

Приложение поиска, анализа и прогнозирования данных в социальных сетях

С. А. Сластников, Л. Ф. Жукова, И. В. Семичаснов

Национальный исследовательский университет Высшая школа экономики, Москва, Россия

Аннотация. В статье описывается веб-сервис, предназначенный для поиска, загрузки и анализа данных из социальных сетей и мессенджеров, а также демонстрируется его применение для исследования сообществ социальной сети «ВКонтакте». Веб-сервис позволяет выявлять типичные профили пользователей сообществ, оценивать эмоциональную окраску постов и комментариев, а также прогнозировать вектор развития сообществ. Описанный в работе веб-сервис обладает обширными функциональными возможностями и оригинальной нейросетевой моделью классификации текстов произвольной длины по эмоциональной окраске. Представлены примеры работы инструмента для анализа развития сообществ автомобильных брендов. Анализ охватывает миллионные аудитории подписчиков, десятки тысяч постов и сотни тысяч комментариев, что подтверждает релевантность выборок и достоверность результатов.

Ключевые слова: социальная сеть, сообщества, анализ данных, нейронная сеть, профиль, публикация, эмоциональный окрас, тональность, прогнозирование.

DOI 10.14357/20718632240110

EDN YMIDVN

Введение

Социальные сети и мессенджеры уже давно перестали быть просто средством личной коммуникации. Сегодня они представляют собой платформы для профессиональных и экспертных сообществ, СМИ, государственных учреждений и бизнеса. С помощью накопленных данных в социальных сетях можно анализировать общественные реакции на события, бренды и многое другое, а также прогнозировать поведение аудитории. Исследования в различных областях, таких как социология, психология, образование и государственное управление, непрерывно проводятся на основе этих данных [1-3].

Уже давно исследователи задаются вопросом о возможности составления психологического портрета человека на основе информации, содержащейся в его профиле в социальной сети.

Эти исследования начались еще задолго до того, как социальные сети завоевали многомиллионную аудиторию пользователей. К примеру, в 2013 году ученые из Кембриджского университета [4] доказали, что, анализируя «лайки», отражающие предпочтения пользователей, можно вывести некий портрет потребителя, который включает расовую принадлежность, политические взгляды и многое другое, даже то, что не указано в его профиле. Кадровые агентства и специалисты по подбору персонала также часто анализируют подписки соискателей в социальных сетях [5], исходя из сферы их интересов, чтобы сделать множество выводов о личности соискателя: его личностные, политические, религиозные взгляды, отношение к обществу, государству, семейным ценностям, склонность к насилию, девиантному поведению, уровень экспрессивности и другие характеристики.

* Статья подготовлена в рамках стратегического проекта «Национальный центр научно-технологического и социально-экономического прогнозирования» интегрированной программы развития ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики»

Однако большинство исследований в области социальных сетей сталкиваются с ограничениями: это ограниченный объем анализируемых данных. Такая ситуация не всегда позволяет делать корректные выводы о большой аудитории и пропускает интересные частные особенности, которые не выражены в основной массе данных. В данной статье представлен разработанный инструмент для поиска, загрузки и анализа данных из социальных сетей с целью выявления профилей пользователей сообществ, определения эмоциональной окраски постов и комментариев, а также прогноза развития сообществ. С помощью данного инструмента было проведено исследование сообществ автомобильных брендов в социальной сети "ВКонтакте".

Статья организована следующим образом. В первой главе описана архитектура и функциональные особенности сервиса. В главе 2 описан модуль анализа тональности сообщений. Глава 3 иллюстрирует применение описанного сервиса для анализа активности в сообществах социальной сети «ВКонтакте» в сфере автомобильного рынка. Основные выводы и краткое описание результатов исследования представлены в Заключение.

1. Описание сервиса

Инструмент представляет собой веб-сервис, позволяющий единообразно взаимодействовать с самыми популярными социальными сетями – "ВКонтакте", Instagram, Facebook¹, Twitter, и мессенджер Telegram. Инструмент обладает следующим набором функций:

- поиск текстовой информации по заданным критериям поиска из социальных сетей;
- визуализация извлеченных данных;
- определение эмоциональной окраски текстов и комментариев;
- формирование различных форм отчетов с результатами поиска;
- выполнение поисковых запросов по заданному расписанию;
- автоматическое формирование и отправка отчетов пользователю.

Разработанное веб-приложение имеет микросервисную архитектуру, представленную на Рис. 1.

