DOI 10.37882/2223-2966.2024.7-2.25

РАЗРАБОТКА ПРОГРАММНЫХ СРЕДСТВ СБОРА И АНАЛИЗА ДИНАМИКИ ВРЕМЕННЫХ РЯДОВ НА ОСНОВЕ АКТИВНОСТИ ПОЛЬЗОВАТЕЛЕЙ В МАСС-МЕДИА[1]

DEVELOPMENT OF SOFTWARE TOOLS FOR COLLECTING AND ANALYZING THE DYNAMICS OF TIME SERIES BASED ON USER ACTIVITY IN MASS MEDIA

K. Otradnov
S. Strashnov
V. Kalinin

Summary. This paper is devoted to the study of data collection and analysis of user behavior in social media. The study describes several methods developed using the Python language that allow not only to collect data but also to analyze time series, such as the Hurst method. The focus is on identifying fractal characteristics and the degree of stationarity of distributions of time series parameters.

Keywords: time series dynamics, data analysis, Python, Hurst method, fractal analysis, stationary distributions.

Отраднов Константин Константинович
старший преподаватель, МИРЭА — Российский технологический университет, г. Москва
strashnov_sv@pfur.ru
Страшнов Станислав Викторович
доцент, кандидат технических наук, заведующий кафедрой, Российский университет дружбы народов им. Патриса Лумумбы г. Москва
strashnov_sv@pfur.ru
Калинин Владимир Николаевич
педагог ДО, Российский университет дружбы народов им. Патриса Лумумбы г. Москва
kalinin_vn@pfur.ru

Аннотация. Работа посвящена исследованию процессов сбора и анализа данных о поведении пользователей в социальных медиа. В исследовании описан ряд методов, разработанных с использованием языка Python, позволяющих не только осуществлять сбор данных, но и проводить анализ временных рядов, например, таких, как метод Хёрста. Основное внимание уделяется выявлению фрактальных характеристик и степени стационарности распределений параметров временных рядов.

Ключевые слова: динамика временных рядов, анализ данных, Python, метод Хёрста, фрактальный анализ, стационарные распределения.

Введение

В современном информационном обществе масс-медиа и социальные сети играют ключевую роль в формировании и распространении контента, а также взаимодействии между пользователями. Значительный объем данных, генерируемый пользователями в масс-медиа, представляет собой ценный источник информации о их активности, предпочтениях, тенденциях и социокультурных характеристиках.

Разработка программных средств, способных собирать и анализировать динамику временных рядов на основе активности пользователей в социальных медиа, становится предметом все более широкого интереса как в академической, так и в прикладной областях.

Данный подход позволит извлекать информацию и метаданные, необходимые для понимания поведения пользователей, прогнозирования тенденций и разработки эффективных стратегий в маркетинге, социологии, политике и других областях.

В данном контексте организация сбора данных, их обработка, анализ, визуализация и интерпретация становятся ключевыми этапами разработки программных средств, способных эффективно работать с динамикой временных рядов, основанных на активности пользователей в социальных медиа.

В данной работе будет рассмотрен обзор существующих подходов и методов, а также предложен собственный подход к решению данной проблемы с использованием современных методов сбора, предобработки, хранения данных и прогнозирования временных рядов.

Обзор связанных работ

Статья [1] обращает внимание на важность эффективной характеристики само подобных и регулярных

[1] Работа выполнена при финансовой поддержке Российского научного фонда (РНФ), грант № 23–21–00153 «Анализ и моделирование динамики нестационарных временных рядов фрактальных процессов с реализацией памяти (последействия) и самоорганизацией на основе использования дифференциальных уравнений с дробными производными».

паттернов во временных рядах, обладающих краткосрочной и долгосрочной памятью, в различных областях деятельности в условиях постоянно меняющегося и сложного мирового ландшафта.

Особое внимание уделяется анализу динамики возникновения временных срезов для точного, эффективного и своевременного прогнозирования волатильных состояний экономической среды, которая сама по себе представляет собой сложную систему. Для эффективного управления выбором данных и достижения надежных прогнозов критическое значение имеет характеристика сложности и самоподобия в финансовом принятии решений.

Статья [1] предлагает анализ на основе двух основных подходов. Первый подход включает использование экспоненты Хёрста [2], рассчитываемой методом рескейлед-ранга (R/S) и энтропии вейвлет-преобразования для улучшения точности прогнозирования долгосрочного тренда на финансовых рынках. Второй подход включает применение алгоритмов искусственных нейронных сетей (ANN) — Feed forward back propagation (FFBP), Cascade Forward Back Propagation (CFBP) и алгоритм векторного квантования обучения (LVQ) для прогнозирования целей.

Кроме того, в исследовании [3] авторы отмечают, что временные ряды социальных процессов, которые наблюдаются на практике, обладают фрактальные свойства. Они описывают динамику систем, обладающих памятью и способных к самоорганизации.

Например, [3, 4], анализ зависимости между математическим ожиданием и дисперсией амплитуд изменений в этих временных рядах в зависимости от интервала времени расчета показывает сложные взаимосвязи. Дисперсия, например, изменяется в соответствии с размером «скользящего» окна, следуя закону, аналогичному корню дробной степени, что является существенным отличием от нормального закона распределения.

А в статье [5] представлен систематический обзор пакетов Python с упором на анализ временных рядов. Целью является предоставление обзора различных задач анализа временных рядов и методов предварительной обработки данных, а также обзор характеристик развития пакетов.

Авторы классифицировали пакеты в соответствии с реализованными задачами анализа, методами подготовки данных и средствами оценки полученных результатов (методы и доступ к данным для оценки). Мы также рассмотрели аспекты документации, лицензий, размер сообщества пакетов и используемые зависимости.

Результаты авторов показывают, что прогнозирование является наиболее часто реализуемой задачей, что половина пакетов предоставляют доступ к реальным данным или позволяют генерировать синтетические данные, и что многие пакеты зависят от нескольких библиотек (наиболее используемыми являются numpy, scipy и pandas).

Разработка программного обеспечения для сбора данных

Разработка программного обеспечения для сбора и анализа динамики временных рядов на основе активности пользователей в массмедиа возможно с использованием готовых и эффективных инструментов для сбора данных. В связи с чем, основным языком для написания программного обеспечения был выбран Python [6–8].

Один из ключевых компонентов такого программного обеспечения будет парсер, который отвечает за сбор и структурирование данных из социальных сетей.

Преимущества Python для написания веб-скраперов заключается в том, что:

1. Python обладает простым и читаемым синтаксисом, что делает его идеальным для быстрого создания и дальнейшей поддержки веб-скрапера;
2. Python имеет огромное количество библиотек для веб-скрапинга, таких как Beautiful Soup [9] (bs4), lxml [10], requests [11], Selenium [12] и др;
3. Python легко интегрируется с другими инструментами и технологиями, такими как базы данных, фреймворки для анализа данных и визуализации.

Более того, в проектируемой системе каждый этап получения, обработки и сохранения информации может быть организован как отдельный модуль, функционирующий как самостоятельная единица и построенный на основе принципов микро сервисной архитектуры [13].

К примеру, парсер можно реализовать на основе библиотеки:

1. requests, для отправки HTTP-запросов на веб-страницы и получения HTML-кода;
2. BeautifulSoup, для парсинга HTML и извлечения нужной информации из веб-страниц.

А модуль обработки временных рядов, на основе библиотеки:

1. pandas, для работы с данными в формате таблицы, представленными в виде временных рядов.
2. numpy, для выполнения математических операций и работы с многомерными массивами, что может быть полезно при анализе временных рядов.

3. matplotlib или seaborn, для визуализации временных рядов и результатов анализа.
4. statsmodels или scikit-learn, для статистического анализа временных рядов, построения моделей прогнозирования или выполнения других аналитических задач.

