

УДК 004.9

АРХИТЕКТУРА МОДУЛЯ ПРОГНОЗИРОВАНИЯ В ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЕ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

А. М. Кузнецов, аспирант ТГТУ, Тамбов, Россия;
orcid.org/0000-0003-2640-3417, e-mail: 068t@mail.ru

Представлена архитектура модуля прогнозирования в информационно-аналитической системе (ИАС) социологических исследований. Рассмотрена интеграция современных подходов и методов предиктивной аналитики для анализа и прогнозирования социальных процессов, включая линейную и логистическую регрессии, модели временных рядов (ARIMA, SARIMA, SARIMAX), методы машинного обучения и нейронные сети (LSTM, BERT). Модуль обеспечивает хорошее качество прогнозов, гибкость и адаптивность анализа с учетом сложных зависимостей в данных.

Ключевые слова: информационно-аналитическая система (ИАС), социологические исследования, модуль, прогнозирование, нейронные сети, LSTM, BERT, машинное обучение, временные ряды, регрессионный анализ, SARIMAX.

DOI: 10.21667/1995-4565-2024-90-77-90

Введение

Социологические исследования, основанные на данных опросов, требуют эффективных методов для предсказания социальных изменений и тенденций. Предиктивная аналитика (Predictive Analytics) позволяет прогнозировать события на основе анализа исторических данных с использованием математических моделей. Ключевые этапы предиктивной аналитики включают сбор данных, исследовательский анализ и предиктивное моделирование, что требует высокоточной информации и больших объемов данных [1].

В рамках предиктивного анализа применяются различные модели. Линейная регрессия используется для выявления линейных зависимостей между переменными, в то время как логистическая регрессия прогнозирует бинарные или категориальные исходы. Анализ временных рядов, включающий модели ARIMA и методы экспоненциального сглаживания, используется для прогнозирования на основе временных данных [2].

Вычислительная социология представляет собой новую область количественной социологии, которая описывает, объясняет и предсказывает социальные явления и процессы с использованием сложных моделей и инструментов социальных вычислений [3]. В последние десятилетия быстрый рост информации и данных создал беспрецедентные возможности для социальных наук. Машинное обучение предоставляет социологам новые инструменты для анализа и прогнозирования, но требует внимательного подхода и интеграции с традиционными социологическими методами, обеспечивая тем самым всесторонний и обоснованный подход к исследованию социальных явлений [4].

Санкции, введенные против России рядом стран, ограничили доступ ко многим западным облачным сервисам, платформам машинного обучения и аналитики данных (например, Google Cloud, AWS, Azure), а также к программным продуктам, таким как SPSS, SAS и другие [5]. Использование же языков программирования (например Python или R) для решения задач обработки социологических опросов, сложно для исследователей-социологов. В этой связи важно создать программный продукт, который не будет зависеть от зарубежных технологий и сможет обеспечить полную независимость от внешних поставщиков и рисков, связанных с санкциями.

Задача исследования – разработка архитектуры отечественного модуля прогнозирования в информационно-аналитической системе (ИАС) социологических исследований, включающего современные подходы и методы предиктивной аналитики для анализа и прогнозирования социальных процессов.

Теоретическая часть ***Анализ существующих методов прогнозирования***

В социологических исследованиях используется широкий спектр методов предиктивного анализа, включая традиционные статистические подходы и современные методы машинного обучения. Тем не менее, многие из них сталкиваются с проблемами интерпретации результатов и адаптации под специфические задачи социологии. Линейная и логистическая регрессии, хотя и просты в использовании и интерпретации, ограничены в способности выявлять сложные нелинейные зависимости [6-8]. Модели временных рядов, такие как ARIMA, SARIMA и SARIMAX, не учитывают изменение структуры данных и зависимость от внешних факторов, что может привести к ошибкам в прогнозах при изменении социально-экономической ситуации [9, 10]. Методы машинного обучения, такие как случайные леса и градиентный бустинг, требуют значительных вычислительных ресурсов и специализированных навыков для настройки параметров, что ограничивает их применение исследователями без технической подготовки [4, 11, 12].

Для преодоления этих ограничений в разрабатываемом модуле прогнозирования были выбраны методы линейной и логистической регрессии, модели временных рядов и нейронные сети LSTM, которые обеспечивают баланс между интерпретируемостью, адаптивностью и вычислительной эффективностью. Линейные и логистические модели служат базовым уровнем для простых прогнозов, а нейронные сети LSTM учитывают временные зависимости и нелинейные отношения в данных, что критически важно для прогнозирования сложных социальных процессов [13-15]. Такой подход позволяет не только повысить точность и надежность прогнозов, но и обеспечивает удобство использования модуля для исследователей, не обладающих продвинутыми навыками программирования и создания отечественного программного продукта, особенно в условиях формирования цифрового суверенитета России.

Архитектура модуля

В ходе данного исследования был разработан комплексный модуль прогнозирования в информационно-аналитической системе (ИАС) социологических исследований. Архитектура комплексного модуля включает несколько компонентов, предназначенных для сбора, обработки данных, аналитического моделирования и графического представления результатов (рисунок 1). Для реализации архитектуры использовались: Python – для модулей анализа данных и JavaScript – для визуализации результатов.

Ядро модуля состоит из двух основных классов: DataManager – отвечает за загрузку и предварительную обработку данных. Функции данного класса включают загрузку данных и их предобработку, что является критически важным для подготовки данных к анализу. ConfigManager – управляет конфигурациями системы, загружая и применяя настройки подключения к базе данных и онтологиям. Это позволяет адаптировать систему под различные исследовательские задачи и условия работы.

Компонент комплексного модуля DataProcessingModule анализирует поступившие данные и выбирает наиболее подходящий метод прогнозирования. Компонент StatisticalModule выполняет анализ данных с использованием статистических методов, таких как линейная и логистическая регрессии.

Компонент SimpleForecastingModule осуществляет простое прогнозирование на основе закона больших чисел (ЗБЧ) и центральной предельной теоремы (ЦПТ). Компонент TimeSeriesModule производит анализ и прогнозирование временных рядов на основе модели ARIMA, а SeasonalModule анализирует данные на предмет сезонных изменений и их влияние на результаты на основе моделей SARIMA и SARIMAX. Компонент LSTMModule использу-

ет модели LSTM для анализа числовых и категориальных данных с длительными временными зависимостями, а BERTLSTMModule анализирует текстовые открытые ответы с использованием технологий BERT и LSTM.

