ИССЛЕДОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И СБОРА ДАННЫХ ДЛЯ АНАЛИЗА ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ

ПЫЛАЕВ Кирилл Дмитриевич
студент
Московский технический университет связи и информатики
г. Москва, Россия

В статье анализируется методы сбора данных пользователей социальных сетей. Обозреваются решения по сбору данных для анализа поведения пользователей в социальных сетях, исследуется адаптация решений для зарубежных социальных сетей к отечественным аналогам. Вносятся новые предложения для корректировки анализа данных на основе актуальных обновлений социальных сетей. Подчеркивается важность применения современных технологий для повышения точности результата.
Ключевые слова: социальные сети, датасет, анализ данных, методы сбора данных, машинное обучение.

Социальные сети являются одними из самых популярных сервисов в интернете. Они используются для общения, знакомств, чтения новостей. Пользователям данных ресурсов предоставляется возможность размещать разнообразный контент. Мы привыкли видеть это в формате публикаций в сообществах, блога на личных страницах.

Информация может предоставляться в виде текста, изображений, аудио и видео. В виду популярности подобных сервисов мы имеем доступ к огромному количеству открытых данных. Потенциально, многие из этих данных можно использовать для интеллектуального анализа. Подобные исследования социальных сетей приобретают всё большую актуальность в связи с возрастанием количества открытой информации и обостряющейся необходимостью обеспечения безопасности населения и мониторинга общественных настроений.

С динамическим развитием систем рекламы и персонального подбора новостей, разработчики и аналитики стали чаще пользоваться открытыми данными пользователями. Анализ данных пользователей может быть полезен для добавления нового функционала, улучшения качества сервисов (развитие алгоритмов персонального подбора контента и рекламы), определение мошенников и фейковых пользователей.

В данной работе будут проанализированы методы сбора открытых данных пользователей социальных сетей для дальнейшего анализа поведения пользователей.

Методы сбора и предварительной обработки данных играют важную роль в обеспечении качества исследования. Сбор данных может осуществляться с помощью автоматизированных скриптов или специализированных инструментов для работы с API социальных сетей. После сбора данных необходимо провести их предварительную обработку, которая может включать очистку от шума, обработку пропущенных значений, нормализацию и стандартизацию. Важным аспектом является также структурирование данных в формат, удобный для последующего анализа, например, в виде таблиц или графовых структур [9].

При сборе данных необходимо учесть несколько факторов:

– Этические аспекты и вопросы конфиденциальности при работе с пользовательскими данными. Исследователи должны строго соблюдать законодательство о защите персональных данных, в России это закон «О персональных данных» № 152-ФЗ. Это включает в себя получение информированного согласия пользователей, если это применимо, анонимность данных для защиты личной информации, а также обеспечение безопасного хранения и обработки данных.

– Важно учитывать потенциальные преду-

беждения и ограничения в данных. Например, данные могут не в полной мере представлять все демографические группы или могут быть искажены из-за особенностей алгоритмов социальных сетей [11]. Исследователи должны критически оценивать репрезентативность своих данных и учитывать возможные искажения при интерпретации результатов.

– При работе с большими объемами данных также необходимо учитывать технические аспекты, такие как масштабируемость методов обработки и анализа. Это может потребовать использования распределенных вычислительных систем или облачных технологий для эффективной обработки данных.

– Применение методов машинного обучения для анализа поведения пользователей социальных сетей открывает широкие возможности для понимания и прогнозирования пользовательских действий. Этот подход включает в себя несколько ключевых направлений, каждое из которых вносит существенный вклад в общую картину анализа [1].

Алгоритмы кластеризации играют важную роль в выявлении групп пользователей со схожим поведением. Наиболее распространенными методами являются K-means, иерархическая кластеризация и DBSCAN [7]. Такая сегментация помогает маркетологам и разработчикам контента более точно таргетировать свои сообщения и функции платформы.

Методы классификации широко применяются для прогнозирования поведения пользователей. Наиболее эффективными показали себя алгоритмы Random Forest, Gradient Boosting и Support Vector Machines (SVM) [5].

На основе изучения открытых данных иностранных социальных сетей, можно сделать, что для анализа поведения пользователей выбирают посты схожие по определенной тематике (политика, реклама и т. д.) и производят отбор текстовых данных из комментариев пользователей с дальнейшим использованием методов обработки естественного языка. Использование методов обработки естественного языка (NLP) стало неотъемлемой частью анализа текстового контента в социальных сетях [6]. Популярные методы включают анализ тональности (sentiment analysis), извлечение ключевых тем (topic modeling) и классификацию текста. В сети можно встретить множество открытых данных пользователей с определением тональности текста. Пример продемонстрирован на рисунке 1.

