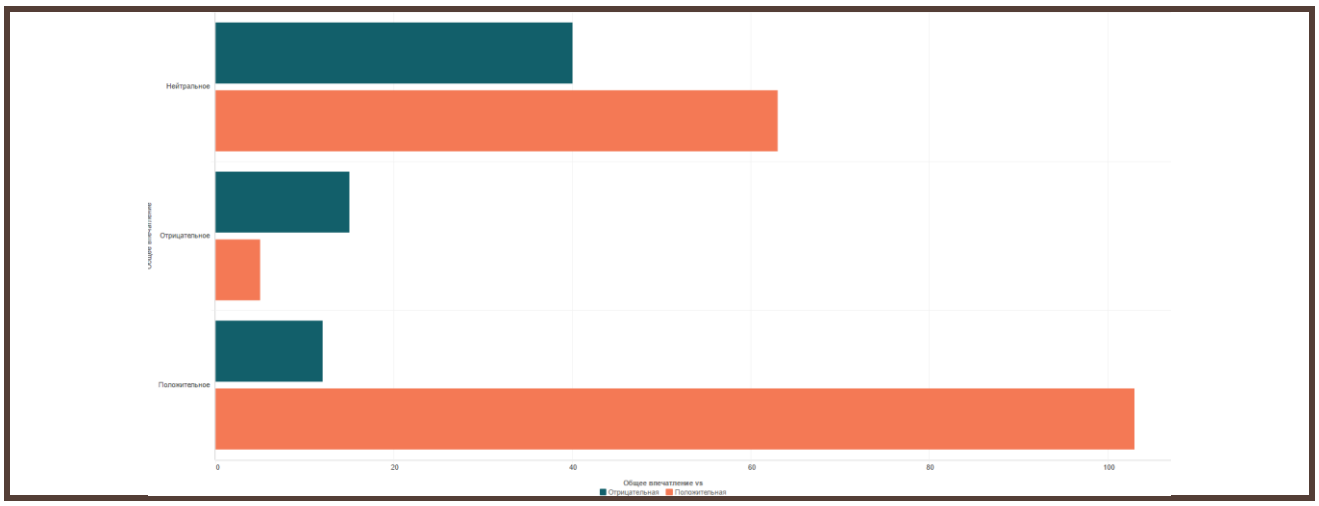

Рисунок 8 – Пример результата классификации текстового контента по типам тональности в PolyAnalyst

5

Расчет индекса онлайн активности пользователей в региональных сообществах;

$$\frac{like + k_1 * repost + k_2 * comments}{\text{Общее кол-во подписчиков в сообществах региона}}$$

k_1 – коэффициент, учитывающий вес «репостов» в соотношении с количеством «лайков»

k_2 - коэффициент, учитывающий вес комментариев в соотношении с количеством «лайков»

Рисунок 9 - Формула расчета индекса онлайн-активности

После расчета индекс онлайн-активности пользователей регионах можно отследить динамику мнений в регионе по той или иной теме и выявить какие публикации вызывают всплески активности.