Запрос на поиск по социальным сетям приходит с клиентской стороны, в качестве которой выступает веб-сайт или Telegram-бот, с указанием перечня аккаунтов (пабликов), по которым необходимо осуществить поиск, и поисковым

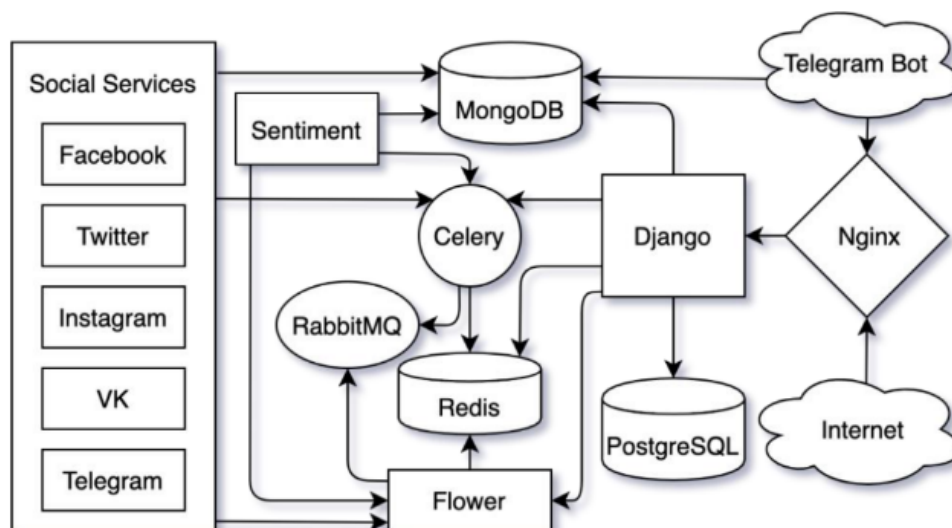


Рис. 1. Общая архитектура веб-сервиса

¹ Facebook, Instagram — проект Meta Platforms Inc., деятельность которой в России запрещена

запросом. Главным сервисом является Django-приложение с gunicorn-сервером, который объединяет все составные части между собой. Он принимает запросы от сервера Nginx, который выступает в качестве прокси-сервера для балансировки нагрузки запросов с клиентской стороны. Django-сервис взаимодействует с СУБД PostgreSQL для хранения данных о пользовательских аккаунтах и настройках, данных авторизации и управлением стоп-задачами на поиск, которые создаются пользователями, а также с MongoDB для работы с данными полученными из социальных сетей. Управлением и мониторингом задач поиска занимается Flower, который принимает HTTP запросы из Django приложения. Он осуществляет коммуникацию и синхронизацию между процессами Celery. Передача данных между процессами осуществляется через RabbitMQ и Redis. Сервисы для поиска и определения эмоциональной окраски состоят из запущенных Celery процессов, которые слушают соответствующие очереди сообщений RabbitMQ и сохраняют результаты своей работы в MongoDB. После завершения они отправляют сообщение сервису сентимент-анализа для определения эмоциональной окраски, которая после обработки всех текстов из социальных сетей отправляет сообщение на создание агрегированных отчетов в формате xlsx и pdf и их последующей отправке пользователям на почту и в Telegram.

Благодаря тому, что каждый процесс обработки запускается отдельным процессом в контейнере под Docker-compose'ом и слушает очередь, возможно масштабирование нагрузки и распараллеливание вычислений в зависимости от доступных вычислительных мощностей, с помощью запуска дополнительных контейнеров.

2. Модуль сентимент-анализа

Разработанный веб-сервис включает модуль, позволяющий определять эмоциональную окраску русскоязычных текстов. Всего выделяется 6 классов тональности:

- негатив – тексты с негативной эмоциональной окраской;
- нейтральность – тексты без ярко выраженной эмоциональной окраски;

- позитив – тексты с положительной эмоциональной окраской;
- пропуск – непонятные тексты, тексты на иностранных языках;
- этикет – приветствия, поздравления;
- юмор – тексты, имеющие выраженную юмористическую окраску.

Для конвертации текстовых данных в числовые использовались предобученные эмбединги – ELMo [6], которые строятся, принимая во внимание контекст слова. Такие эмбединги обучаются на большом количестве текстов, не привязываясь к конкретной задаче. После того, как каждое слово было преобразовано в вектор, весь текст может быть представлен в виде матрицы и восприниматься как изображение. Таким образом становится возможным применение к тексту моделей, которые обычно работают с изображениями. В данной работе в качестве модели для классификации выбрана сверточная нейронная сеть (CNN). Этот тип нейронных сетей имеет матрицы-фильтры, задача которых распознавать различные шаблоны в рассматриваемом объекте. В случае работы с текстами фильтры имеют такую же ширину, как и размерность векторного представления слова, а высота матриц-фильтров может варьироваться и отвечает за длину отыскиваемого шаблона. Так, например, фильтр высотой 3 будет анализировать сочетания из трех слов. Подсчитывается корреляция между различными частями текста и множеством фильтров, после чего полученные числа передаются в полносвязный слой, где преобразуются в вероятности принадлежности объекта к тому или иному классу. Результаты экспериментов с моделями показали, что для корректного определения эмоциональной окраски любого текста из социальных сетей сначала необходимо выявить тип этого текста: новость (пост) или комментарий, и уже после этого строить соответствующие эмбединги и применять модель для классификации. В связи с этим, была обучена простая и быстрая модель (LGBM Classifier [7]) для классификации текстов по их типу после предварительной TF-IDF (Term Frequency - Inverse Document Frequency) векторизации [8]. Определить тип текста существенно легче, чем эмоциональную окраску, поэтому такие простые методы показывают весьма



Рис. 2. Общая структура модуля классификации текста

высокие результаты (качество выше 0,99). Таким образом, предлагаемая система имеет структуру, представленную на Рис. 2.