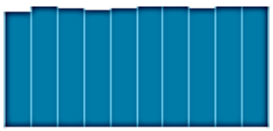
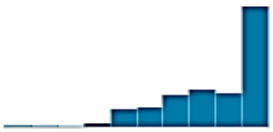
# Unnamed: 0 Index	Text Post by the user	Sentiment Referred sentiment is the post	Timestamp Timestamp of posting
 0 736	707 unique values	Positive 6% Joy 6% Other (646) 88%	 2010-05-15 2023-10-22
0	Enjoying a beautiful day at the park!	Positive	2023-01-15 12:30:00
1	Traffic was terrible this morning.	Negative	2023-01-15 08:45:00
2	Just finished an amazing workout! 💪	Positive	2023-01-15 15:45:00
3	Excited about the upcoming weekend getaway!	Positive	2023-01-15 18:20:00
4	Trying out a new recipe for dinner tonight.	Neutral	2023-01-15 19:55:00
5	Feeling grateful for the little things in life.	Positive	2023-01-16 09:10:00
6	Rainy days call for cozy blankets and hot cocoa.	Positive	2023-01-16 14:45:00
7	The new movie release is a must-watch!	Positive	2023-01-16 19:30:00
8	Political discussions heating up on the timeline.	Negative	2023-01-17 08:00:00
9	Missing summer vibes and beach days.	Neutral	2023-01-17 12:20:00

Рисунок 1. **Пример открытого датасета реакций пользователей с тональностью комментариев (сервис Kaggle)**

Анализ тональности текста является одной из востребованных NLP-задач. Однако, несмотря на ее частотное появление в исследованиях, она имеет достаточно много преломлений: например, алгоритм может «захватывать» и оценивать эмоциональную окраску текста пользователя, а может «склоняться» к определению, был ли комментарий положительный по отношению к первоначальному контексту. Кроме того, тональность, содержащуюся в тексте, можно анализировать на уровне бинарной классификации (основа – выделение полярных пар, к примеру, «негативный-позитивный»), или детализировать на несколько классов (самая распространенная стратегия выделить три класса – «негативный», «нейтральный» и «позитивный», также часто используют пяти-

балльную шкалу – очень положительный, положительный, нейтральный, отрицательный и крайне отрицательный текст) [8].

Нейронные сети, демонстрируют впечатляющие результаты в моделировании сложных поведенческих паттернов пользователей. Рекуррентные нейронные сети (RNN) и их вариации, такие как LSTM (Long Short-Term Memory), особенно эффективны для анализа последовательных данных, таких как история просмотров или покупок пользователя [8].

Другой важный аспект – это динамическая природа социальных сетей. Поведение пользователей может быстро меняться под влиянием внешних факторов, таких как глобальные события или изменения в алгоритмах платформ. Для увеличения точности можно производить обучение на основе глобального набора данных на основе набора популярных социальных сетей. Таким образом можно учесть другие виды онлайн деятельности пользователей. Для структуризация подобного набора данных применяется сетевая база знаний (NKB). Данная модель содержит в себе многоуровневую сеть с локальной базой знаний, представленной в виде ориентированного графа, где вершины представляют собой пользователей, а ребра, их отношения. Т. е. данный метод так же основана на определении реакции пользователей на посты и определении типа личности пользователя, но сбор данных не ограничен единым источником информации [9].

Анализ результатов и практическое применение методов машинного обучения в контексте изучения поведения пользователей социальных сетей представляют собой заключительный и, пожалуй, наиболее важный этап исследования. Этот процесс включает в себя оценку эффективности различных методов, интерпретацию полученных результатов и их практическое применение в различных сферах.

Оценка эффективности методов машинного обучения является критическим шагом для понимания их реальной ценности. Для этого используются различные метрики, такие как точность, полнота, F1-мера для задач классификации, или среднеквадратичная ошибка для задач регрессии [4]. Важно также проводить валидацию и тестирование на независимых данных для обеспечения надежности результатов. При этом необходимо учитывать не только статистические показатели, но и вычислительную сложность алгоритмов, их масштабируемость и способность работать с большими объемами данных в режиме реального времени. Важно не только выявить паттерны в данных, но и объяснить их значение в контексте поведения пользователей. Например, кластеризация пользователей может выявить группы с различными интересами или моделями поведения, что может быть использовано для персонализации контента. Анализ временных рядов активности пользователей может показать циклические паттерны или тренды, которые могут быть связаны с внешними факторами или изменениями в алгоритмах социальных платформ.