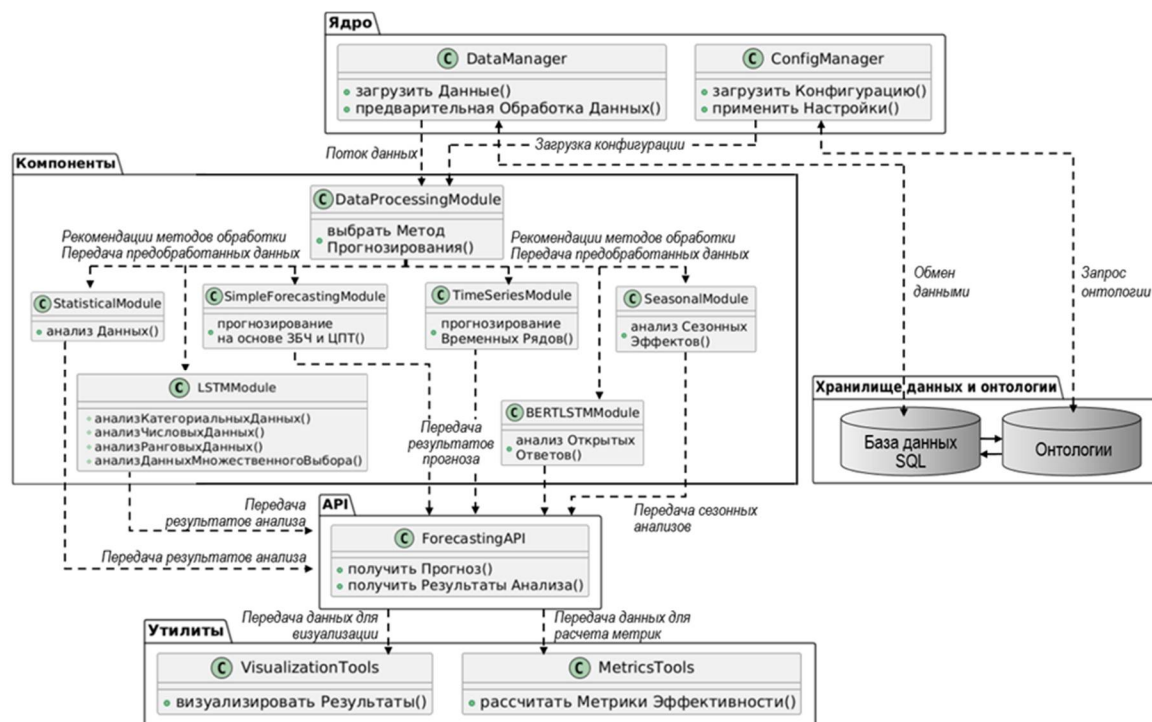


Рисунок 1 – Архитектура комплексного модуля прогнозирования

Figure 1 – Architecture of forecasting framework ork

ForecastingAPI предоставляет интерфейсы для получения прогнозов и результатов анализа, что позволяет пользователям ИАС получать необходимую информацию в удобной форме.

Утилита VisualizationTools предназначена для визуализации результатов анализа; а MetricsTools используется для расчета метрик эффективности работы моделей.

DataManager также взаимодействует с внешними системами, такими как SQL база данных и система управления онтологиями, для извлечения и интеграции данных, необходимых для анализа.

Архитектура обеспечивает модулю прогнозирования в системе социологических исследований гибкость в подходах к разнообразным научным и прикладным исследованиям, гарантируя при этом высокую степень точности и способность к адаптации при анализе комплексных социологических данных.

Модуль прошел экспериментальную апробацию в составе ИАС социологических исследований [16], показав достаточно высокую точность прогнозов.

Реализация комплексного модуля прогнозирования и используемые технологии.

Прогнозирование на основе закона больших чисел и центральной предельной теоремы

Прогнозирование на основе закона больших чисел и центральной предельной теоремы основано на том, что при увеличении размера выборки распределение средних значений приближается к нормальному, что позволяет делать достоверные прогнозы на основе выборочных данных.

В компоненте SimpleForecastingModule реализован алгоритм автоматического формирования репрезентативной выборки респондентов (автовыборка), учитывающий демографические и социальные параметры, такие как пол, возраст, тип населенного пункта, социальное положение и уровень образования. Алгоритм контролирует квоты по каждой категории, определяя необходимое количество респондентов для репрезентативной выборки, пересчи-

тывает частные и общие коэффициенты, обеспечивая пропорциональное представительство категорий. UML-диаграмма деятельности отражает этот процесс (рисунок 2).

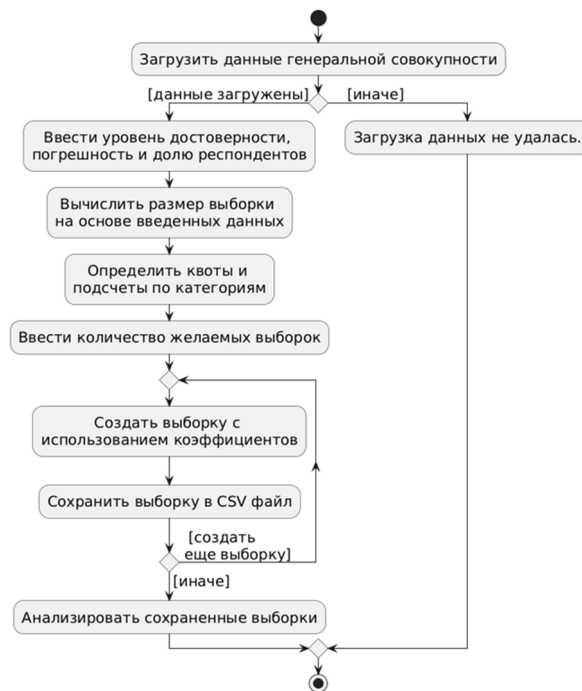


Рисунок 2 – Алгоритм автоматического формирования репрезентативной выборки респондентов

Figure 2 – Algorithm for automatic formation of a representative sample of respondents

При добавлении респондентов в выборку алгоритм минимизирует отклонения и ошибки, соответствуя требованиям закона больших чисел и обеспечивая применение центральной предельной теоремы для последующего анализа данных. Алгоритм используется, когда количество опрошенных респондентов превышает заданный объем выборки, а также для контроля набора респондентов. Точность прогнозирования напрямую зависит от числа опрошенных респондентов, при увеличении размера выборки распределение средних значений приближается к нормальному, что позволяет делать достоверные прогнозы на основе выборочных данных (рисунок 3).

Формирование выборки осуществляется на основе отбора анкет с учетом весовых коэффициентов, которые рассчитываются для каждой категории респондентов. Более высокие весовые коэффициенты повышают вероятность выбора анкеты. Функция `df.sample(n = sample_size, weights = coefficient')` в библиотеке `pandas` (Python) выбирает `sample_size` количество случайных элементов из датафрейма `df`. Элементы выбираются не с равной вероятностью, а с вероятностями, определяемыми весами (коэффициентами), указанными в столбце `coefficient`. Это обеспечивает создание статистически обоснованных и репрезентативных выборок, что важно для точности последующих анализов и выводов.

Оптимальное количество выборок определяется исходя из целей исследования, доступных ресурсов, желаемой точности и уровня доверия к результатам. Например, размер выборки n для генеральной совокупности $N0$ рассчитывается по формуле:

$$n = \frac{Z^2 \cdot p(1-p)}{\delta^2},$$

где n – размер выборки, Z – значение z-критерия, соответствующее желаемому уровню доверия (например, $Z = 1,96$ для уровня доверия 95 %), p – предполагаемая пропорция интересующего признака в генеральной совокупности (если неизвестна, используется $p = 0,5$, что дает максимальный размер выборки), δ – допустимая ошибка выборки (например, 0,05 или 5 %).