Для обучения моделей изначально было сформирован большой датасет из 5 известных наборов данных: RuSentiment [9], Uniform RuSentiment [10], FUN [11], Combined RuSentiment and FUN, News. Первые эксперименты показывали, что модель слишком часто предсказывает нейтральный класс. Поэтому было принято решение расширить классы «Позитив», «Негатив» и «Этикет» путём добавления в них постов с Pikabu [12]. Каждый пост Pikabu имеет набор тэгов, по которым можно автоматически определить его эмоциональную окраску. Была реализована функция для парсинга постов и сформированы новые обучающая и тестовая выборки. Поскольку посты с Pikabu имеют ярко выраженную окраску, на них была дообучена модель для комментариев. Чтобы уменьшить количество нейтральных предсказаний новостной модели, ее переобучили, подобрав веса для

каждого класса так, чтобы уравновесить их. Таким образом, модель штрафует сильнее, если признает позитивный или негативный текст нейтральным, и слабее, если признает нейтральный текст позитивным или негативным. Итоговое качество построенного модуля классификации представлено в Табл. 1.

Представленные результаты показывают довольно высокое качество модели с учетом наличия 6 классов и без использования привязки текстов к какой-то конкретной тематике. Более подробное описание разработанной модели можно найти в работе [10].

Помимо качества модели, важным аспектом для использования ее в рамках сервиса является и скорость работы. Зависимость скорости работы от количества слов в тексте определялось с использованием видеокарты NVIDIA Tesla K80. Время для построения эмбедингов и определения типа сообщения для текста длиной 200 слов не превосходит 0,6 секунды.

Табл. 1. Качество нейросетевой модели классификации текстов

	Precision	Recall	F1 score
Юмор	0.83	0.6	0.7
Негатив	0.75	0.87	0.81
Позитив	0.75	0.72	0.73
Нейтральность	0.7	0.73	0.71
Пропуск	0.53	0.42	0.47
Этикет	0.91	0.89	0.9
Macro AVG	0.73	0.71	0.72
Weighted AVG	0.74	0.74	0.74
Accuracy	0.74	0.74	0.74

3. Анализ сообществ автомобильных брендов в социальной сети «ВКонтакте»

В рамках данного раздела проводится исследование сегмента крупнейших дилеров авторынка РФ, которые были отобраны на основании опубликованной статистики за 2021 год [13]. Рассматриваемые сообщества представлены в

Табл. 2 – это топ 20 лидеров рейтинга в сегментах бюджетного, среднего и бизнес-класса, суммарная аудитория подписчиков составляет более 3,5 млн. пользователей.

3.1 Профиль подписчиков

Исследуем профиль подписчиков данных сообществ на основе данных, представленных в социальной сети. К данным, которые заполнены

Табл. 2. Список сообществ

Название группы Вконтакте	Ссылка	Число подписчиков
Лада	https://vk.com/lada	200195
Киа Россия	https://vk.com/kiarussia	273790
Hyundai Russia	https://vk.com/hyundairussia	442882
Volkswagen Russia	https://vk.com/volkswagenrussia	230123
ŠKODA Russia	https://vk.com/skodarussia	206380
Официальное сообщество Nissan в России	https://vk.com/nissanrussia	127878
Официальное российское сообщество Mazda	https://vk.com/mazda	201512
Официальная страница Renault Россия	https://vk.com/renaultru	311043
Официальная страница BMW Россия	https://vk.com/bmw	157916
Официальная страница Audi Россия	https://vk.com/audi	187186
Официальная страница Toyota Россия	https://vk.com/toyotarussia	195866
Официальная страница Lexus Россия	https://vk.com/lexusrussia	42822
Официальное сообщество Haval Россия	https://vk.com/havalrussia	59779
Официальное сообщество Geely Россия	https://vk.com/geely_official	43263
Официальное сообщество Chery Россия	https://vk.com/cheryrussia	280963
АО «Мерседес-Бенц РУС»	https://vk.com/mercedesbenzrussia	86766
Импортер и дистрибьютор легковых автомобилей Volvo	https://vk.com/volvocarsrus	67913
Официальное сообщество Ford Россия	https://vk.com/ford	242254
Официальное сообщество дистрибьютора автомобилей Subaru в России и Республике Беларусь	https://vk.com/subaru_russia	79319
Новостной канал Mitsubishi Motors в России	https://vk.com/mitsucarrus	142119
	Суммарная аудитория	3579969

у абсолютного большинства подписчиков можно отнести следующие: пол, возраст, город. Распределение по полу показало ожидаемые результаты: 80% аудитории составляют мужчины: сфера автомобилей традиционно считается более интересной мужчинам. Интересно отметить, что всего у 0.1% подписчиков пол в профиле не указан. Далее рассмотрим распределение аудитории по городам. Для этого были отобраны только города с более чем 250 тыс. подписчиков. Диаграмма распределения представлена на Рис. 3.

В данном распределении можно отметить, что аудитория распределяется довольно равномерно по крупнейшим городам России пропорционально количеству населения.