Одной из самых популярных сетей в странах СНГ является «ВКонтакте» [3]. У данного приложения так же есть официальный VK API, который позволяет собирать открытые данные пользователей, в том числе их реакции на тематические посты, не прибегая к усложненным алгоритмом с HTML парсингом. Данное API включает в себя множество функций:

1. Предоставление информации о группах (описание, количество участников и т. д.).

2. Получение статистических данных сообщества, что позволяет извлечь аудиторию и эффективность контента.

3. Получение данных об активности пользователей.

С помощью данного интерфейса можно провести анализ поведения пользователей данной социальной сети, собирая открытые данные о реакциях на тематические посты, новости и прочее [2].

Но несмотря на популярность сервиса «ВКонтакте» в странах СНГ, относительно аналогов, не так много сформированных датасетов, созданных для дальнейшего анализа пользователей.

Для упрощения классификации пользователей можно воспользоваться партнерским сервисом SegmentoTarget. Для дальнейшего анализа необходимо собрать информацию о реакции пользователей на определенные посты [10]. К списку реакций относятся:

– информация из текстового комментария;

– одобрение записи по маркеру «мне нравится»;

– добавление записи в личные закладки;

– поделиться записью с друзьями;

– игнорирование записи.

Сервис «ВКонтакте» имеет активную поддержу и регулярные обновления, в следствии чего можно сделать вывод что большинство собранных данных и исследований потеряли свою актуальность и эффективность. Например, отметки «мне нравится» были изменены на реакции, в следствии чего принимать эту метку исключительно как положительную реакцию больше нельзя. Так же из-за обновленной системы рекомендательного подбора сложно определить личную вовлеченность пользователей в предлагаемом контенте. Данные стоит учитывать для дальнейшей работы по сбору датасета.

Для подтверждения гипотезы, был создан скрипт для парсинга данных с помощью VK API. Для тестирования было выбрано новостное сообщество: «Новости первого канала». Были взяты и оформлены в EXCEL-таблицу данные по актуальным постам сообщества. Результаты сформированной таблицы представлены на рисунке 2.

	ID_записи	ID владельца	Количество реакций	Положительные	Отрицательные	Количество комментариев	Количество репостов	Дата публикации	Текст поста
0	1054872	-49388814	1316	1296	20	576	286	2024-03-18 02:52:53	Прямой эфир Первого канала ⬇
1	1142129	-49388814	1	1	0	1	0	2024-12-26 14:20:11	на идет с 1966 года, и в нем нет грандиозных те
2	1142119	-49388814	22	22	0	1	0	2024-12-26 14:13:40	
3	1142116	-49388814	7	7	0	3	13	2024-12-26 14:11:31	ва VPN возможно, а вот реализовать их сбор —
4	1142114	-49388814	3	3	0	0	0	2024-12-26 14:11:03	рустального на своем автомобиле умышленно
5	1142113	-49388814	2	2	0	0	1	2024-12-26 14:10:22	о сумму свыше миллиона рублей. При задерж
6	1142106	-49388814	14	13	1	1	0	2024-12-26 14:02:02	ас важно вернуть городу и его жителям ощуще
7	1142099	-49388814	19	19	0	0	1	2024-12-26 13:48:12	ого совета: Владимир Путин выступил с кратки
8	1142087	-49388814	44	44	0	3	1	2024-12-26 13:27:15	сте с мамой, бабушкой и братом побывала на р
9	1142075	-49388814	30	30	0	5	1	2024-12-26 12:49:05	рытым каналам обращения наших французски
10	1142059	-49388814	53	53	0	10	2	2024-12-26 11:40:09	оспитале имени Мандрыка их вручил заместит
11	1142042	-49388814	109	109	0	18	13	2024-12-26 11:15:30	Украины.«В результате проведенных мероприя
12	1142039	-49388814	38	36	2	2	12	2024-12-26 11:10:34	омки одного БПЛА упали на территорию проми
13	1142038	-49388814	21	21	0	8	1	2024-12-26 11:09:08	озят потерей единства.«Начало выборной кам
14	1141963	-49388814	46	46	0	18	4	2024-12-25 23:46:06	ль Минпромторга Антон Алиханов.«Елка жела
15	1141962	-49388814	28	27	1	2	1	2024-12-25 23:43:38	а», где десять цифровых героев сразятся за гла
16	1141961	-49388814	87	86	1	11	8	2024-12-25 23:42:49	на Михала.«Мы не готовы. Это совершенно оч
17	1141922	-49388814	166	142	24	41	19	2024-12-25 21:24:38	га слова собеседника, недавно общавшегося с
18	1141916	-49388814	115	93	22	18	21	2024-12-25 20:47:11	мстве, спасатели сразу же приступили к тушен
19	1141876	-49388814	61	60	1	0	0	2024-12-25 15:49:16	ау для помощи выжившим в авиакатастрофе р

Рисунок 2. **Таблица данных о стене сообщества ВКонтакте**

Из полученных данных видно, что последние обновления ВКонтакте добавили более гибкий инструментарий для определения эмоциональной реакции пользователей на определенные посты.