Использование баз данных временных рядов

Базы данных временных рядов играют ключевую роль в хранении, управлении и анализе данных, которые меняются во времени.

Одни из самых популярных решений для работы с временными рядами:

1. InfluxDB — является высокопроизводительной, распределенной базой данных, специализированной на хранении временных рядов.
2. TimescaleDB — является расширением для PostgreSQL, предназначенным для работы с временными рядами.
3. Apache Cassandra — распределенная NoSQL база данных, которая хорошо подходит для хранения временных рядов.

Базы данных временных рядов применяют, когда есть упорядоченные по времени данные с временными метками, такие как метрики от инфраструктуры, данные датчиков, различные метаданные и др.

Основные преимущества баз данных временных рядов:

Данные временных рядов всегда собираются на протяжении определенного периода времени;

Данные из рабочих нагрузок являются новыми и записываются как вставки. Существующие данные не обновляются путем замены значений;

Когда данные записываются, они автоматически назначаются последнему интервалу времени.

В нашем исследовании было принято решение выбрать TimescaleDB, так как она является расширением для базы данных PostgreSQL, а также:

1. Обеспечивает возможность использования привычного SQL для работы с данными временных рядов, что упрощает разработку и анализ;
2. Масштабируется вертикально и горизонтально, что позволяет поддерживать большие объемы данных и высокую производительность;
3. Имеет встроенную поддержку различных временных функций и агрегаций, что облегчает аналитику и обработку временных данных;
4. Полностью совместима с экосистемой PostgreSQL, что обеспечивает доступ к большому количеству инструментов и библиотек для анализа данных.

Структура базы данных для временных рядов приведена на рисунке 1.

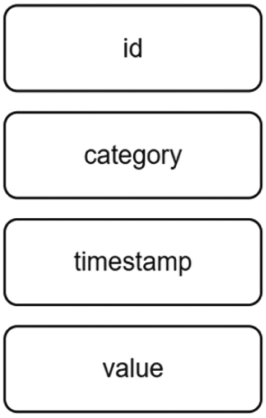

Рис. 1. Поля базы данных временных рядов

База данных временных рядов для каждой записи собирает следующую информацию: категория (уникальный идентификатор временного ряда), время (или шаг интервала записи), значение.

Анализ временных рядов

Полученные данные из социальных сетей можно представить (визуализировать) в виде временных рядов и после провести их анализ. Для примера, с новостного портала «РИА Новости» (https://ria.ru/) были взяты, собраны и обработаны все реакции пользователей за последние 2,5 года (882 дня), рисунок 2.

С данного портала были получены информации о реакциях: «Нравится», «ХаХа», «Удивительно», «Грустно», «Возмутительно», «Не нравится», которые возможно оставить после прочтения новости на портале внизу страницы (рисунок 3)

В качестве первичного анализа динамики временного ряда и определения его особенностей можно использовать метод нормированного размаха Хёрста [2].

Этот метод может быть применен для определения фрактальной размерности D временных рядов и однозначных самоаффинных кривых [14], а сам метод основан на анализе размаха случайной величины и её среднеквадратичного отклонения [15]. Данный метод дает возможность выявить фрактальные характеристики временных рядов и классифицировать их тип поведения.

Алгоритм метода Хёрста приведен ниже (Алгоритм 1). Результаты приведены на рисунке 4.

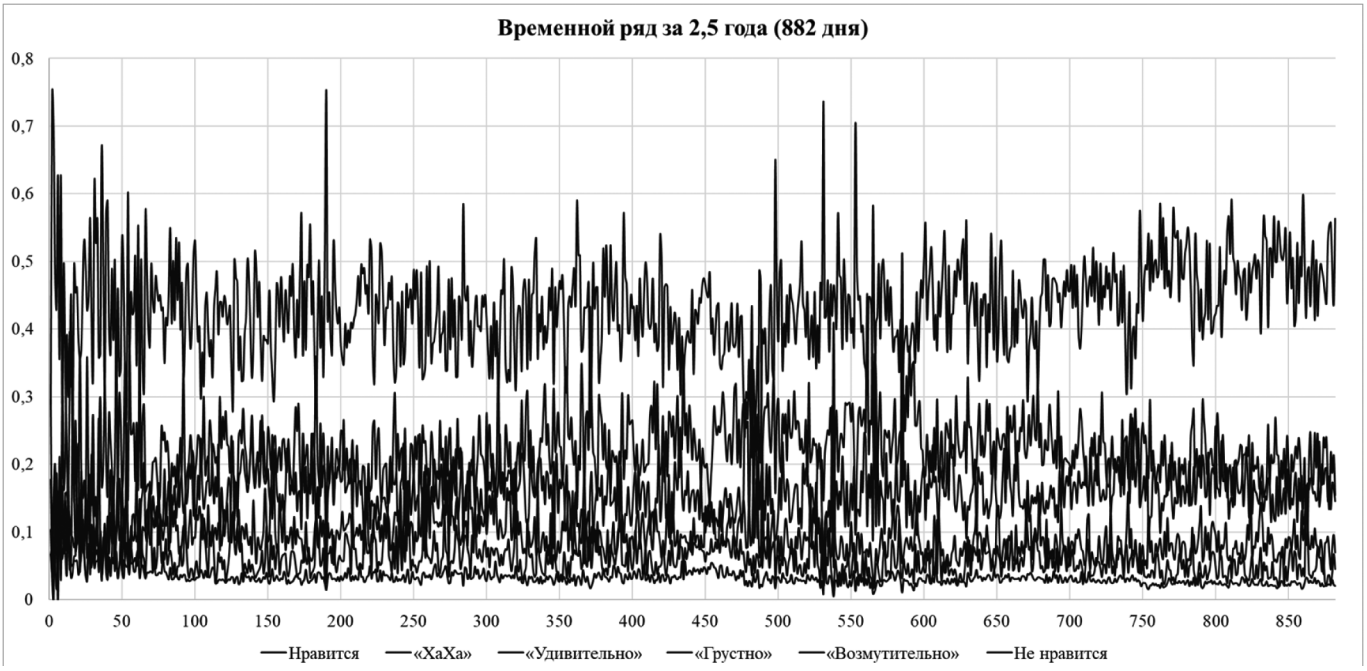

Рис. 2. Временной ряд активности пользователей новостного портала «РИА Новости» за 2,5 года

Рис. 3. Блок реакций под каждой статьей на сайте «РИА Новости»

Алгоритм 1. Алгоритм метода Хёрста (R\S анализ)	
Ввод:	$\xi(\tau)$ — временной ряд, при $\tau \in [1,t]$
1.	Среднее значение — $\langle \xi \rangle$
2.	$\langle \xi \rangle = \frac{1}{t}\sum_{\tau=1}^{t}\xi(\tau)$
3.	$\delta(\tau,t) = \sum_{i=1}^{\tau}\left(\xi(i) - \langle \xi \rangle\right)$
4.	$R(t) = \max_{1\le\tau\le t}\delta(\tau,t) - \min_{1\le\tau\le t}\delta(\tau,t)$
5.	$S(t) = \sqrt{\frac{1}{t}\sum_{i=1}^{\tau}\left(\xi(i) - \langle \xi \rangle\right)^2}$
6.	$\frac{R(t)}{S(t)} \sim t^H$

Наличие изломов в зависимости $\frac{R(t)}{S(t)}$ может свидетельствовать о наличие характерных временных масштабов и/или периодичностей. Величина коэффициента Хёрста позволяет дать классификацию временных рядов по характеру их поведения [16].