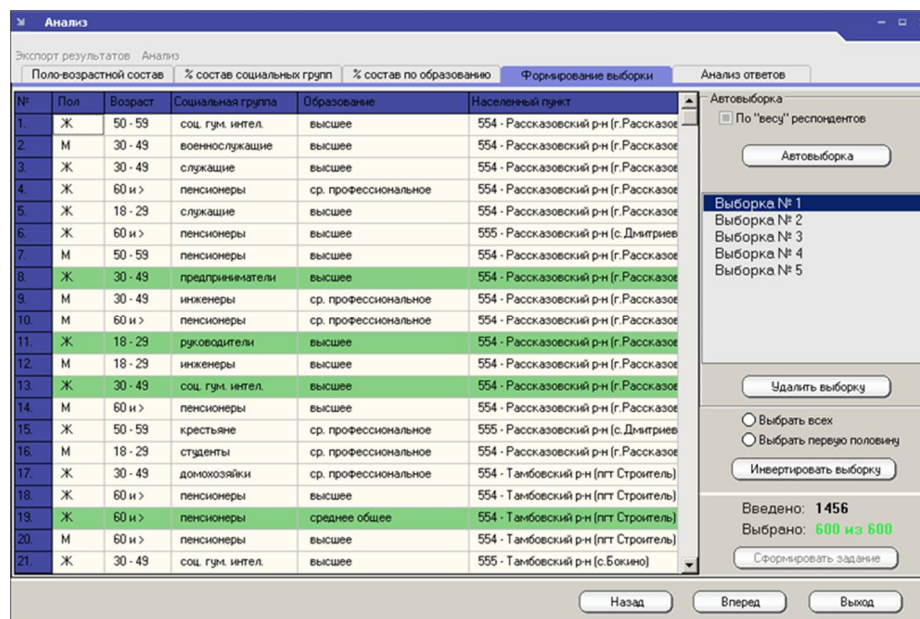


Рисунок 3 – Окно интерфейса «Формирование автовыборки респондентов»
Figure 3 – Interface window «Forming respondents auto-selection»

Для каждой категории респондентов рассчитывается количество необходимого числа анкет с учетом коэффициентов. Таким образом, алгоритм обеспечивает репрезентативное покрытие всех категорий генеральной совокупности, минимизируя искажения, возникающие при случайной выборке.

Прогнозирование на основе моделей ARIMA, SARIMA и SARIMAX

Модели ARIMA (Авторегрессионная интегрированная модель скользящего среднего), SARIMA (Сезонная ARIMA) и SARIMAX (Сезонная ARIMA с экзогенными переменными) широко используются для анализа временных рядов, особенно если данные демонстрируют явные тренды и сезонные колебания. Эти модели использованы при разработке компонент TimeSeriesModule (ARIMA) и SeasonalModule (SARIMA и SARIMAX) в ИАС социологических исследований для прогнозирования уровня удовлетворенности граждан услугами и анализа социально-экономических показателей, а также для учета внешних факторов, таких как экономические и социальные индикаторы [8].

Так как компоненты TimeSeriesModule (ARIMA) и SeasonalModule (SARIMA и SARIMAX) похожи по архитектуре, то в статье рассмотрена архитектура SeasonalModule, отражающая комплексный процесс обработки данных и моделирования. Архитектура этой компоненты состоит из нескольких основных блоков (рисунок 4). Главный блок инициирует процессы, управляет качеством данных на каждом этапе и принимает корректирующие меры в случае обнаружения проблем. Блок обработки данных состоит из модулей загрузки данных, предобработки данных и интеграции экзогенных (внешних) данных. Включает процедуры проверки качества данных и интеграции экзогенных данных, чтобы убедиться в их пригодности для дальнейшего анализа. Экзогенные переменные обрабатываются в отдельном блоке, который включает загрузку экзогенных данных и их предобработку с последующей проверкой качества. Обработанные экзогенные данные затем передаются в блок интеграции данных для совместного анализа. Блок статистического анализа отвечает за проверку стационарности временных рядов (анализ стационарности), проведение теста Дики – Фуллера для формальной оценки стационарности, построение коррелограмм для анализа автокорреляций и контроль качества анализа с валидацией результатов. Этот блок проверяет корректность данных перед их использованием для подбора параметров модели.

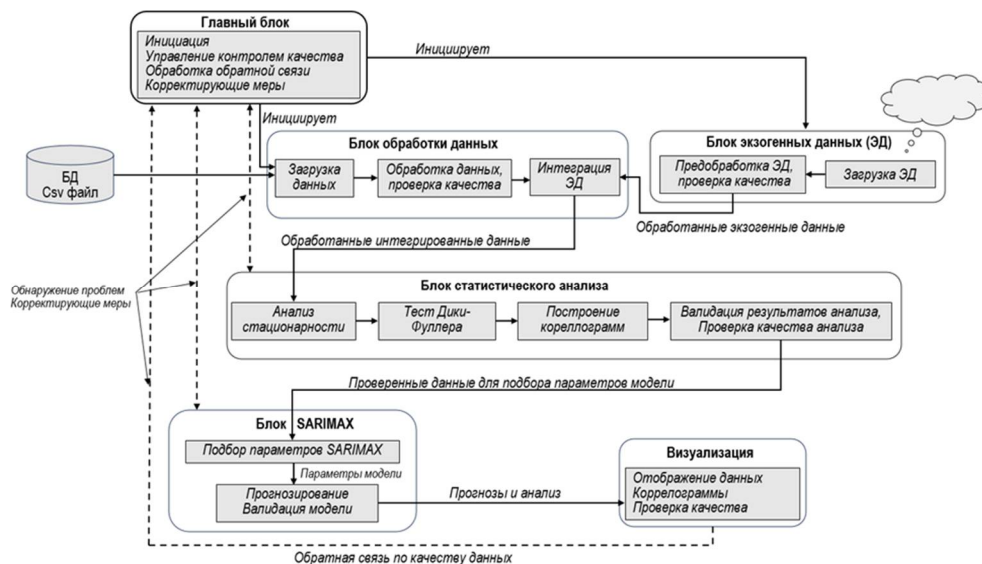


Рисунок 4 – Архитектура компоненты SeasonalModule (SARIMA и SARIMAX)
Figure 4 – Architecture components of SeasonalModule (SARIMA and SARIMAX)

В блоке SARIMAX реализованы подбор оптимальных параметров модели с использованием информационного критерия Акаике (AIC) и прогнозирование на основе этой оптимальной модели. Затем осуществляются валидация модели, проверка точности прогнозов на тестовых данных. Блок визуализации занимается отображением данных, включая визуализацию коррелограмм и результатов проверки качества. Этот блок также обеспечивает обратную связь по качеству данных, позволяя системе корректировать ошибки и улучшать точность прогнозов.