В данной работе были проанализированы методы сбора и анализа данных поведения пользователей в социальных сетях. Так же на основе изученного материала было изучено актуальное решение о возможности исследование данных в отечественных сервисах, на примере сервиса «ВКонтакте».

В заключение, методы сбора данных для анализа поведения пользователей в социальных сетях играют ключевую роль в понимании современных тенденций коммуникации и взаимодействия. Независимо от того, используются ли опросы, аналитика контента или технологии отслеживания действий пользователей, каждый из этих подходов предоставляет уникальные возможности для получения ценной информации. Правильно подобранные методы позволяют не только понять предпочтения и интересы аудитории, но и предсказать ее поведение, что является важным аспектом для бизнес-стратегий и маркетинга. Однако важно помнить о необходимости соблюдения этических норм и защиты личных данных пользователей.

СПИСОК ЛИТЕРАТУРЫ

1. *Браницкий А.А., Дойникова Е.В., Котенко И.В.* Использование нейросетей для прогнозирования подверженности пользователей социальных сетей деструктивным воздействиям // Информационно-управляющие системы. – 2020. – № 1(104). – С. 24-33.

2. *Гасанов И.З., Ликсаков М.В.* Эффективная работа с данными сообществ на примере API Вконтакте // Инновации и инвестиции. – 2023. – № 6. – С. 144-146.

3. *Глинская И.Ю., Воронина Л.А.* Роль социальных сетей в формировании приоритетов молодежи // Рекламный вектор-2020: smart-коммуникации. – 2020. – С. 130-134.

4. *Олисеенко В.Д., Абрамов М.В., Тулупьев А.Л.* Нейронные сети lstm и gru в приложении к задаче многоклассовой классификации текстовых постов пользователей социальных сетей // Вестник ВГУ. Серия: Системный анализ и информационные технологии. – 2021. – № 4. – С. 130-141.
5. *Попова Е.П., Леоненко В.Н.* Прогнозирование реакции пользователей в социальных сетях методами машинного обучения // Научно-технический вестник информационных технологий, механики и оптики. – 2020. – Т. 20. – № 1. – С. 118-124.
6. *Чижик А.В., Мельникова С.А., Захаров В.П.* Социальное картирование на основании анализа тональности комментариев в социальных сетях // International Journal of Open Information Technologies. – 2022. – Т. 10. – № 11. – С. 75-80.
7. *Adnan M.M.J., Hemmje M.L., Kaufmann M.A.* Social Media Mining to Study Social User Group by Visualizing Tweet Clusters using Word2Vec, PCA and K-Means //BIRDS+ WEPIR@ CHIIR. 2021. P. 40-51.
8. *de Oliveira N. R. et al.* Identifying fake news on social networks based on natural language processing: trends and challenges // Information. 2021. T. 12. № 1. P. 38.
9. *Gallo F. R. et al.* Predicting user reactions to Twitter feed content based on personality type and social cues // Future Generation Computer Systems. 2020. T. 110. P. 918-930.
10. *Kalabikhina I. E. et al.* The measurement of demographic temperature using the sentiment analysis of data from the social network VKontakte // Mathematics. – 2021. T. 9. № 9. P. 987.
11. *Kauer T. et al.* The public life of data: Investigating reactions to visualizations on reddit // Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. P. 1-12.

RESEARCH OF MACHINE LEARNING METHODS AND DATA COLLECTION FOR ANALYZING THE BEHAVIOR OF SOCIAL NETWORK USERSC

PYLAEV Kirill Dmitrievich
Student
Moscow Technical University of Communications and Informatics
Moscow, Russia

The article analyzes the methods of collecting data from users of social networks. Data collection solutions for analyzing user behavior on social networks are reviewed, and the adaptation of solutions for foreign social networks to their domestic counterparts is investigated. New proposals are being made to adjust the data analysis based on current social media updates. The importance of using modern technologies to improve the accuracy of the result is emphasized.
Keywords: social networks, dataset, data analysis, data collection methods, machine learning.