Так, на рисунке 4, при расчете зависимости логарифма $\frac{R(t)}{S(t)}$ от логарифма величины выборки уровней временного ряда (τ) можно получить следующие линейные уравнения:

1. для эмоции «нравится» $y = 0{,}26x + 0{,}56$ со значением коэффициента корреляции $R^2 = 0.95$;
2. для эмоции «ХаХа» $y = 0{,}21x + 0{,}77$ со значением коэффициента корреляции $R^2 = 0.97$;
3. для эмоции «Удивительно» $y = 0{,}26x + 0{,}57$ со значением коэффициента корреляции $R^2 = 0.98$;
4. для эмоции «Грустно» $y = 0{,}21x + 0{,}76$ со значением коэффициента корреляции $R^2 = 0.96$;
5. для эмоции «Возмутительно» $y = 0{,}27x + 0{,}55$ со значением коэффициента корреляции $R^2 = 0.98$;
6. для эмоции «Не нравится» $y = 0{,}27x + 0{,}51$ со значением коэффициента корреляции $R^2 = 0.98$;

При $H = 0{,}5$ — временной ряд носит случайный процесс с независимыми приращениями, где события случайны и некоррелировны.

При $0 < H < 0{,}5$ — временной ряд является антиперсистентным, и рост в прошлом означает уменьшение

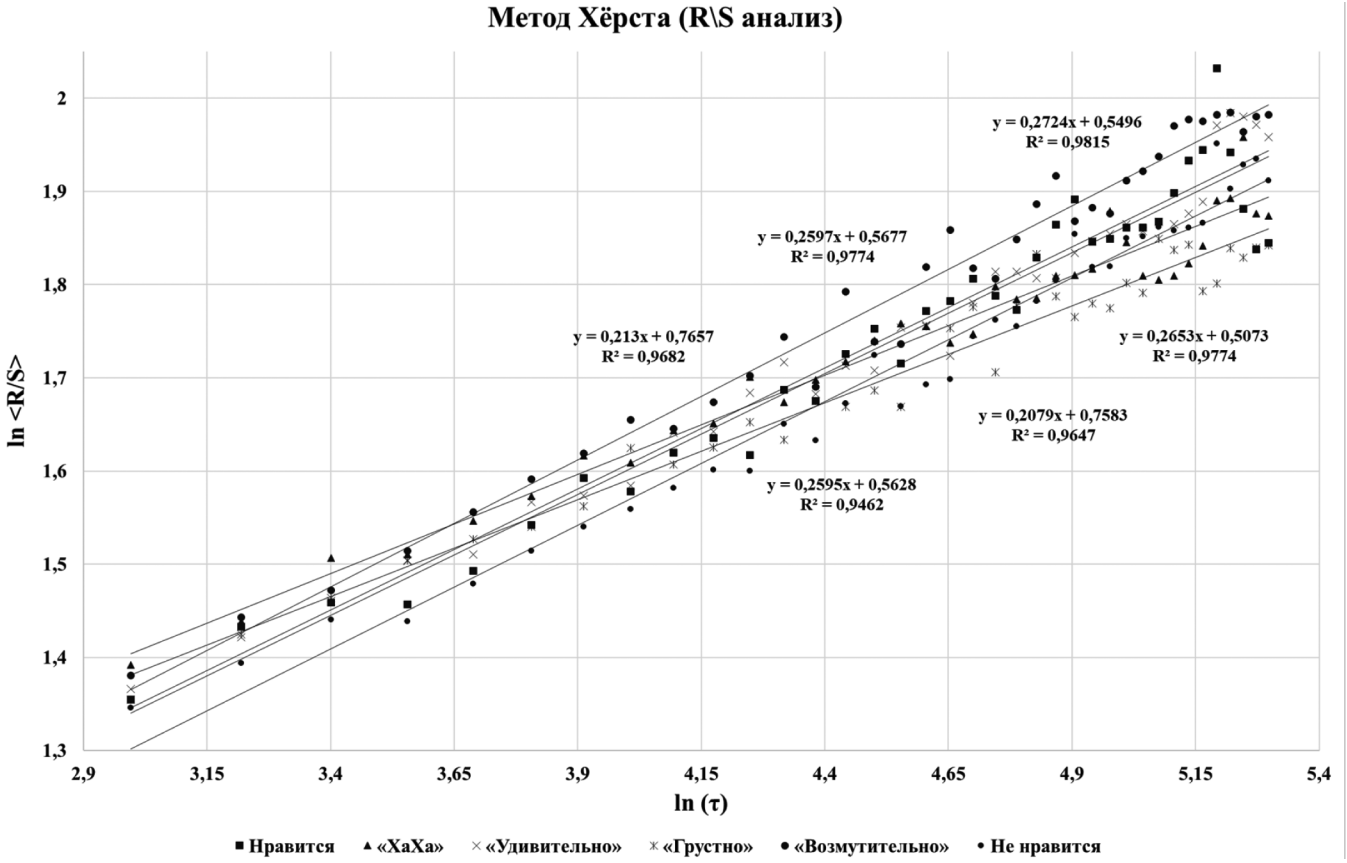

Рис. 4. Определение показателя Хёрста для всех временных рядов эмоций с портала «РИА Новости»

в будущем, и наоборот. То есть, временной ряд склонен в ближайшее время поменять тренд.

При 0,5 < *H* < 1 — временной ряд является персистентным, т.е. трендоустойчивым. Если ряд возрастает, то вероятно, что он сохранит данный тренд еще какое-то время.

Во всех случаях коэффициент Хёрста меньше 0,5 и, следовательно, наблюдаемые временные ряды является антиперсистентным. Поскольку величины коэффициента Хёрста существенно отлична от 0,5, то из этого следует, что структура данных рядов обладает фрактальностью, а описываемые им процессы имеют краткосрочную память [16].

Из эмпирического закона Хёрста следует, что $\frac{R(t)}{S(t)} \sim t^H$, где: *H* — показатель Хёрста, связанный с коэффициентом фрактальности размерностью *D*, связанной соотношением $D = 2 - H$

Для дальнейшей обработки наблюдаемых временных рядов можно определить величины амплитуды изменения эмоционального отношения пользователей сетевых новостных ресурсов, можно использовать алгоритм 2.

Алгоритм 2. Алгоритм определения величины амплитуды изменения характеристик	
Ввод:	$\xi(\tau)$ — временной ряд, при $\tau \in [1,t]$
1.	Define $\omega = 200$ (max значение скользящего)
2.	For w in range (ω)
3.	For each *i* in $\tau \in [1, t-\omega]$
4.	$\xi'(i) = \xi(i+w) - \xi(i)$
5.	$\mu(\tau,\omega) = \frac{\sum_{i=1}^{N}\tau_i}{N}$ математическое ожидание
6.	$\sigma^2(\tau,\omega) = \frac{\sum_{i=1}^{N}\{\tau_i - \mu(\tau_i)\}^2}{N}$ дисперсия
7.	$As(\tau,\omega) = \frac{\sum_{i=1}^{N}\{\tau_i - \mu(\tau)\}^3}{N\{\sqrt{\sigma^2(\tau)}\}^3}$ асимметрия
8.	$Ex(\tau,\omega) = \frac{\sum_{i=1}^{N}\{\tau_i - \mu(\tau)\}^4}{N\{\sigma^2(\tau)\}^2} - 3$ эксцесс

Исследование зависимости математического ожидания, дисперсии, асимметрии и эксцесса амплитуд отклонений активности пользователей позволяет определить, являются ли изучаемые временные ряды стационарными или нестационарными. Результаты приведены на рисунках 5–8.