Ниже приведен пример прогнозирования оценки качества медицинского обслуживания населения Тамбовской области (ежеквартально выборочно проводился опрос населения по 1000 человек). Рассмотрим результаты работы компонента SeasonalModule (SARIMA и SARIMAX). На рисунке 5 представлены коррелограммы автокорреляционной функции (ACF) и частичной автокорреляционной функции (PACF), которые используются для идентификации сезонных и автокорреляционных свойств временного ряда.

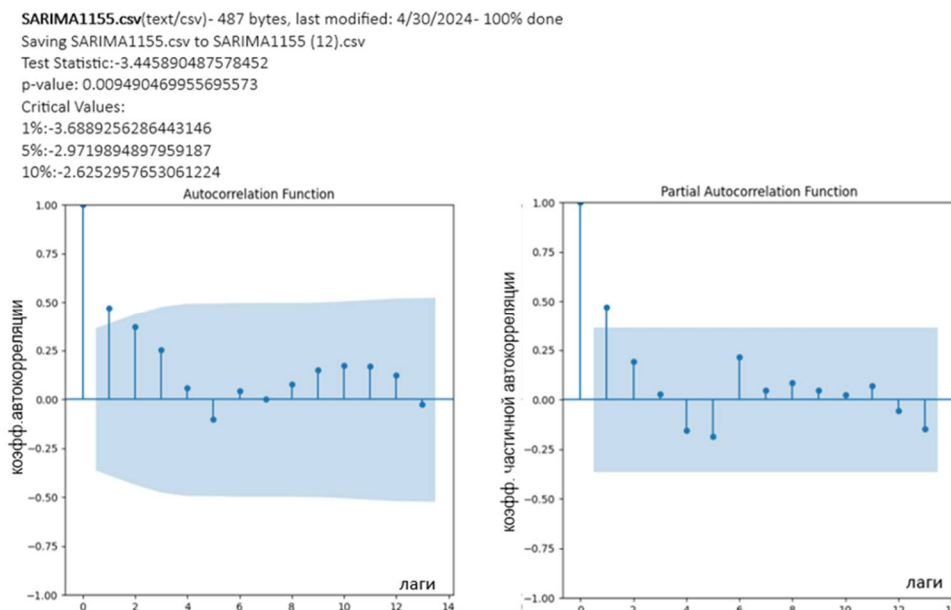


Рисунок 5 – Коррелограммы ACF и PACF для идентификации параметров модели
Figure 5 – ACF and PACF corellograms for identification of model parameters

Параметры теста Дики – Фуллера (Test Statistic: -3.445890487578452, p-value: 0.009490469955695573) показывают, что ряд стационарен, так как тестовая статистика меньше критических значений на уровнях 1 %, 5 % и 10 %. Это позволяет утверждать, что временной ряд готов для дальнейшего моделирования, включая учет сезонных компонентов.

Результаты подгонки и прогнозирования модели SARIMAX показаны на рисунке 6. Временной ряд разделен на обучающую и тестовую выборки. График демонстрирует, что модель SARIMAX успешно адаптируется к историческим данным (синяя и зелёная линии), с учетом экзогенных переменных и сезонных эффектов. Прогнозы, выполненные моделью (красная линия), показывают ожидаемое поведение системы в будущем, учитывая внешние факторы, влияющие на целевую переменную. Это видно по различным цветам, обозначающим данные для обучения, тестирования и прогнозы внутри и вне выборки.

Проверка качества модели осуществлялась по метрикам точности MSE, RMSE, MAE, MAPE, значения которых показаны внизу на рисунке 6 и подтверждают адекватность модели. Как следует из анализа этих показателей, среднеквадратичная ошибка (RMSE = 0,59) указывает на то, что среднее отклонение прогнозов от фактических данных составляет около 5-6 %. Средняя абсолютная ошибка (MAE = 5,14 %) свидетельствует о том, что модель не допускает крупных выбросов ошибок. Коэффициент средних абсолютных процентных отклонений (MAPE = 6,61 %) указывает на среднее отклонение прогнозов на 6,61 % от фактических значений, что является удовлетворительным результатом для моделей временных рядов.

Таким образом можно сделать вывод, что модель SARIMAX показывает достаточно точные результаты, с ошибкой около 5-6 %. График прогнозов (рисунок 6) демонстрирует хорошее соответствие модели с реальными данными.

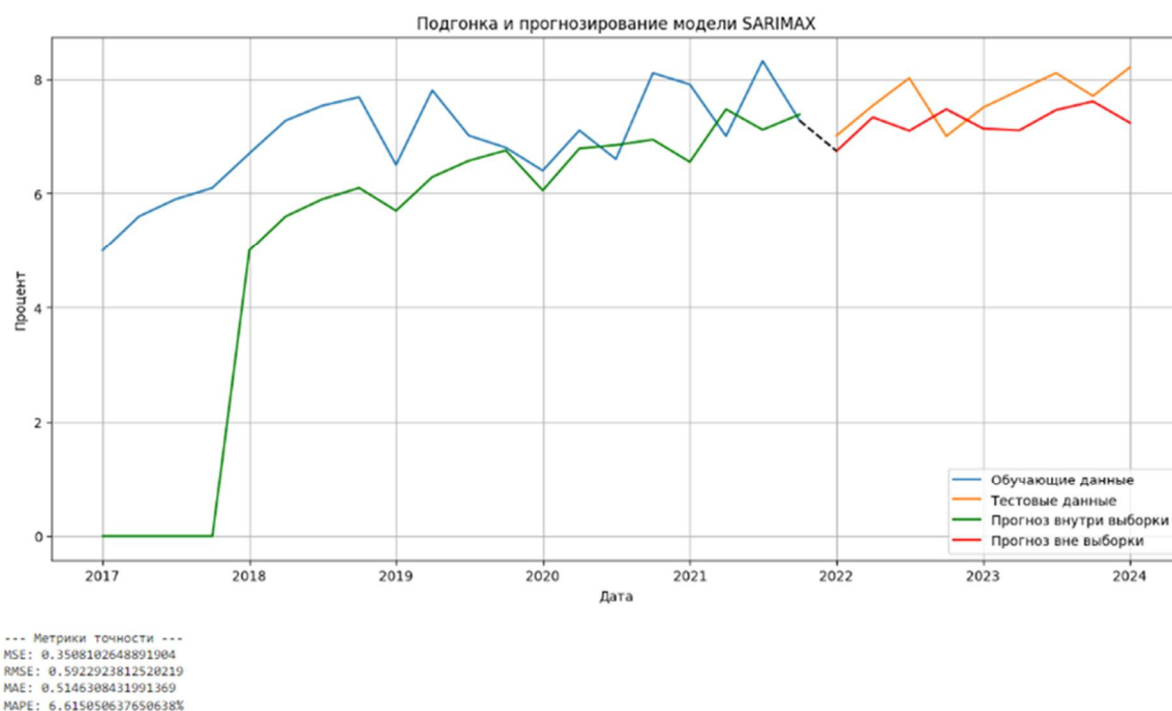


Рисунок 6 – Результаты прогнозирования компонента SeasonalModule
Figure 6 – The results of forecasting SeasonalModule component

Работа компонента SeasonalModule, включая модели SARIMA и SARIMAX, демонстрирует высокую адаптивность к сезонным изменениям и внешним влияниям. Модели успешно используют как исторические данные, так и экзогенные переменные для точного прогнозирования временных рядов, что особенно важно для социологических исследований, где влияние сезонных и внешних факторов может быть значительным.