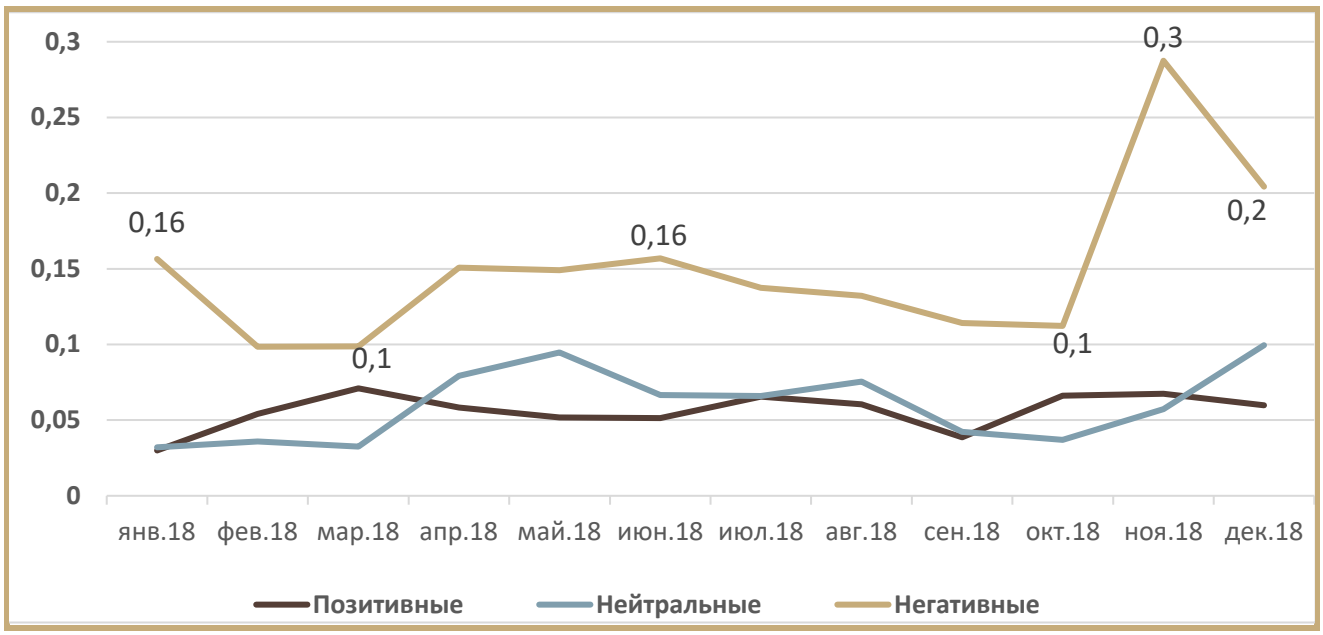

Рисунок 10 - Активность: категория «Инфраструктура» (Томская область)

Рисунок 11 - Пост, вызвавший всплеск активности: Инфраструктура (Томская область)

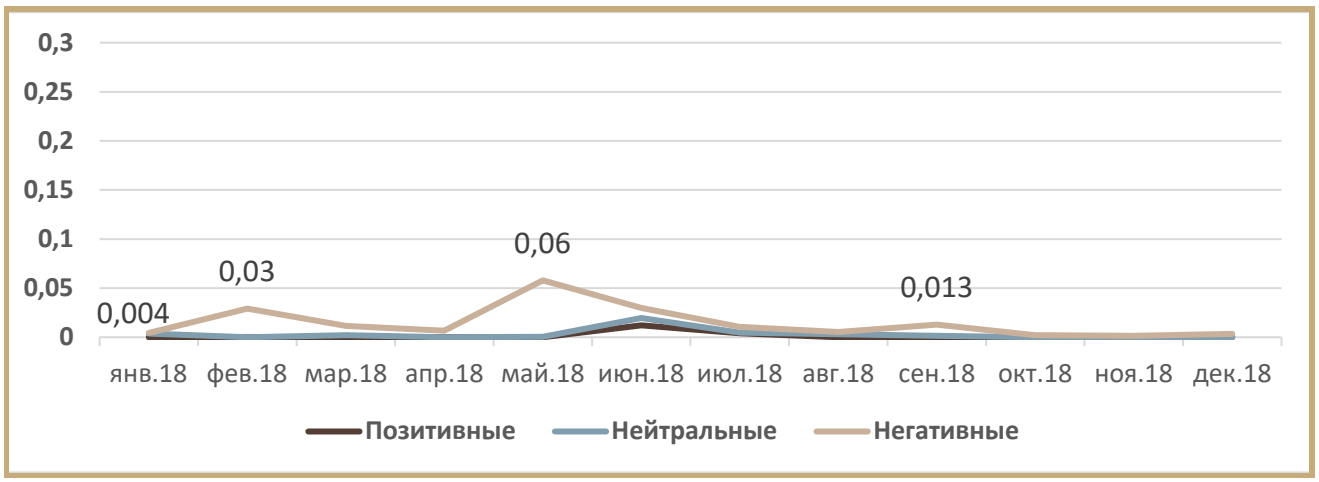

Рисунок 12 - Активность: категория «Протестный потенциал» (Томская область)