Помимо анализа активности пользователей с помощью метода Хёрста, можно применить метод исключения тренда (алгоритм 3,4) и метод вычисления индекса фрактальности (алгоритм 5). Метод исключения тренда является еще одним способом, позволяющим рассчитать показатель Хёрста

Алгоритм 3. Метод наименьших квадратов (LSM)	
Ввод:	$\xi(\tau)$ — временной ряд, при $\tau \in [1,N]$
1.	$a = \dfrac{N * \sum_{i=1}^{N}\left(\tau_i * \xi(\tau_i)\right) - \sum_{i=1}^{N}\tau_i * \sum_{i=1}^{N}\xi(\tau_i)}{N * \sum_{i=1}^{N}\left(\tau_i^2\right) - \left(\sum_{i=1}^{N}\tau_i\right)^2}$
2.	$b = \dfrac{\sum_{i=1}^{N}\xi(\tau_i) - a * \sum_{i=1}^{N}\tau_i}{N}$

Алгоритм 4. Алгоритм метода исключения тренда	
Ввод:	$\xi(\tau)$ — временной ряд, при $\tau \in [1,N]$; t — количество интервалов
1.	For $\xi(\tau)_i$-th part in range $\dfrac{N}{\tau}$ do:
2.	Define int a_i and $b_i \leftarrow LSM\left(\xi(\tau)_i\right)$
3.	$F_i^2(t) = \dfrac{1}{t}\sum_{\tau=i-t+1}^{(i+1)t}\left(\xi(\tau) - y_i(\tau)\right)^2$, where $y_i(\tau) = \tau * a_i + b_i$
4.	$F(t) = \sqrt{\dfrac{t}{N}\sum_{i=0}^{\frac{N}{t}-1}F_i^2(t)}$

Строя зависимость log$F$(τ) и og(τ), методом наименьших квадратов определяется наклон аппроксимирующей прямой полученной зависимости, который оценивает значение показателя Хёрста H. Тогда фрактальная размерность временного ряда: $D = 2 - H$.

Кроме того, фрактальный анализ находит практическое применение в изучении взаимосвязей во временных рядах. Если фрактальная размерность составляет

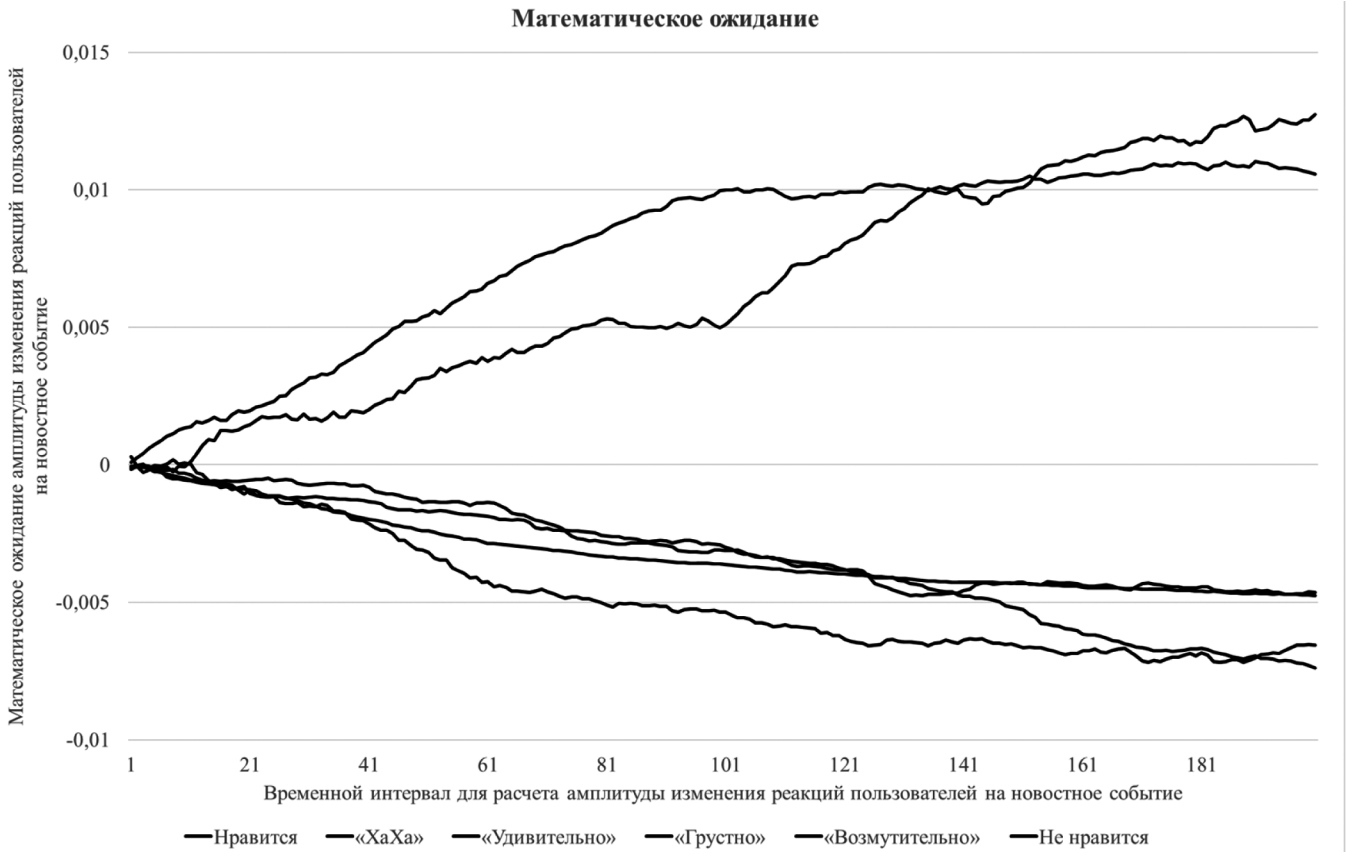

Рис. 5. Зависимость величины математического ожидания амплитуд от временного интервала