Прогнозирование на основе модели LSTM

Модель Long Short-Term Memory (LSTM) – рекуррентная нейронная сеть, способная запоминать долгосрочные зависимости, т.е. она хорошо приспособлена для анализа и прогнозирования временных рядов. Это особенно важно при работе с социологическими данными, где текущее состояние сильно зависит от предшествующих событий и трендов.

Основными преимуществами LSTM перед SARIMAX являются способность обрабатывать сложные долгосрочные зависимости без явно выраженной сезонности и обучение на последовательностях данных. Однако SARIMAX лучше справляется с временными рядами, содержащими чётко выраженную сезонность и внешние воздействия.

Проект компонента LSTM показан на рисунке 7 и включает несколько классов, обеспечивающих полный цикл обработки данных (загрузка и предобработка, построение и обучение модели, визуализация результатов).

Класс «ОбработчикФайлов» отвечает за загрузку данных, а «ПредобработчикДанных» занимается их подготовкой для анализа. «СоздательНабораДанных» формирует обучающие и тестовые выборки, а «LSTMМодель» предоставляет методы для разработки, компиляции и обучения модели.

Для демонстрации результатов используются классы «Постобработчик» и «Визуализатор», которые возвращают данные к исходному масштабу и создают графическое отображение прогноза.

Структура нейронной сети состоит из входного слоя, LSTM слоя с 50 нейронами, слоя Dropout с вероятностью отключения 18 %, и полносвязного выходного слоя для генерации предсказаний. В процессе обучения структура нейронной сети не меняется, но веса и параметры сети подстраиваются под данные. Слои Регуляризация и Dropout используются для предотвращения переобучения, улучшая обобщающую способность модели.

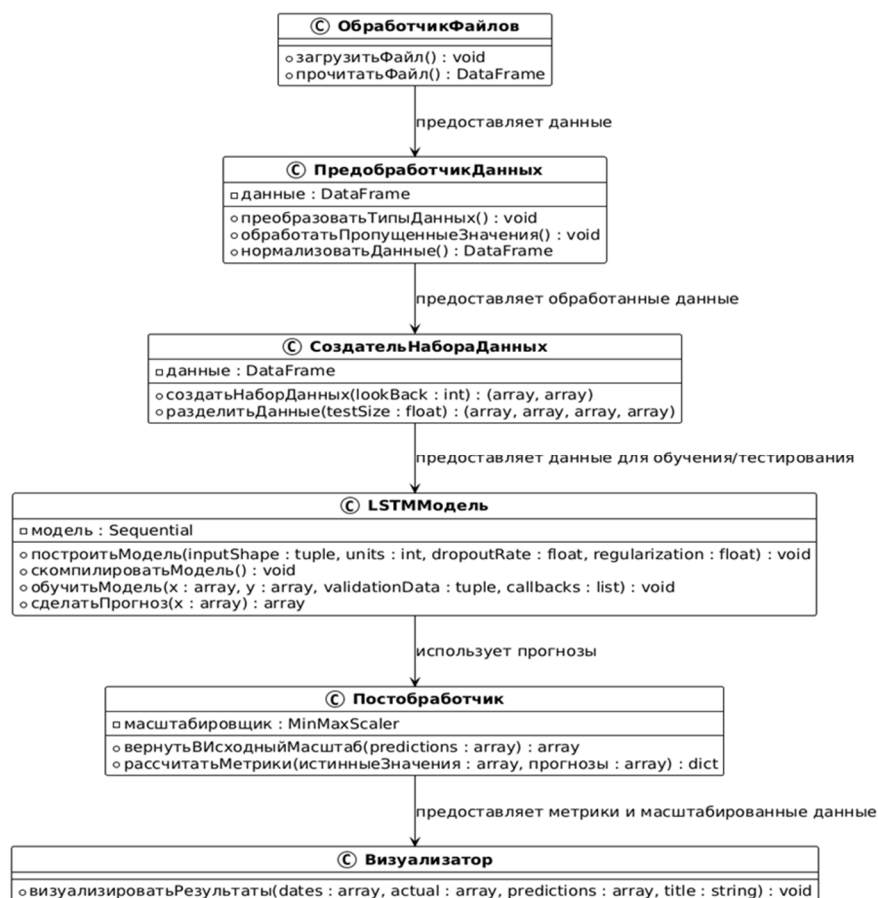


Рисунок 7 – Диаграмма классов компонента LSTM

Figure 7 – Class diagram of LSTM component

Для реализации компонента LSTM был использован язык программирования Python и библиотеки, показанные на рисунке 8.

Результат прогнозирования нейронной сети представлен на рисунке 9. По мере эксплуатации этой модели планируется дальнейшее совершенствование ее характеристик.

В ходе обучения модель показала хорошие результаты на тренировочных данных, однако наблюдалось небольшое снижение производительности на тестовых данных. Метрики ошибок, такие как MSE и MAE, остаются низкими, что свидетельствует о высоком качестве предсказаний модели.

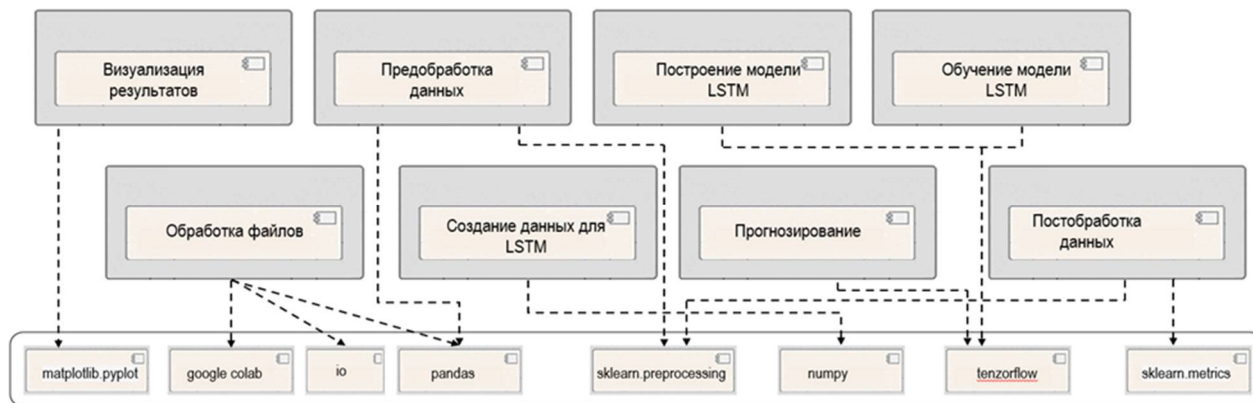


Рисунок 8 – Библиотеки Python, использованные в компоненте LSTM

Figure 8 – Python libraries used in LSTM component

Training Data Metrics:
MSE: 0.06317859405703335
MAE: 0.1984813827938504
R²: 0.9820347626795932
RMSE: 0.25135352405930844
Testing Data Metrics:
MSE: 0.06470824731214658
MAE: 0.23881912231445357
R²: 0.8849311523207174
RMSE: 0.25437815808780945