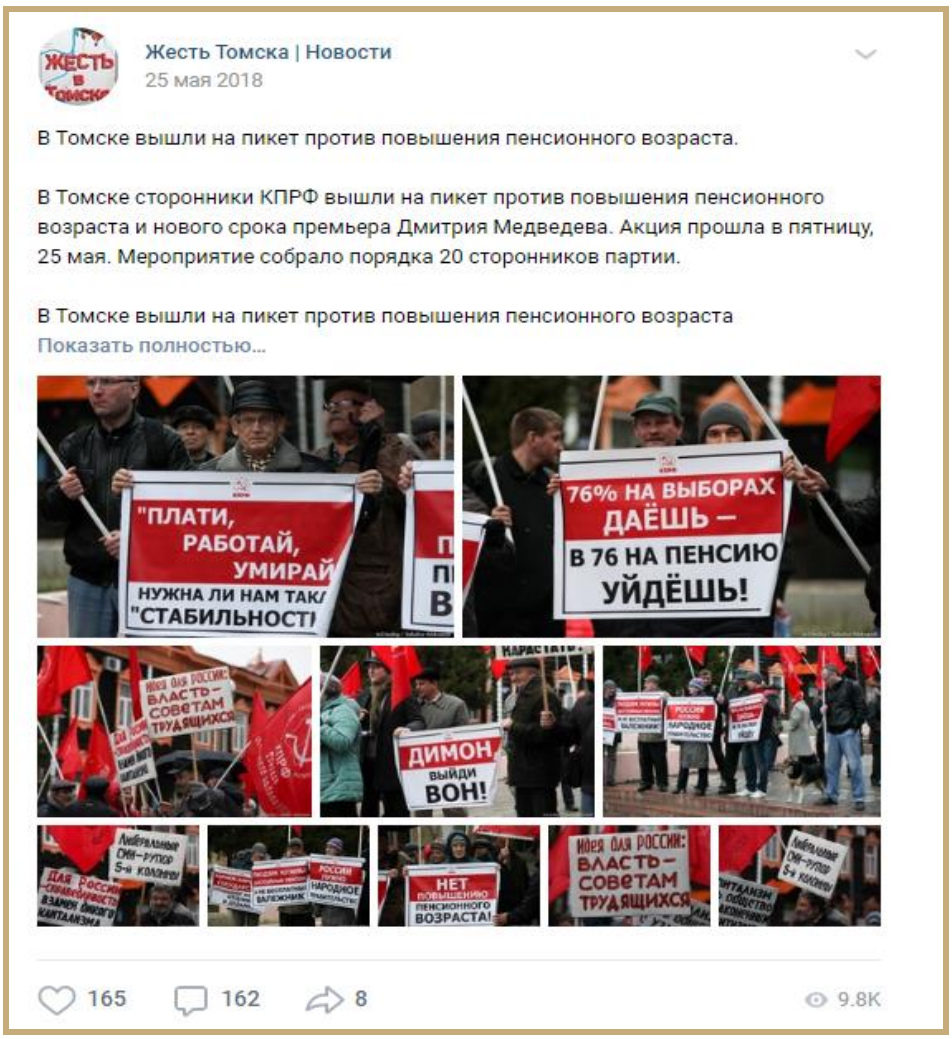

Рисунок 13 - Пост, вызвавший всплеск активности: Протестный потенциал (Томская область)

Расчет индекса субъективного качества жизни населения регионов.

$$\frac{I_{1(+)} + I_{2(+)} + I_{3(+)} \ldots + I_{n(+)}}{12} - \frac{I_{1(-)} + I_{2(-)} + I_{3(-)} \ldots + I_{n(-)}}{12}$$

I – среднемесячный индекс

Рисунок 14 - Формула расчета индекса субъективного качества жизни

Стоит отметить, что отобранные регионы предварительно были разделены на национальные и не национальные. На основе полученных данных был получен индекс субъективного благополучия по каждому региону Российской Федерации:

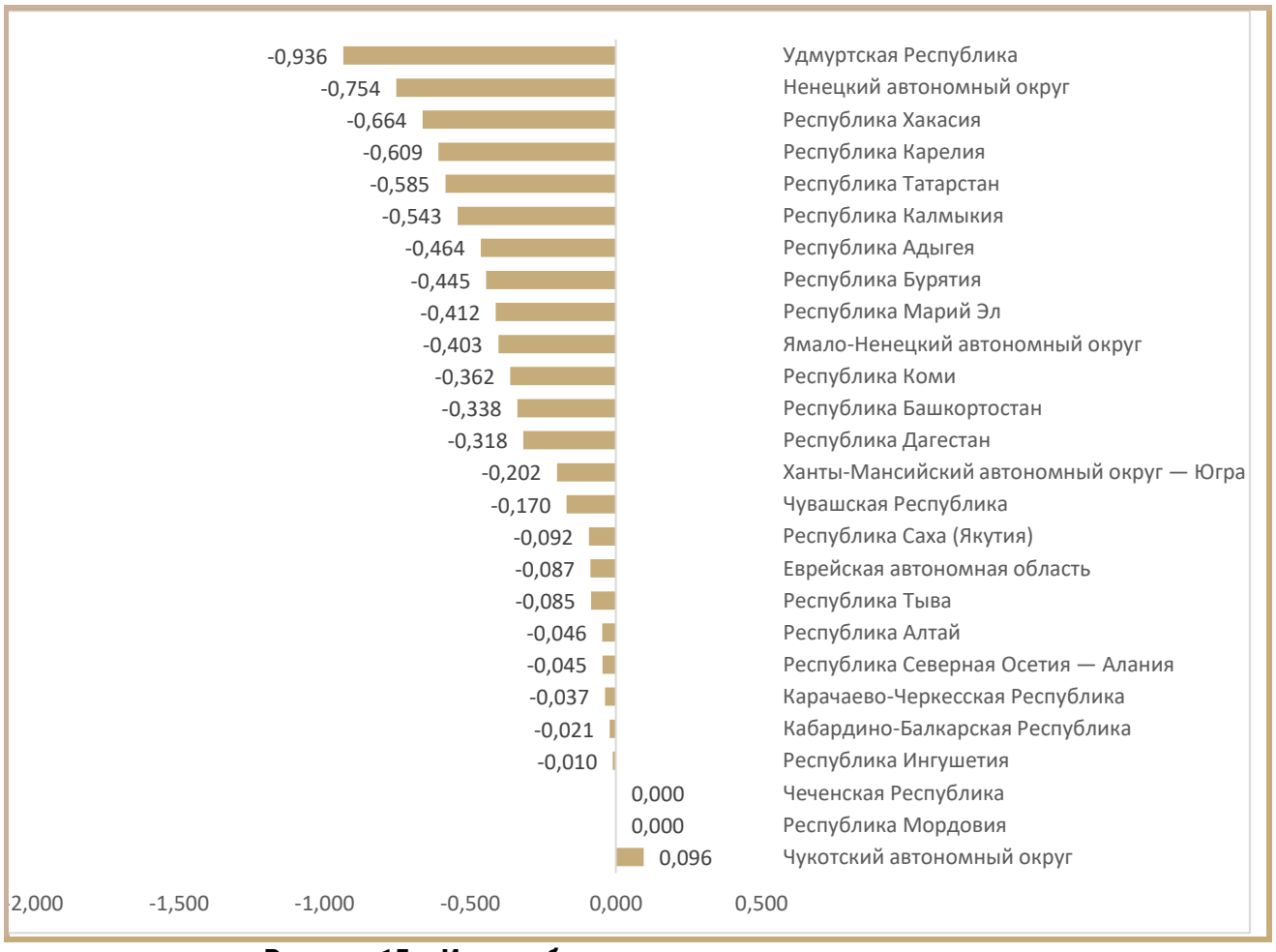

Рисунок 15 – Индекс благополучия в национальных регионах