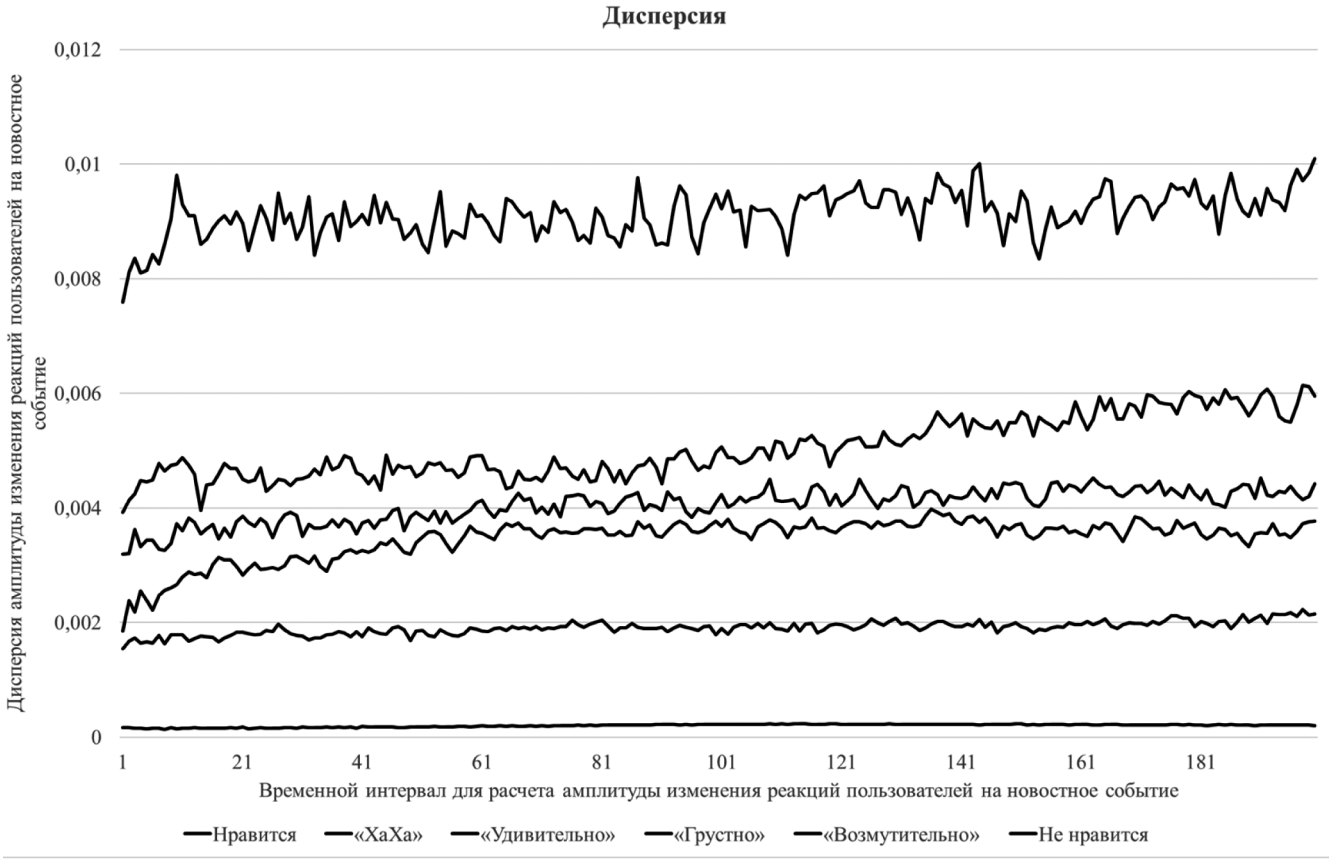

Рис. 6. Зависимость величины дисперсии амплитуд от временного интервала

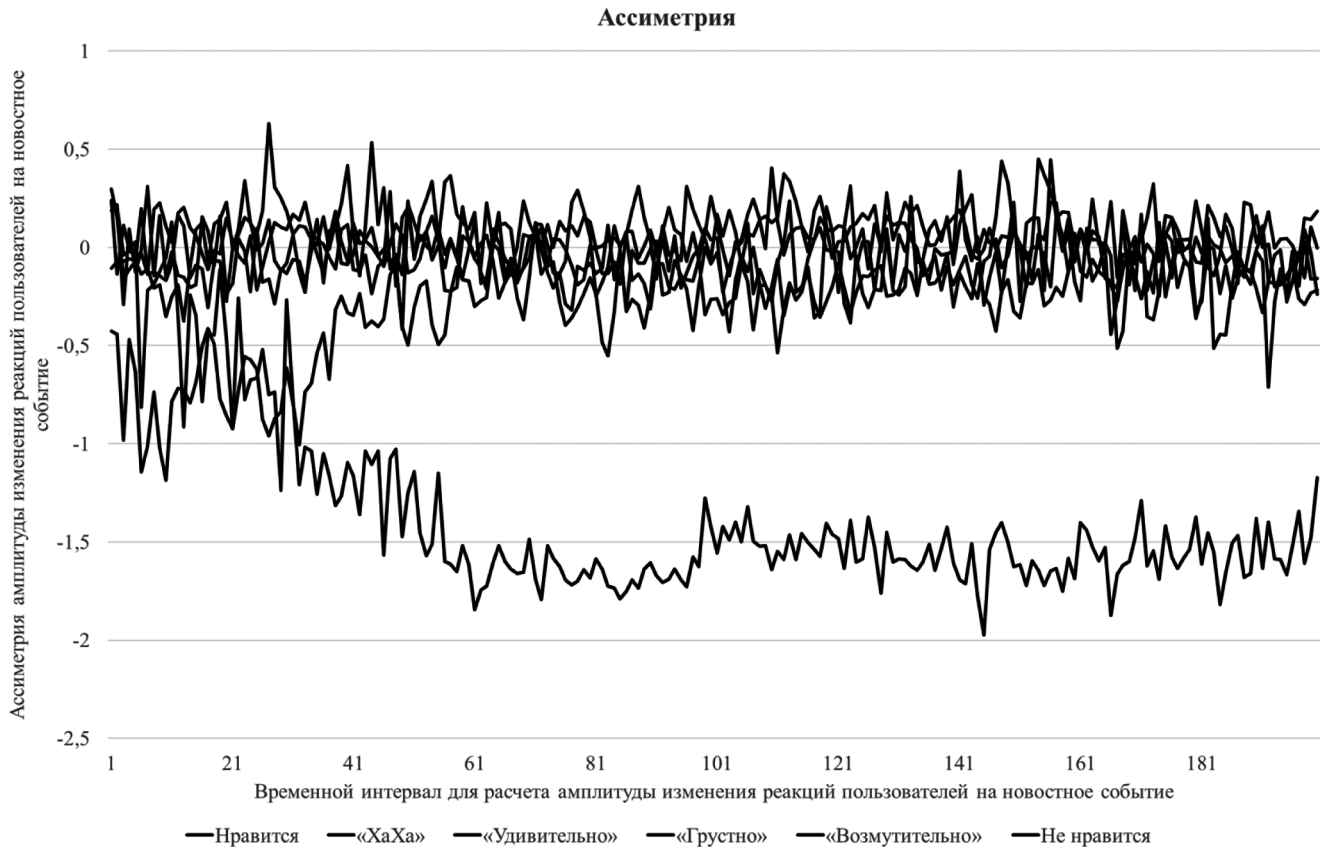

Рис. 7. Зависимость величины ассиметрии амплитуд от временного интервала

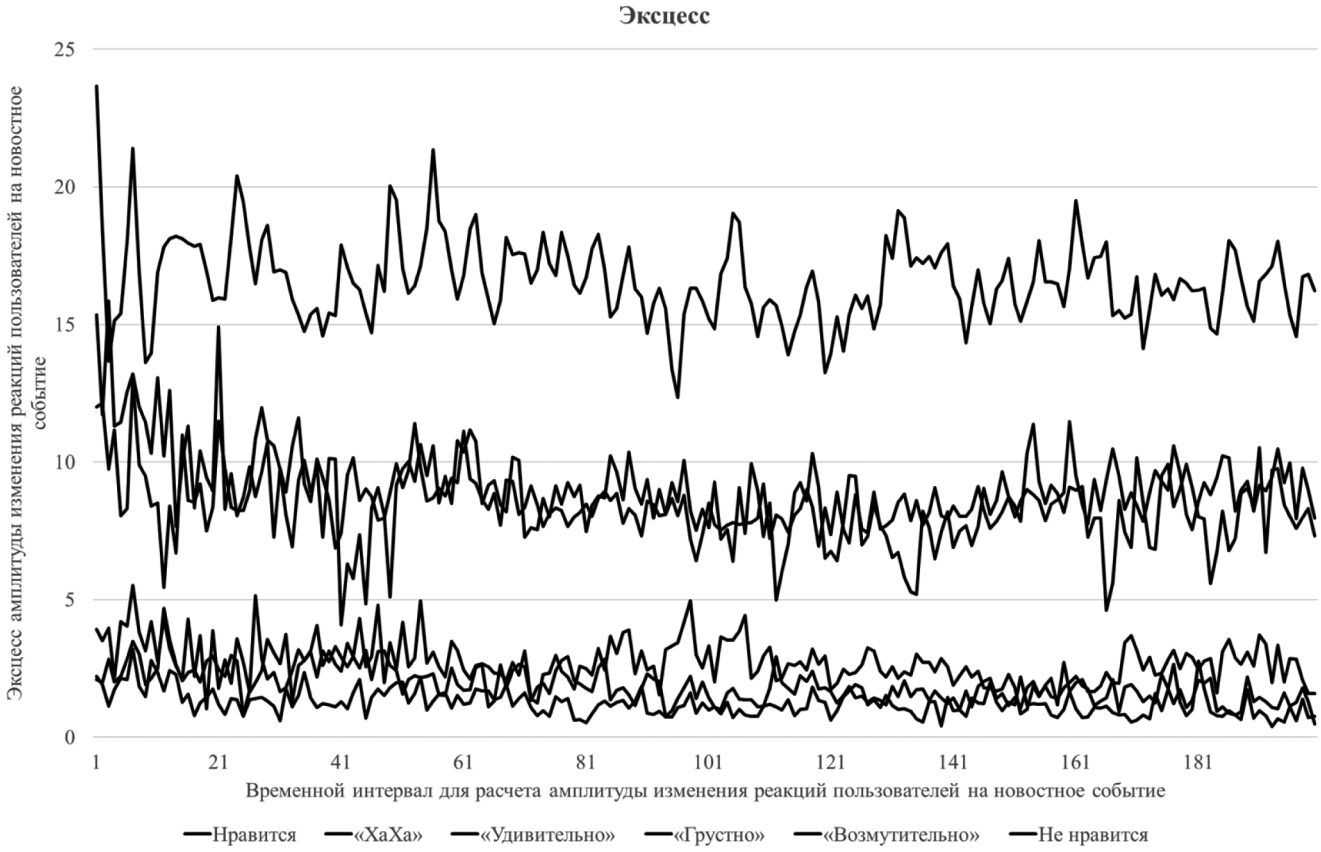

Рис. 8. Зависимость величины эксцесса ожидания амплитуд от временного интервала

1,5, то приращения в ряду являются независимыми. Значение фрактальной размерности меньше 1,5 указывает на персистентный ряд с эффектами «долговременной памяти», а значение больше 1,5 соответствует антиперсистентному поведению временного ряда. [17]

Алгоритм 5. Алгоритм вычисления индекса фрактальности	
Ввод:	$\xi(\tau)$ — временной ряд, при $\tau \in [1, N]$
1.	Define $\omega_m, m = 2^n$
2.	For each ω_m calculate $A_i(\varepsilon)$
3.	$A_i(\varepsilon) = \max_{\tau_{i-1} \le t \le \tau_i} \xi - \min_{\tau_{i-1} \le t \le \tau_i} \xi$
4.	$V_x(\varepsilon) = \sum_{i=1}^{m} A_i(\varepsilon)$
5.	$\text{draw} \log V_x(\varepsilon),\ \log\varepsilon$
6.	$a, b = FSM()$
7.	$\mu = -a, D_\mu = \mu + 1$

Кроме того, можно построить гистограммы распределения величин амплитуд изменения активности пользователей для скользящего окна, равного одному дню (рисунок 9), 10 (рисунок 10), 50 (рисунок 11), 100 (рисунок 12).

Амплитудные распределения демонстрируют резкие пики, высота которых остается практически постоянной независимо от интервала времени анализа. Даже при увеличении длительности интервала времени ширина гистограммы может возрастать, однако высота пиков и их положение относительно нуля остаются практически неизменными. Это поведение типично для стационарных распределений.

Заключение

Выводы по обработке данных.

Обработка и анализ наблюдаемых данных позволяет сделать ряд выводов:

1. Временные ряды, описывающие рассмотренные процессы — являются нестационарными;
2. Анализ наблюдаемых временных рядов показывает, что описываемые ими процессы обладают краткосрочной памятью ($H < 0{,}5$);
3. В распределении амплитуд наблюдается небольшая величина асимметрии и распределение амплитуд является почти симметричным.

В итоге, в представленной работе была приведена архитектура сбора данных временных рядов и их последующий анализ, так, была изучена активность пользова-