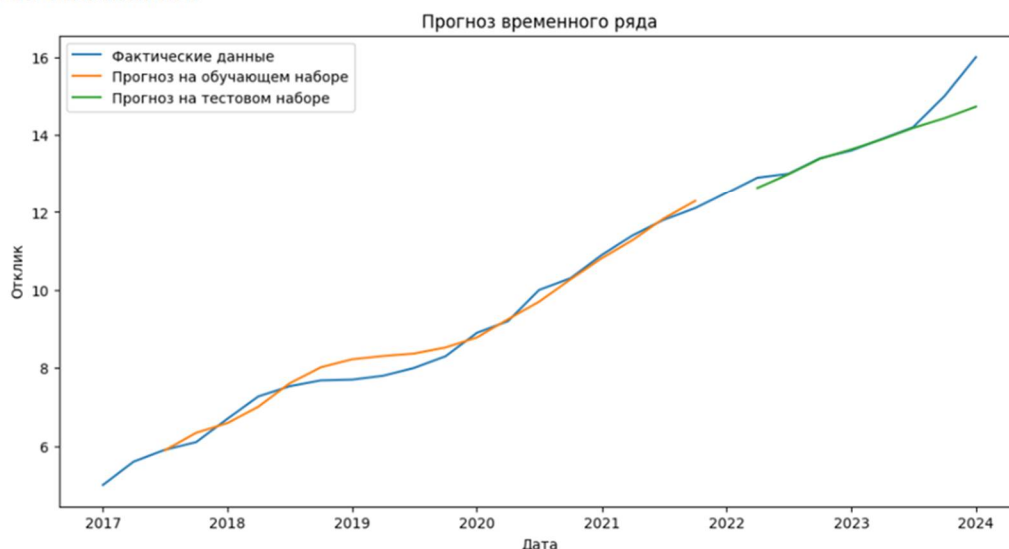


Рисунок 9 – Результаты прогнозирования категориальных данных опросов

Figure 9 – The results of forecasting categorical survey data

Проектирование компонента BERTLSTM для обработки текстовых ответов респондентов

Социологические опросы часто генерируют большие объемы данных, включающие ответы как на открытые, так и на закрытые вопросы. Эти данные содержат сложные шаблоны и зависимости, что требует параллельной обработки результатов в ИАС социологических ис-

следований для эффективного прогнозирования и поддержки принятия управленческих решений.

Для обработки текстовых ответов, особенно открытых вопросов, целесообразно использование языковой модели BERT (Bidirectional Encoder Representations from Transformers) [17]. BERT обеспечивает глубокое понимание языкового контекста, что значительно улучшает качество анализа текстовых данных.

В настоящее время компонент BERTLSTM находится в разработке. Последовательность обработки данных, иллюстрирующая интеграцию методов обработки закрытых и открытых вопросов, показана на рисунке 10.

Можно выделить следующие этапы обработки данных в этом компоненте.

На начальном этапе происходит ввод данных в компонент BERTLSTM, где смешанные данные опросов, включающие как текстовые, так и числовые ответы, загружаются для дальнейшей обработки. Следующий этап включает предобработку открытых ответов, где осуществляются очистка текста, приведение его к нижнему регистру, удаление стоп-слов и лемматизация. Эти процессы направлены на улучшение качества текстовых данных и их подготовку для последующего анализа.

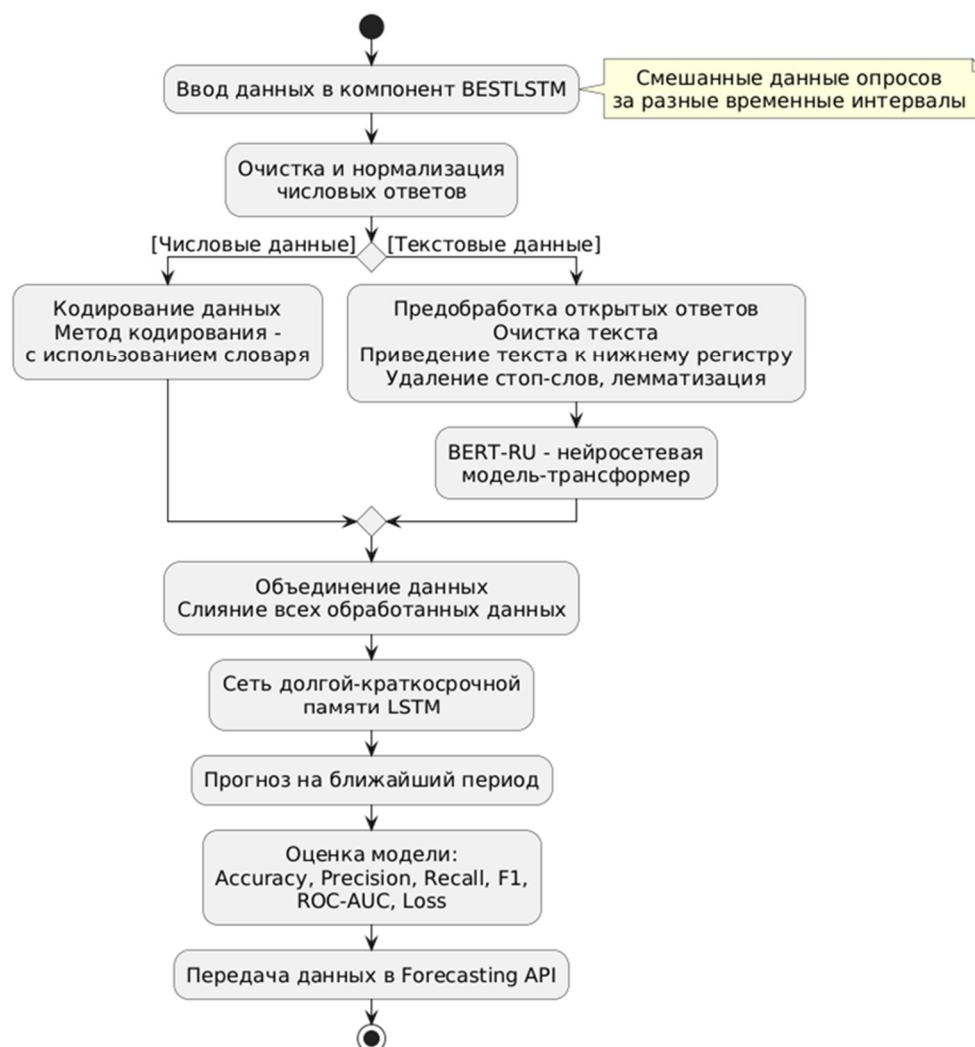


Рисунок 10 – Алгоритм процесса обработки данных в компоненте BERTLSTM

Figure 10 – Algorithm for data processing in BERTLSTM component

После предобработки текстовых данных применяется нейросетевая модель-трансформер BERT-RU, которая преобразует текстовые ответы в векторные представления. Этот шаг поз-

воляет учитывать контекст и смысловую нагрузку текста, что особенно важно для глубокого анализа данных.