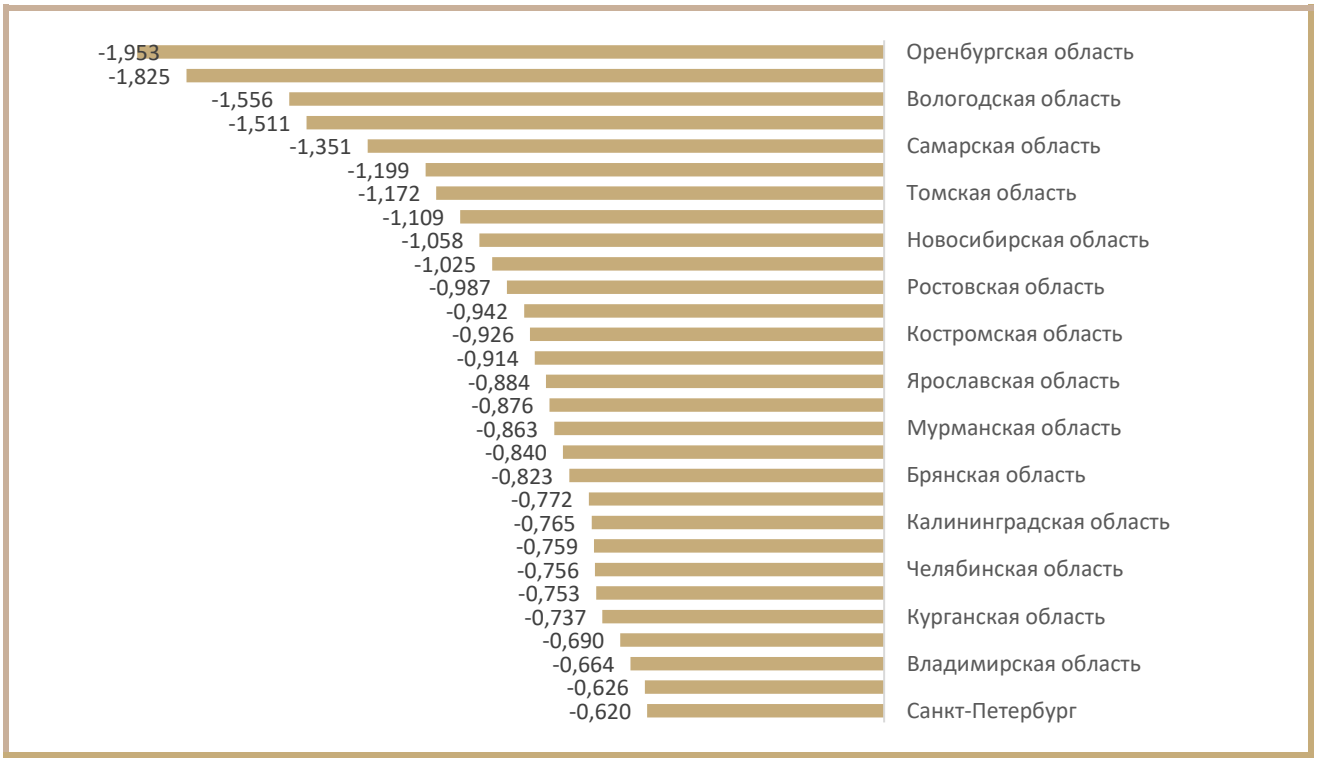

Рисунок 16 – Индекс благополучия в не национальных регионах (Часть 1)