Исследование данных о возрасте показало следующие статистику: общий диапазон от 14 до 91 года, средний возраст 33 года, мода и медиана распределения при этом совпадают и составляют 32 года. В целом можно утверждать, что распределение подчиняется нормальному закону распределения.

Далее исследуем, насколько подписчики заинтересованы отраслью в целом, либо же привержены исключительно одному бренду. Для этого, рассмотрим распределение подписчиков по количеству одновременного нахождения в пабликах (подписках) среди 20 рассматриваемых групп. Более половины пользователей подписаны на одно сообщество. Среднее значение

выборки составляет 1.3, мода и медиана – 1. Исходя из этого, можно сделать вывод о том, что средний участник сообщества состоит в группе конкретного автопроизводителя, с большой вероятностью именно того, чьим автомобилем владеет.

Тем не менее, попробуем все же классифицировать участников сообществ по числу подписок. Для этого разделим весь диапазон на 3 группы:

- малая группа от 1 до 3 подписок - предположительно относится к людям, имеющим в семье одну или несколько машин и интересующимся лишь конкретными производителями;
- средняя группа от 4 до 9 подписок - предположительно состоит из любителей автомобилей разного типа и марок;
- крупная группа от 10 до 20 подписок - предположительно состоит из людей, занимающихся автомобилями в рамках профессиональной деятельности.

На Рис. 4 представлено распределение по городам в третьей группе. Сравнив данное распределение с общим распределением аудитории по городам (Рис. 3) можно отметить следующие особенности: 1) лидерство двух крупнейших городов (Москва, Санкт-Петербург) сохранилось, однако в общей пропорции произошло снижение на 20-25%; 2) появилась группа «новых» городов (Ижевск, Тольятти, Владимир, Северодвинск), не входящих существенно в общее

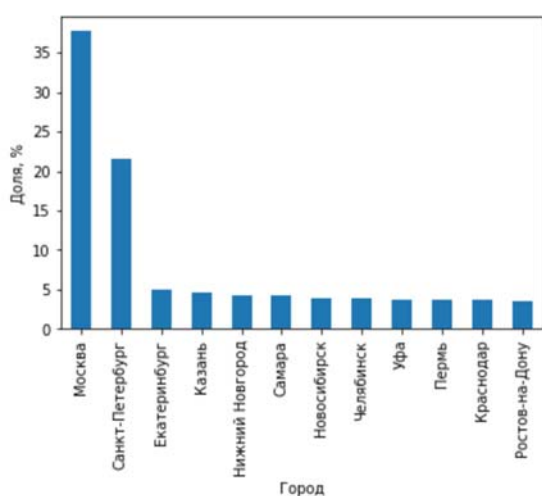


Рис. 3. Распределение аудитории по городам

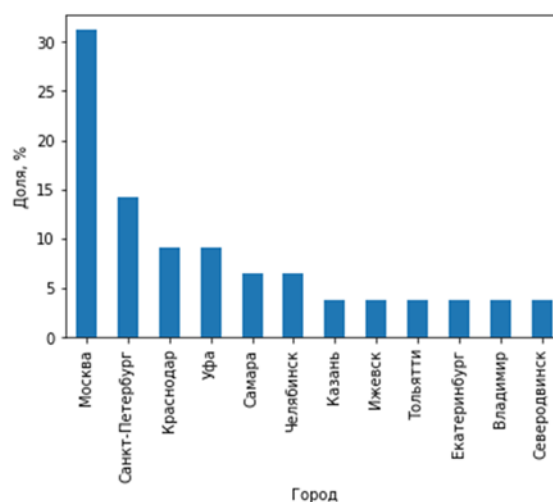


Рис. 4. Распределение по городам пользователей, состоящих в 10-20 группах

распределение. В «новые» города попали ключевые российские центры активности в сфере автопрома. Попадание в рейтинг Тольятти и Ижевска легко объяснимо наличием больших автозаводов, но в отношении Владимира и Северодвинска того же утверждать нельзя.

3.2 Анализ постов

Выгрузим и проанализируем посты всех рассматриваемых пабликов за календарный год: с 01 сентября 2021 года до 31 августа 2022 года. За данный период было опубликовано 4628 постов. Распределение эмоциональной окраски постов по классам представлено на Рис. 5.

В силу того, что администраторы пабликов заинтересованы в продвижении своего бренда, кажется очевидным, что большинство постов и должно иметь позитивную окраску. Интересно отметить при этом, что негативных постов было найдено существенно больше, чем юмористических. Посмотрим пример поста, классифицированного нейронной сетью как пост с отрицательной коннотацией (Рис. 6.).

С точки зрения семантического анализа, механизм которого, заложенный через машинное обучение нейросети, используется для определения коннотации текста, формулировка «забыть сделать что-либо» имеет негативную эмоциональную окраску. Это связано с подавляющей частотой употребления данной языковой конструкции в текстах, содержащих негативные смыслы – сожаление, гнев, осуждение и т. п. В позитивных смысловых образованиях использование данной модели совершенно нетипично.