Одновременно с этим происходит кодирование категориальных данных, что включает преобразование очищенных и нормализованных числовых данных в числовые представления с использованием словаря. Этот процесс необходим для унификации и стандартизации данных перед их объединением. После подготовки текстовых и числовых данных происходит объединение данных, где все обработанные данные сливаются в единый массив. На этапе анализа используется сеть долгосрочной кратковременной памяти (LSTM), которая позволяет выявлять временные зависимости и тренды в данных.

Завершающим этапом является прогноз на ближайший период, где обученная модель используется для предсказания результатов опросов на основе объединенных данных. После формирования прогноза осуществляется оценка качества и точности модели с использованием метрик Accuracy, Precision, Recall, F1, ROC-AUC и Loss. Итоговые данные передаются в ForecastingAPI, который предоставляет интерфейсы для доступа к прогнозам и аналитическим данным.

Апробация модуля прогнозирования на многолетних данных опросов населения

Проверка работы комплексного модуля прогнозирования проводилась на большом массиве исторических данных по определению рейтингов политических партий в Тамбовской области за период с 2017 по 2023 годы. Для обеспечения репрезентативного покрытия всех категорий генеральной совокупности был использован алгоритм автоматического формирования репрезентативной выборки, что позволило сформировать ежемесячные выборки по 110-120 респондентов. Данные каждого месяца представляют собой процент голосов, поданных за партию №1¹.

Прогнозирование проводилось с использованием моделей SARIMAX и LSTM. Интеграция моделей в разработанном модуле позволяет эффективно обрабатывать временные ряды, содержащие сложные нелинейные зависимости и сезонные эффекты, характерные для социологических данных. LSTM включает один слой с 50 ячейками, за которым следует слой Dropout с коэффициентом 0,2 для предотвращения переобучения. Входные данные имеют форму, определяемую числом временных шагов, а функция рекуррентного отключения нейронов в LSTM установлена на уровне 0,18. На выходе используется слой с одним нейроном и линейной функцией активации, что позволяет моделировать сложные зависимости и тренды временного ряда. Такая конфигурация обеспечивает достаточную точность предсказаний и способность эффективно обрабатывать нелинейные и сезонные паттерны в данных. Параметры модели SARIMAX были выбраны как $(p = 1, d = 1, q = 1)$ и сезонное разложение $(P = 1, D = 1, Q = 1, m = 4)$. Результаты прогнозирования представлены на рисунке 11.

Результаты показали, что модель LSTM значительно превосходит SARIMAX по ключевым метрикам качества как на обучающем, так и на тестовом наборах данных. Модель LSTM на обучающей выборке продемонстрировала низкие значения MSE (6,528) и MAE (1,805), а также высокий коэффициент детерминации R^2 (0,897). На тестовой выборке: показатели MSE (1,033) и MAE (0,803) еще ниже, т.е. модель не переобучена и хорошо обобщает закономерности на новых данных. Однако R^2 снизился до 0,552, что может быть связано с изменением динамики данных или внешних факторов.

Модель SARIMAX показала низкую эффективность из-за ограниченной способности улавливать нелинейные паттерны. Таким образом, можно рекомендовать использование модели LSTM для прогнозирования рейтингов предвыборных кампаний в модуле прогнозирования ИАС социологических исследований. Дальнейшее совершенствование модуля прогнозирования планируется за счет оптимизации архитектуры LSTM и SARIMAX и исследования их применимости к другим социологическим данным.

¹ Название партии не раскрывается

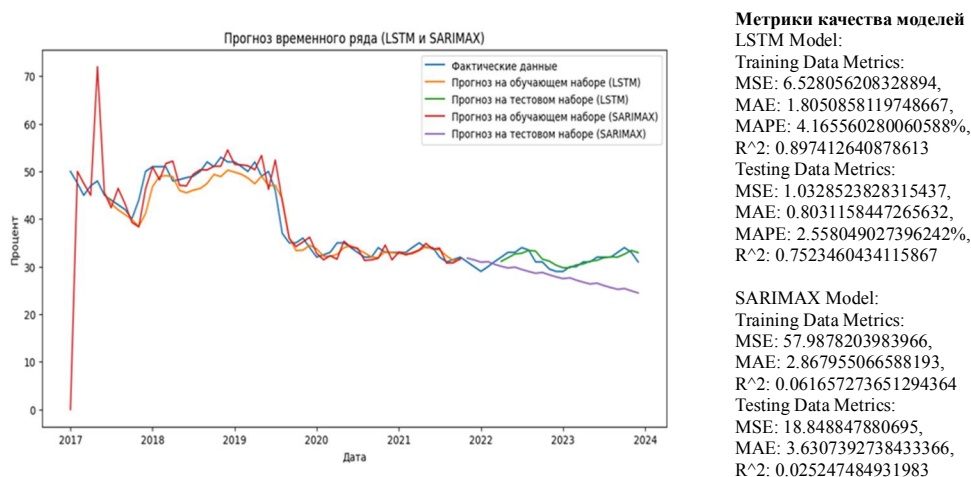


Рисунок 11 – Результаты прогнозирования на модели SARIMAX и на модели LSTM
Figure 11 – The results of forecasting on SARIMAX model and on LSTM model

Заключение

Разработана архитектура модуля прогнозирования для ИАС социологических исследований, учитывающая специфику обработки социологических данных, таких как сложная структура опросов, неоднородность информации, а также влияние социально-демографических факторов на прогнозируемые результаты. Предложенная архитектура объединяет методы статистического анализа, моделирования временных рядов и машинного обучения, что позволяет эффективно решать задачи, связанные с прогнозированием социальных процессов, таких как динамика общественного мнения и сегментация респондентов по социальным группам.

Научная новизна работы заключается в создании единой архитектуры, адаптированной под задачи социологии в условиях ограниченного доступа к зарубежным технологиям, что повышает гибкость и адаптивность системы к изменяющимся требованиям исследований. Интеграция различных методов в единый интерфейс облегчает работу исследователей и упрощает внедрение новых подходов для анализа социологических данных.

Практическая значимость работы заключается в создании доступного инструмента для социологических исследований, снижающего технический барьер для исследователей. Модуль может быть использован для решения различных задач, обеспечивая анализ данных для прогнозирования социальных трендов, полезен для государственных органов, научных учреждений и исследовательских центров.