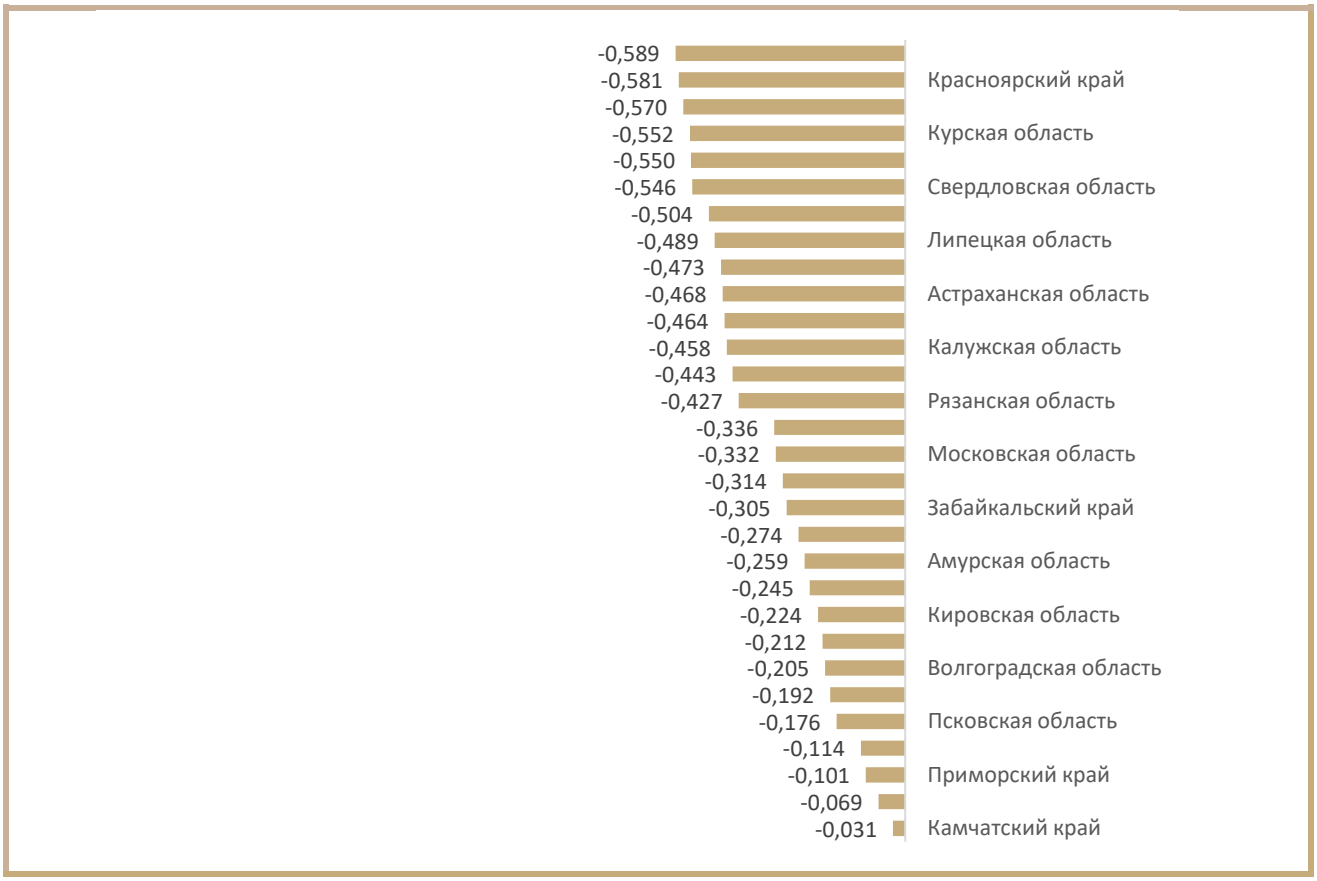

Рисунок 17 – Индекс благополучия в не национальных регионах (Часть 2)

Исследовательский проект по анализу мнений и настроений в регионе также включает в себя выявление лидеров общественного мнения внутри субъекта РФ, которые могут стать эффективным каналом коммуникации и распространения информации среди населения. Разработанная методика включает выполнение работ по выявлению двух типов лидеров мнений:

1. Лидеры общественного мнения среди пользователей

Рисунок 18 - Определение лидеров мнений среди пользователей

Пользователей, генерирующих собственный контент на своей персональной странице достаточно небольшое количество. Например, в Томской области на 2019 год их было всего 5 человек.