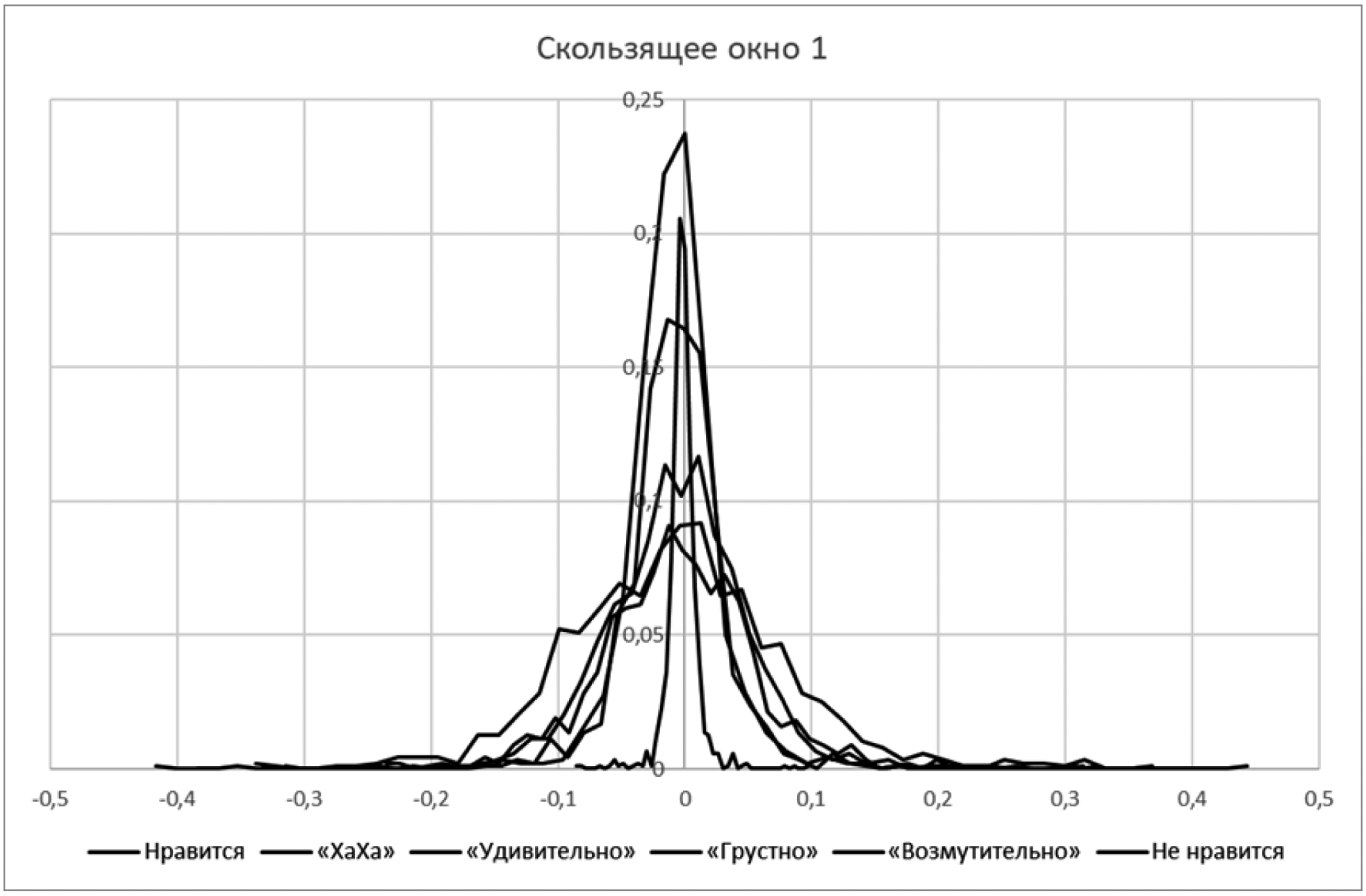

Рис. 9. Гистограммы распределений амплитуд для скользящего окна в 1 день

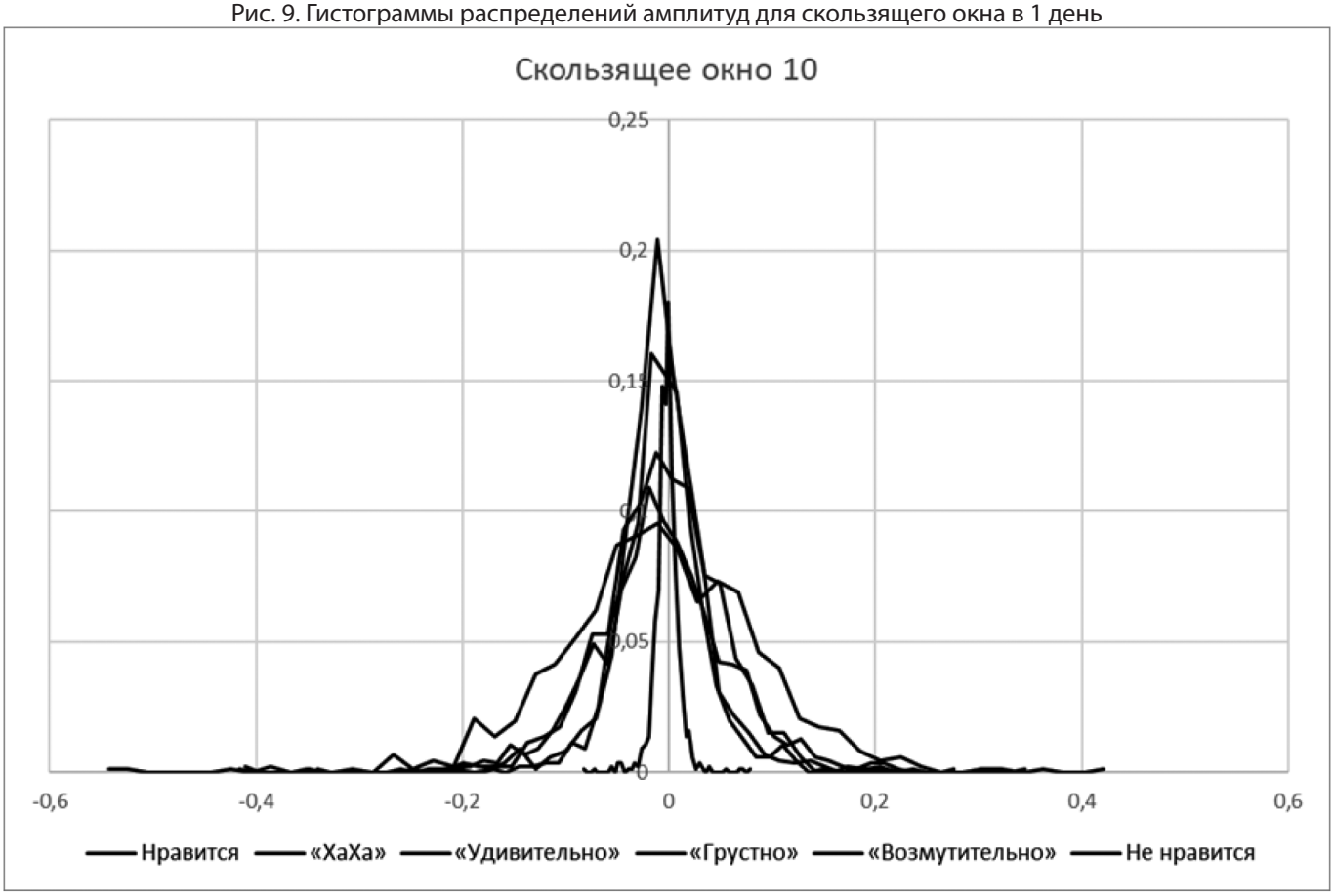

Рис. 10. Гистограммы распределений амплитуд для скользящего окна в 10 дней

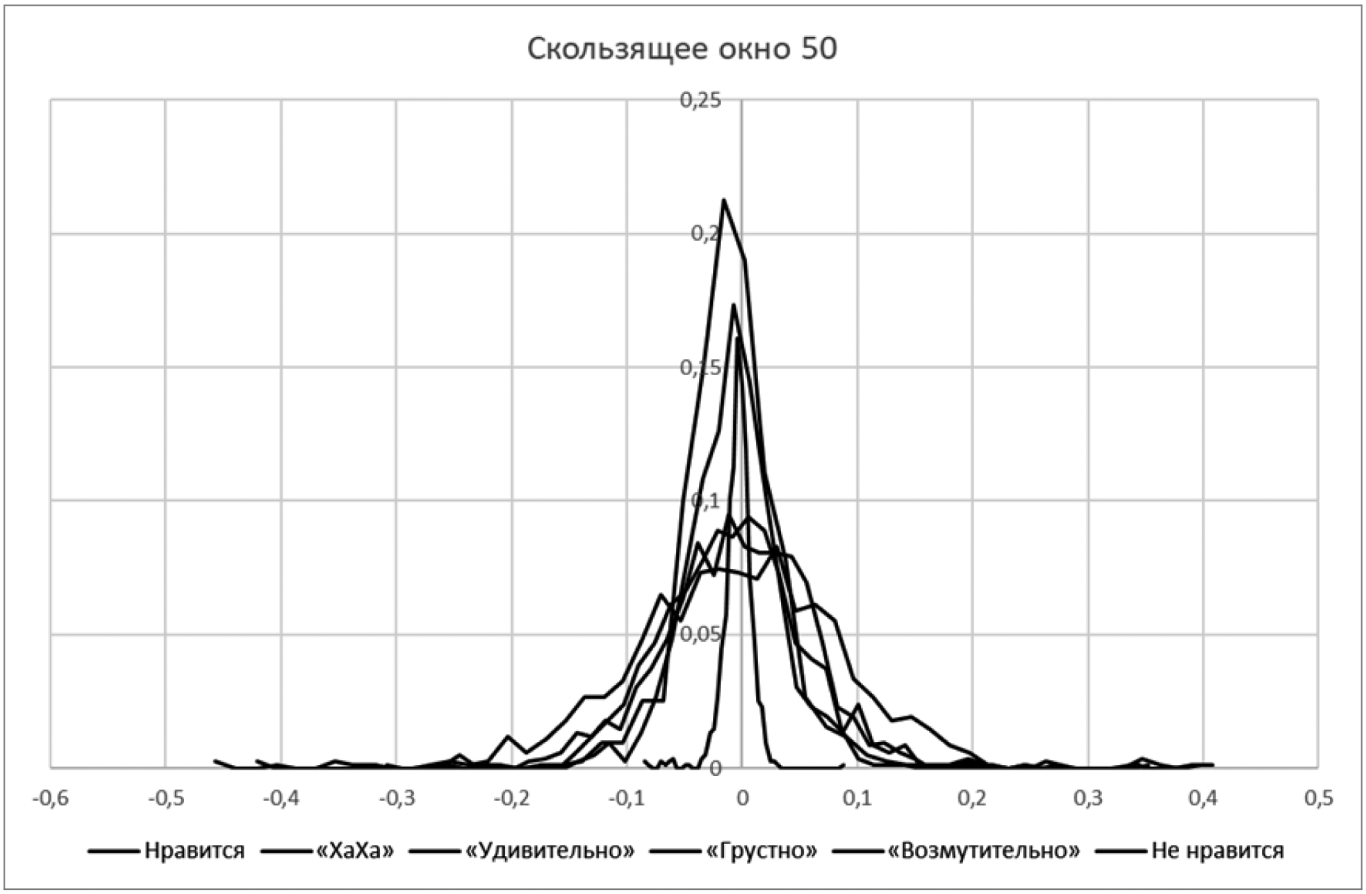

Рис. 11. Гистограммы распределений амплитуд для скользящего окна в 50 дней

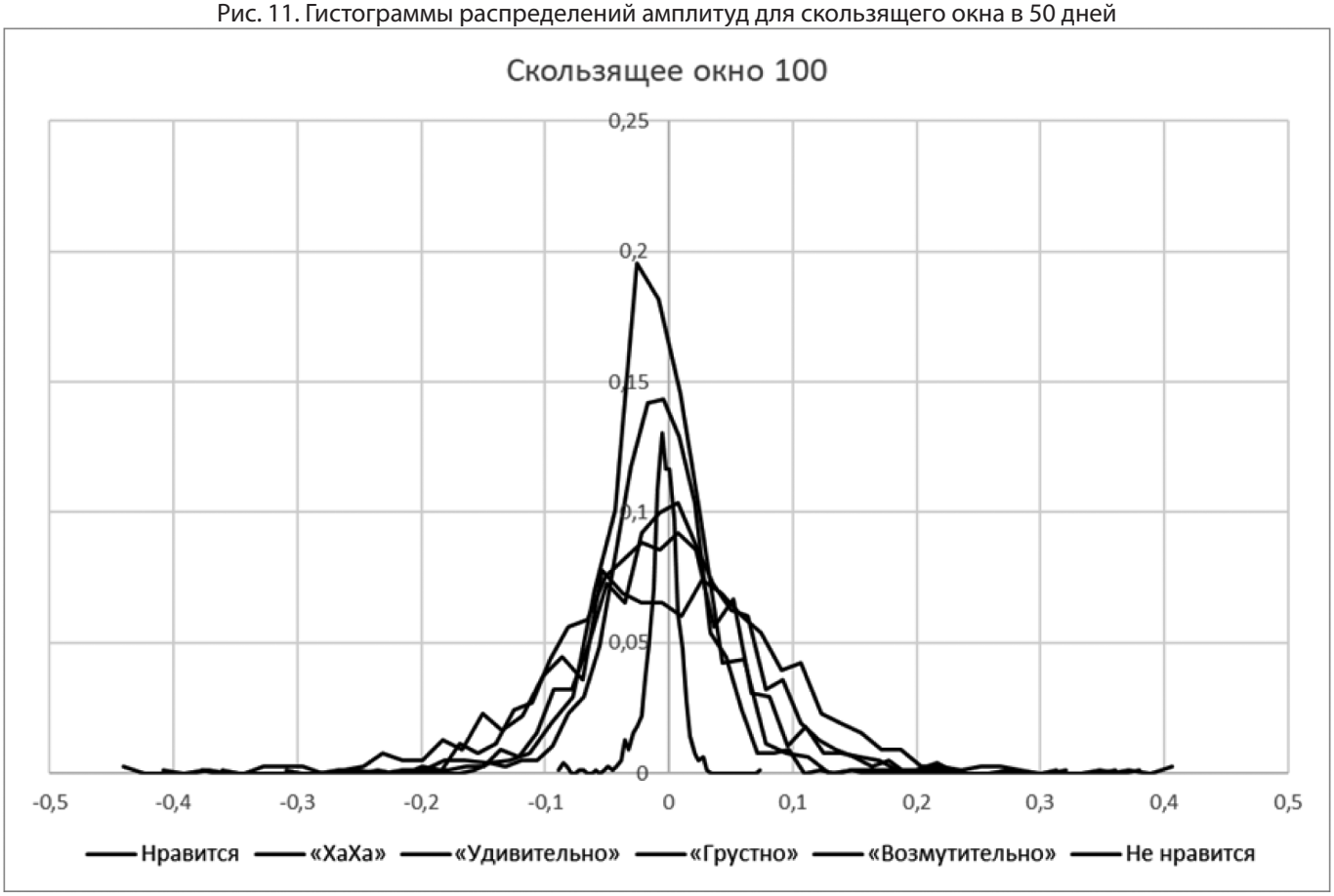

Рис. 12. Гистограммы распределений амплитуд для скользящего окна в 100 дней

телей по комментированию новостей, были построены и обработаны методом нормированного размаха Хёрста временные ряды активностей пользователей.

Исследование методом Хёрста показало, что ряды являются антиперсистентными.

При анализе математического ожидания амплитуд выявлена зависимость от интервала времени расчета этих амплитуд, кроме того, зависимости величин дисперсии от интервала времени имеют сложный нелинейный характер.

ЛИТЕРАТУРА

1. Karaca Y., Baleanu D.A novel R/S fractal analysis and wavelet entropy characterization approach for robust forecasting based on self-similar time series modeling // Fractals. — 2020. — Т. 28. — №. 08. — С. 2040032, https://doi.org/10.1142/S0218348X20400320.
2. H.E. Hurst. Long-term storage capacity of reservoirs. // Transactions of American Society of Civil Engineers. — 1951. — Т. 116. — С. 770.
3. Dmitry Zhukov, Tatiana Khvatova, Leonid Istratov. A stochastic dynamics model for shaping stock indexes using self-organization processes, memory, and oscillations. Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics, ECIAIR 2019, Oxford, UK, 31 October–1 November 2019, pp. 390–401, E-Book ISBN: 978-1-912764-44-0, Book version ISBN: 978-1-912764-45-7.
4. D. Zhukov, T. Khvatova, L. Istratov. Analysis of non-stationary time series based on modelling stochastic dynamics considering self-organization, memory, and oscillations. ITISE 2019 International Conference on Time Series and Forecasting. Proceedings of Papers, 25-27 September 2019, Granada (Spain), Vol. 1, pp. 244–254. ISBN: 978-84-17970-78.