Рассмотрим далее динамику изменения количества постов по месяцам в рамках исследуемого периода (Рис. 7).

Ожидается, что с момента объявления СВО произошло резкое снижение числа постов, так как многие зарубежные бренды приостановили свою деятельность в РФ. На основании распределения можно утверждать, что снижение активности в целом по отрасли примерно двукратное.

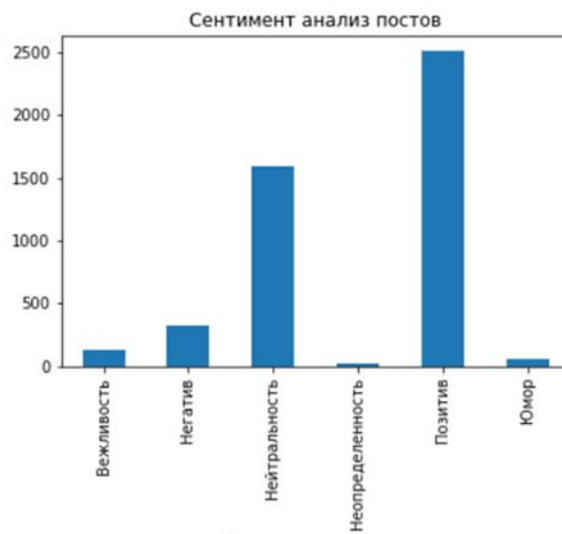


Рис. 5. Распределение эмоциональной окраски

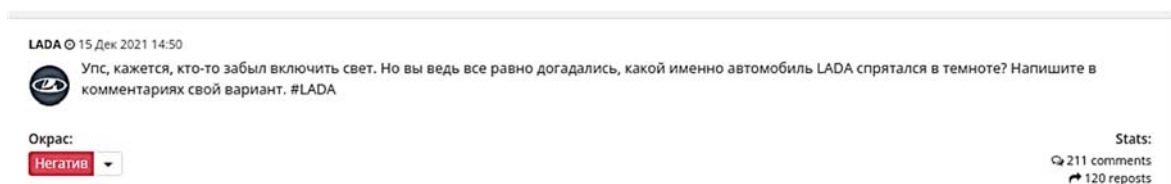


Рис. 6. Пример негативного поста



Рис. 7. Динамика общего числа постов за год

Анализ динамики изменения реакций на посты показал, что если снижение абсолютного количества репостов (Рис. 8) практически повторяет общее снижение числа постов (Рис. 7), то распределение абсолютного числа лайков имеет другой вид (Рис. 9). После «провала» количества лайков в марте 2022 года замечен быстрый возврат к предыдущим показателям. Исходя из этого можно выдвинуть следующую гипотезу: произошла существенная миграция пользователей от сообществ, приостановивших свою деятельность к тем, кто остался на рынке.

Вычислив средние значения показателей постов, получаем: среднее число просмотров поста 130231, среднее число лайков поста 225, среднее число комментариев 36, среднее число репостов

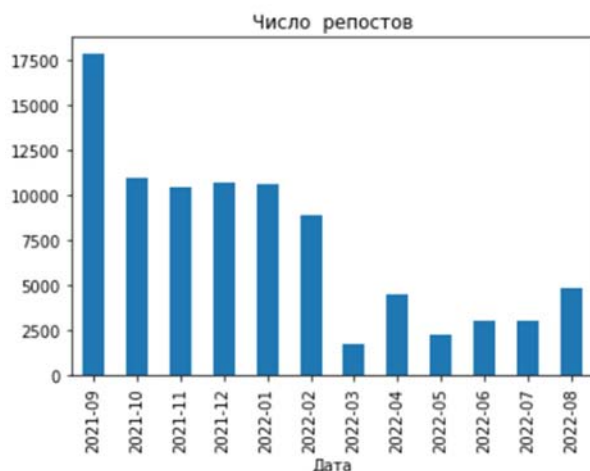


Рис. 8. Динамика общего числа репостов за год

19. Средняя длина поста составляет 393 символа, медиана 296, мода 312. На основании исследований [14; 15] можно сделать вывод о том, что для сообществ данной отрасли характерны существенно более длинные посты, чем считаются оптимальными для рекламы в целом.

3.3. Анализ комментариев

Общее количество комментариев за рассматриваемый период составило 102142 штуки. Рассмотрим распределение эмоциональной окраски комментариев (Рис. 10).

Здесь можно убедиться в явном отличии распределения относительно эмоциональной окраски постов. Ключевой особенностью можно назвать преобладающее число комментариев, помеченных классами «вежливость» и «нейтральность» [16; 17]. Можно выдвинуть гипотезу о том, что большинство комментариев оставлены под постами с всевозможными конкурсами. Выберем топ самых популярных комментариев (Рис. 11) и убедимся в корректности выдвинутой гипотезы.

Кроме этого можно отметить успешные рекламные кампании производителя Chery – не просто единственного бренда попавшего в топ 15, но еще и с упоминаниями в нескольких вариациях. Самые «залайканные» комментарии среди всей выборки имеют либо нейтральную, либо юмористическую окраску. При этом в отличие от предыдущего рейтинга тут уже упоминаются несколько разных брендов - Лада, KIA, Toyota.