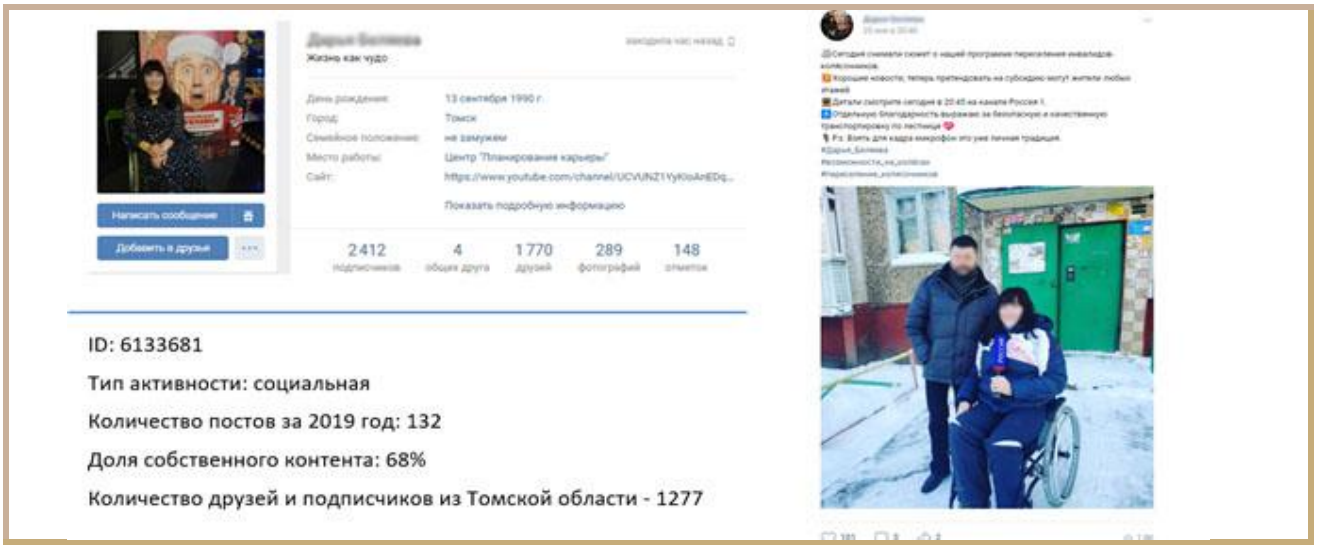

Рисунок 19 - Лидер общественного мнения среди пользователей

2. Лидеры общественного мнения среди авторов

Рисунок 20 - Определение лидеров мнений среди авторов

Стоит отметить, что лидер общественного мнения среди авторов не просто генерирует контент для региональных сообществ, но и получает высокий отклик (лайки, репосты, комментарии) на свои публикации.

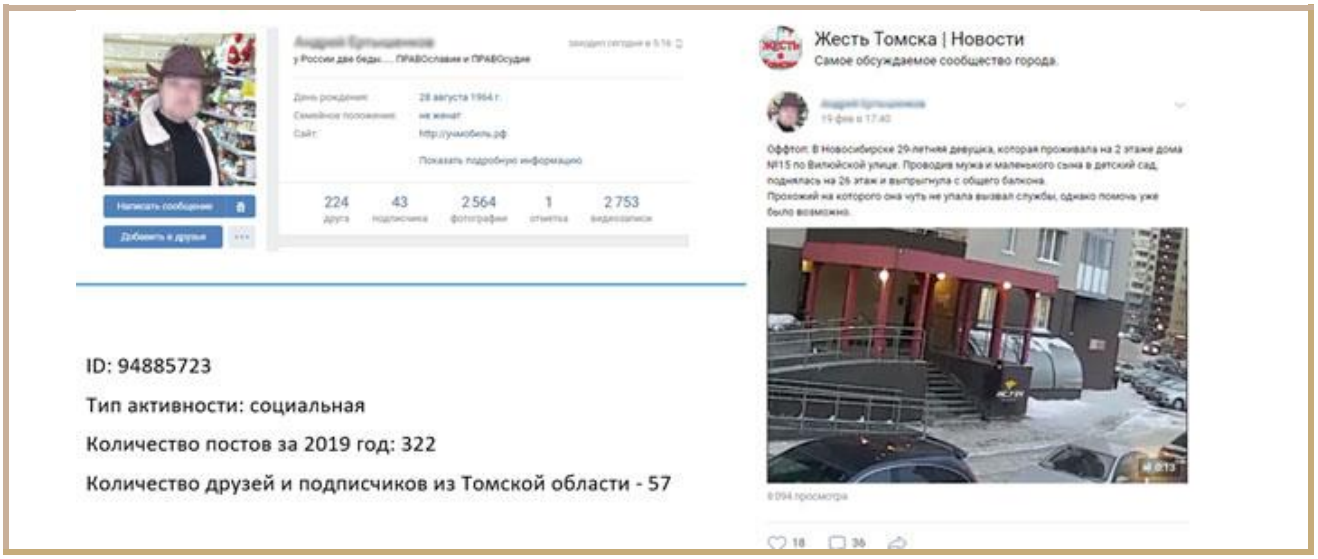

Рисунок 21 - Лидер общественного мнения среди авторов

Таким образом, на основе результатов, полученных в ходе реализации данного исследовательского проекта можно сделать вывод о том, что созданная методика измерения субъективного благополучия актуальна в современных условиях цифровизации и развития информационно-коммуникационных технологий и позволяет оценивать изменений мнений и настроений в регионе. На данный момент совершенствуется алгоритм автоматической «очистки» текстового контента отобранных региональных

сообществ, а также алгоритмы классификации публикаций по категориям и тональностям, что позволит в дальнейшем создать инструмент, позволяющий в современных условиях измерять уровень благополучия населения регионов. В то же время разработанная методика выявления лидеров общественного мнения работает на отдельных регионах и также постоянно совершенствуется, что на данный момент позволяет выявлять лидеров мнений по конкретным тематикам (сферы общественной жизни/тематические категории). Регулярный мониторинг социальных сетей позволит выявлять ключевые инфоповоды и пользователей, способных распространять информацию и коммуницировать с пользователями внутри конкретного региона.

В рамках данного проекта возможно проведение анализа информационного освещения деятельности организаций и отдельных персон. Методология может быть основана на поиске, сборе и классификации сообщений в социальных медиа, касающихся целевого объекта.