5. Siebert J., Groß J., Schroth C.A systematic review of python packages for time series analysis //arXiv preprint arXiv:2104.07406. — 2021, https://doi.org/10.48550/arXiv.2104.07406.
6. McKinley W. Python and data analysis. Per. with English. Slinkin A.A. M.: DMK Press, 2015. 482 pp. ISBN 978-5-97060-315-4.
7. Hellman D. Python Standard Library 3. Reference book with examples. Dialectics. 2nd ed. 2019. 1375 p.
8. Percival G. Python. Development based on testing. 2018. 624 p. ISBN: 978-5-97060-594-3.
9. BeautifulSoup [Электронный ресурс]. URL: https://www.crummy.com/software/BeautifulSoup/ (дата обращения 25.04.2024).
10. LXML [Электронный ресурс]. URL: https://lxml.de/ (дата обращения 25.04.2024).
11. Requests [Электронный ресурс]. URL: https://requests.readthedocs.io/en/latest/ (дата обращения 25.04.2024).
12. Selenium Web Driver [Электронный ресурс]. URL: https://www.selenium.dev/ (дата обращения 25.04.2024).
13. Mitra R, Nadareishvili I. Microservices. From architecture to release. O'Reilly, 2024. 336 p., 978-5-4461-1884-7
14. Латыпов И.А. Фрактальность рекурсивной сети информационнокоммуникационных отношений // Сб. научных статей «Актуальные тенденции социальных коммуникаций: история и современность». — Ижевск, 2013. — С. 149–152.
15. Латыпов И.А. Полисубъектная мультифрактальность информационных отношений в сети: философские аспекты // Вестник Гуманитарного университета. — Екатеринбург, 2014. № 4 (7). С. 80–87. ISSN 2308–8117.
16. Mandelbrot B.B. The Fractal Geometry of Nature. W.H. Freeman, Sun Francisco, 1982.
17. Мансуров А.К. Прогнозирование валютных кризисов с помощью методов фрактального анализа // Проблемы прогнозирования. 2008, №1 (106). — С. 145–158.