Рис. 9. Динамика общего числа лайков за год

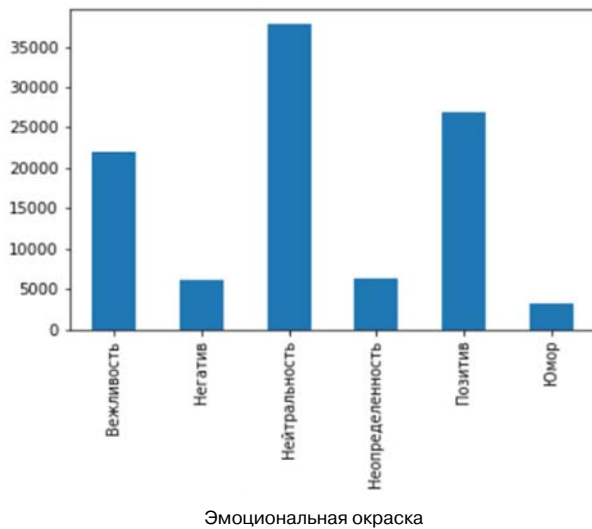


Рис. 10. Распределение эмоциональной окраски комментариев

Текст	count
Удачи	1460
Здравствуй уважаемые участники этого розыгры...	1218
Всем удачи	1060
Участвую	986
Здоровья	962
Счастья	592
Финансового благополучия	521
Желаю себе выиграть айфон	481
👍	474
Хочу поехать в Крым и выбираю Chery Tiggo 7 PRO!	462
Любви	350
Мечтаю на Chery Tiggo 8 PRO съездить в Крым с ...	332
Успехов	265
Мечтаю полюбоваться озером Байкал с Chery Tiggo...	259
Благополучия	251

Рис. 11. Топ 15 самых популярных по содержанию комментариев к постам

Перейдем к рассмотрению динамики комментариев. Явно выраженный всплеск (Рис. 12) свидетельствует о том, что в октябре 2021 года в рассматриваемых сообществах проводились конкурсы, побуждающие пользователей оставлять комментарии под постами для участия в них. В сравнении с резким снижением числа постов с марта 2022 года динамика комментариев немного запаздывает – максимальный провал произошел лишь в мае и июне.

Что касается динамики активности пользователей по часам, то наибольшая активность проявляется в диапазоне 14-19 часов, при этом в оставшуюся часть суток, за исключением ночного времени, распределена достаточно равномерно.

Общее количество пользователей, оставивших комментарии к постам – 32371, что составляет всего лишь порядка 1% от общей аудитории. Среднее количество комментариев среди них – 3,2. Интересно отметить, что 6 из 10 самых активных пользователей (включая первое место с большим отрывом) представляют женский пол, что никак не соответствует общему гендерному распределению подписчиков сообществ. Можно предположить, что столь активное комментирование призвано привлечь внимание мужской аудитории.

Средняя длина комментария составляет 51 символ, медиана – 25 символов, мода – 5. Исходя из этих данных можно сделать вывод, что состав комментариев довольно скуден и часто ограничивается лишь одним словом.



Рис. 12. Распределение числа комментариев к постам за год

3.4. Выводы

Проведенный анализ сообществ и аудиторий социальной сети «ВКонтакте» сегмента автопроизводителей в РФ выявил следующее:

- уход с рынка существенного числа зарубежных компаний заставил подписчиков весьма оперативно переключиться на оставшиеся отечественные и зарубежные бренды, что подтверждается распределением числа лайков, оставленных под постами сообществ, при условии, что существенная их часть перестала вести активность в социальных сетях;
- выявлены некрупные российские города, ведущие активную деятельность в социальных сетях в рассматриваемой сфере – Владимир, Северодвинск, Ижевск, Тольятти;
- установлено, что среди самых активных «комментаторами» постов есть и женщины, возможно с целью привлечения внимания к себе мужской аудитории;
- распределение длины комментариев к постам указывает, что в большинстве они не выражают мнений и рассуждений.

Заключение

В рамках данного исследования описан сервис, позволяющий осуществлять поиск, анализ и прогнозирование на основе данных в социальных сетях и мессенджерах. Был продемонстрирован один из примеров исследований, которые могут быть осуществлены с помощью разработанного инструмента в максимально короткие сроки, покрывая при этом огромные объемы данных. В частности, рассматриваемый случай охватывает миллионные аудитории подписчиков, десятки тысяч постов и сотни тысяч комментариев к ним. Это абсолютно обоснованно позволяет сделать вывод о релеванности выборок и достоверности получаемых результатов. В результате анализа данных в сфере автомобильного рынка были выявлены особенности целевой аудитории, вербального поведения ее членов и степени вовлеченности в период с 01 сентября 2021 года до 31 августа 2022 года.

Стоит отметить, что важным результатом данной работы является не только техническая реализация высоконагруженного сервиса с обширными функциональными возможностями,

но и оригинальная нейросетевая модель классификации текстов произвольной длины и любой тематики по признаку эмоциональной окраски. Обучение и тестирование данной модели проводилось с использованием в том числе и суперкомпьютерного кластера НИУ ВШЭ [18].

Литература

1. Smetanin S. Pulse of the Nation: Observable Subjective Well-Being in Russia Inferred from Social Network Odnoklassniki [Электронный ресурс] // Mathematics. – 2022. – Vol. 10(16). No. 2497. – URL: <https://doi.org/10.3390/math10162947> (дата обращения 19.10.2022).
2. Ковалева Ю.В., Журавлев А.Л. Общественное настроение и субъективность сетевого сообщества в период пандемии COVID-19: на примере социальной сети твиттер// Институт психологии российской академии наук. Социальная и экономическая психология. – 2020. – N 2 (18). – С. 151 - 188.
3. Докука С. В., Валеева Д. Р. Статистические модели для анализа динамики социальных сетей в исследованиях образования// Вопросы образования. – 2015. – N 1. – С. 201 - 213.
4. Социальные сети как инструмент изучения психологического портрета потребителя [Электронный ресурс]. – URL: <https://econet.ru/articles/69899-sotsialnye-seti-kak-instrument-izucheniya-psiologicheskogo-portreta-potrebitelya> (дата обращения 19.10.2022).
5. Анализ личности по социальным сетям как эффективный метод подбора кадров [Электронный ресурс]. – URL: <https://www.4dk.ru/news/d/20181005135017-analiz-lichnosti-po-sotsialnym-setyam-kak-effektivnyy-metod-podbora-kadrov> (дата обращения 19.10.2022).
6. Ilić S. et al. Deep contextualized word representations for detecting sarcasm and irony //arXiv preprint arXiv:1809.09795. – 2018.
7. Ke G. et al. Lightgbm: A highly efficient gradient boosting decision tree //Advances in neural information processing systems. – 2017. – T. 30.
8. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval //Journal of documentation. – 1972. – T. 28. – №. 1. – С. 11-21.
9. Rogers A. et al. RuSentiment: An enriched sentiment analysis dataset for social media in Russian //Proceedings of the 27th international conference on computational linguistics. – 2018. – С. 755-763.
10. Sidorov N., Slastnikov S. Some Features of Sentiment Analysis for Russian Language Posts and Comments from Social Networks //Journal of Physics: Conference Series. – IOP Publishing, 2021. – T. 1740. – №. 1. – С. 012036.
11. Blinov V., Bolotova-Baranova V., Braslavski P. Large dataset and language model fun-tuning for humor recognition //Proceedings of the 57th annual meeting of the association for computational linguistics. – 2019. – С.4027-4032.
12. Пикабу. [Электронный ресурс]. – URL.: <https://pikabu.ru> (дата обращения 29.10.2022).

13. Лидеры и аутсайдеры авторынга РФ в 2021 году [Электронный ресурс]. – URL: <https://www.autostat.ru/infographics/50415/> (дата обращения 23.10.2022).
14. Идеальная длина публикаций для соцсетей [Электронный ресурс]. – URL: <https://rusability.ru/pfanshtil/idealnaya-dlina-publikatsii-dlya-sotssetei/5fd295792dda593c3483df32> (дата обращения 24.10.2022).
15. KPI в SMM: как оценить эффективность продвижения в социальных сетях? [Электронный ресурс]. – URL: <https://vk.com/@smmmad-kpi-v-smm-kak-ocenit-effektivnost-prodvizheniya-v-socialnyh> (дата обращения 24.10.2022).
16. Система мониторинга и анализа социальных медиа и СМИ [Электронный ресурс]. – URL: <https://br-analytics.ru/> (дата обращения 28.10.2022).
17. Медиалогия: мониторинг СМИ и соцсетей, инструмент оценки эффективности коммуникаций [Электронный ресурс]. – URL: <https://www.mlg.ru/> (дата обращения 28.10.2022).
18. Kostenetskiy P.S., Chulkevich R.A., Kozyrev V.I. HPC Resources of the Higher School of Economics // Journal of Physics: Conference Series. – 2021. – Т. 1740. – №1. – С. 012050.

Сластников Сергей Александрович. Национальный исследовательский университет Высшая школа экономики (НИУ ВШЭ), Москва, Россия. Доцент, кандидат технических наук. Область научных интересов: технологии искусственного интеллекта, обработка и анализ данных. E-mail: sslastnikov@hse.ru

Жукова Лилия Фанлевна. НИУ ВШЭ, Москва, Россия. Доцент, кандидат физико-математических наук. Область научных интересов: высокоэффективные параллельные вычисления, анализ данных. E-mail: lfzhukova@hse.ru

Семичаснов Илья Владимирович. НИУ ВШЭ, Москва, Россия. Директор. Область научных интересов: разработка игр, технологии искусственного интеллекта. E-mail: isemichasnov@hse.ru

Application for Data Retrieval, Analysis, and Forecasting in Social Networks

S. A. Slastnikov, L. F. Zhukova, I. V. Semichasnov

National Research University Higher School of Economics, Moscow, Russia

Abstract. In this article, we present a web service designed for searching, extracting, and analyzing data from social networks and messengers, demonstrating its application for studying communities within the "VKontakte" social network. The web service enables the identification of typical user profiles within communities, the assessment of emotional sentiment in posts and comments, as well as the forecasting of community development trends. The described web service boasts extensive functional capabilities and an original neural network model for classifying texts of varying lengths based on emotional sentiment. Examples of the tool's usage are showcased in the analysis of the development of car brand communities. The analysis encompasses millions of subscriber audiences, tens of thousands of posts, and hundreds of thousands of comments, thereby affirming the relevance of the samples and the credibility of the results.

Keywords: social network, communities, data analysis, neural network, profile, publication, emotional sentiment, tone, forecasting.

DOI 10.14357/20718632240110

EDN YMIDVN

References

1. Smetanin S. Pulse of the Nation: Observable Subjective Well-Being in Russia Inferred from Social Network Odnoklassniki. *Mathematics*. 2022;10(16):2497. Available from: <https://doi.org/10.3390/math10162947> [Accessed 19 October 2022].
2. Kovaleva Yu.V., Zhuravlev A.L. Social Mood and Subjectivity of Networks Community during the Pandemic COVID-19: Using the Example of Social Network Twitter. Federal State Financed Establishment of Science Institute of psychology. 2020;2(18):151-188 (In Russ.). doi:10.38098/ipran.sep.2020.18.2.005
3. Dokuka S.V., Valeeva D.R. Statistical Models for Analysis of Social Network Dynamics in Educational Studies. *Voprosy Obrazovaniya*. 2015;1:201-2013. (In Russ.) doi: 10.17323/1814-9545-2015-1-201-213

4. Social networks as a tool for studying the psychological portrait of the consumer. Available from: <https://econet.ru/articles/69899-sotsialnye-seti-kak-instrument-izucheniya-psihologicheskogo-portreta-potrebitelya>. [Accessed 19 October 2022].
5. Personality analysis on social networks as an effective method of recruitment. Available from: <https://www.4dk.ru/news/d/20181005135017-analiz-lichnosti-po-sotsialnym-setyam-kak-effektivnyy-metod-podbor-kadrov>. [Accessed 19 October 2022].
6. Ilić S. et al. Deep contextualized word representations for detecting sarcasm and irony/ arXiv preprint arXiv:1809.09795.2018.
7. Ke G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*. 2017; 30. doi: 10.5555/3294996.3295074.
8. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*. 1972;28(1):11-21. doi: 10.1108/eb026526.
9. Rogers A. et al. RuSentiment: An enriched sentiment analysis dataset for social media in Russian. *Proceedings of the 27th international conference on computational linguistics*. – 2018. – C. 755-763.
10. Sidorov N., Slastnikov S. Some Features of Sentiment Analysis for Russian Language Posts and Comments from Social Networks. *Journal of Physics: Conference Series*. 2021;1740(1): 012036. doi: 10.1088/1742-6596/1740/1/012036.
11. Blinov V., Bolotova-Baranova V., Braslavski P. Large dataset and language model fine-tuning for humor recognition. *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019;4027-4032. doi: 10.18653/v1/P19-1394.
12. Pikabu. Available from: <https://pikabu.ru>. [Accessed 29 October 2022].
13. Leaders and outsiders of the Russian car market in 2021. Available from: <https://www.autostat.ru/infographics/50415/>. [Accessed 23 October 2022].
14. The ideal length of publications for social networks. Available from: <https://rusability.ru/pfanshtil/idealnaya-dlina-publikatsii-dlya-sotssetei/5fd295792dda593c3483df3>. [Accessed 24 October 2022].
15. KPI in SMM: how to evaluate the effectiveness of promotion in social networks? Available from: <https://vk.com/@smmmad-kpi-v-smm-kak-ocenit-effektivnost-prodvizheniya-v-socialnyh>. [Accessed 24 October 2022].
16. Social media and Media monitoring and Analysis system. Available from: <https://br-analytics.ru/>. [Accessed 28 October 2022].
17. Medialogy: monitoring of mass media and social networks, a tool for evaluating the effectiveness of communications. Available from: <https://www.mlg.ru>. [Accessed 28 October 2022].
18. Kostenetskiy P.S., Chulkevich R.A., Kozyrev V.I. HPC Resources of the Higher School of Economics. *Journal of Physics: Conference Series*. 2021;1740(1): 012050. doi: 10.1088/1742-6596/1740/1/012050.

Slastnikov Sergey A. National Research University Higher School of Economics, Moscow, Russia. Associate Professor, Candidate of Technical Sciences. Research interests: artificial intelligence technologies, data processing and analysis. E-mail: sslastnikov@hse.ru

Zhukova Liliia F. National Research University Higher School of Economics, Moscow, Russia. Associate Professor, Candidate of Physical and Mathematical Sciences. Research interests: highly efficient parallel computing, data analysis. E-mail: lfzhukova@hse.ru

Semichasnov Ilya V. National Research University Higher School of Economics, Moscow, Russia. Center Director. Research interests: game development, artificial intelligence technologies. E-mail: isemichasnov@hse.ru