Библиографический список

1. **Шляпентох В.Э.** Проблемы качества социологической информации: достоверность, репрезентативность, прогнозистический потенциал // ЦСП. 2006. С. 547-598.
2. **Vasileiadou E., Vliegthart R.** Studying dynamic social processes with ARIMA modelling // International Journal for Social Research Methodology. 2013. DOI: 10.1080/13645579.2013.81625.
3. **Colbaugh R., Glass K., Bauer T.** Leveraging Sociological Models for Predictive Analytics // ResearchGate. 2012. URL: https://www.researchgate.net/publication/234005594_Leveraging_Sociological_Models_for_Predictive_Analytics (дата обращения: 27.07.2024).
4. **Горошникова Т.А.** Вычислительная социология: сдвиг парадигмы или «эконометрика» социологии // Гуманитарные науки. Вестник Финансового университета. 2021. Т. 11, № 1. С. 37-42. DOI: 10.26794/2226-7867-2021-11-1-37-42.
5. **Воротилова М.Ю.** Анализ текущего состояния импортозамещения в сфере программного обеспечения в России // Экономика и бизнес: теория и практика. 2024. №1-1 (107). URL: <https://cyberleninka.ru/article/n/analiz-tekuschego-sostoyaniya-importozamescheniya-v-sfere-programmnogo-obespecheniya-v-rossii> (дата обращения: 16.09.2024).
6. **Леонова Л.Б.** Корреляционно-регрессионный анализ как инструмент прогнозирования влияния функционирования социально-экономического кластера в сфере ЖКХ на экономику региона //

Вестник Уфимского государственного нефтяного технического университета. 2020. № 3 (33). URL: <https://cyberleninka.ru/article/n/korrelyatsionno-regressionnyy-analiz-kak-instrument-prognozirovaniya-vliyaniya-funktsionirovaniya-sotsialno-ekonomicheskogo> (дата обращения: 06.07.2024).

7. **Maravelakis P.** The use of statistics in social sciences // *Journal of Humanities and Applied Social Sciences*. 2019. Т. 1, № 2. С. 87-97. DOI: 10.1108/JHASS-08-2019-0038.

8. **Воробьев А.А., Ноздрачева Е.Д.** Методика прогнозирования результатов выборов с использованием комбинирования методов аналитического и трендового моделирования // *Мониторинг общественного мнения: экономические и социальные перемены*. 2022. № 2. С. 24-41. DOI: 10.14515/monitoring.2022.2.1921.

9. **Фелькер М.Н., Чеснов В.В.** Исследование влияния изменения параметров модели ARIMA на качество прогноза для коротких наборов данных // *Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника*. 2021. № 3. URL: <https://cyberleninka.ru/article/n/issledovanie-vliyaniya-izmeneniya-parametrov-modeli-arima-na-kachestvo-prognoza-dlya-korotkih-naborov-dannyh> (дата обращения: 30.07.2024).

10. **Heiberger R.H.** Applying Machine Learning in Sociology: How to Predict Gender and Reveal Research Preferences // *Köln Z Soziol.* 2022. Т. 74, № Suppl 1. С. 383-406. DOI: 10.1007/s11577-022-00839-2.

11. **Garip F., Macy M.W.** Machine Learning in Sociology: Current and Future Applications // *The Oxford Handbook of the Sociology of Machine Learning*. 2023. URL: <https://doi.org/10.1093/oxfordhb/9780197653609.013.11> (дата обращения: 01.08.2024).

12. **Siami-Namini S., Tavakoli N., Siami Namin A.** A Comparison of ARIMA and LSTM in Forecasting Time Series // *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018. Pp. 1394-1401. DOI: 10.1109/ICMLA.2018.00227.

13. **Li K., Qian N.** Application of improved LSTM model and visualization analysis in bike-sharing demand prediction // *2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*. 2023. P. 656-661. DOI: 10.1109/CIPAE60493.2023.00127.

14. **Chen S., Han X., Shen Y., Ye C.** Application of Improved LSTM Algorithm in Macroeconomic Forecasting // *Computational Intelligence and Neuroscience*. 2021. DOI: 10.1155/2021/4471044.

15. **Luque C., Sosa J.** Bayesian analysis for social science research // *Model-Assisted Statistics and Applications*. 2024. DOI: 10.3233/MAS-231460.

16. **Кузнецов А. М.** Проактивная система обработки информации социологических исследований // Свидетельство о государственной регистрации программы для ЭВМ № 2024617985. А. М. Кузнецов. Заявка № 202461581. Дата поступления: 22 марта 2024 г. Зарегистрировано в Реестре программ для ЭВМ: 08 апреля 2024 г.

17. **Larson J., & al.** A Comparison of LSTM and BERT for Small Corpus // *arXiv preprint arXiv:2009.05451*. 2020. URL: <https://arxiv.org/abs/2009.05451>.

UDC 004.9

FORECASTING MODULE FRAMEWORK ARCHITECTURE IN SOCIOLOGICAL RESEARCH INFORMATION-ANALYTICAL SYSTEM

A. M. Kuznetsov, graduate student, TSTU, Tambov, Russia;
orcid.org/0000-0003-2640-3417, e-mail: 068t@mail.ru

This article presents the framework architecture for the forecasting module in the sociological research information-analytical system (IAS). The integration of modern approaches and predictive analytics methods for analyzing and forecasting social processes is considered, including linear and logistic regression, time series models (ARIMA, SARIMA, SARIMAX), machine learning methods, and neural networks (LSTM, BERT). The module provides good forecast quality, flexibility and adaptability of analysis taking into account complex dependencies in the data.

Keywords: Information-analytical system (IAS), sociological research, framework, forecasting, neural networks, LSTM, BERT, machine learning, time series, regression analysis, SARIMAX.

DOI: 10.21667/1995-4565-2024-90-77-90

References

1. **Shlyapentokh V.E.** Problemy kachestva sotsiologicheskoy informatsii: dostovernost', reprezentativnost', prognosticheskiy potentsial. TsSP. 2006, pp. 547-598. (in Russian).
2. **Vasileiadou E., Vliegthart R.** Studying dynamic social processes with ARIMA modelling. *International Journal for Social Research Methodology*. 2013. DOI:10.1080/13645579.2013.81625.
3. **Colbaugh R., Glass K., Bauer T.** Leveraging Sociological Models for Predictive Analytics. ResearchGate. 2012. URL: https://www.researchgate.net/publication/234005594_Leveraging_Sociological_Models_for_Predictive_Analytics (дата обращения: 27.07.2024).
4. **Goroshnikova T.A.** Vychislitel'naya sotsiologiya: sdvig paradigmy ili «ekonometrika» sotsiologii. *Gumanitarnye nauki. Vestnik Finansovogo universiteta*. 2021. vol. 11, no 1. pp. 37-42. DOI: 10.26794/2226-7867-2021-11-1-37-42. (in Russian).
5. **Vorotilova M.Yu.** Analiz tekushchego sostoyaniya importozameshcheniya v sfere programmogo obespecheniya v Rossii. *Ekonomika i biznes: teoriya i praktika*. 2024. № 1-1 (107). URL: <https://cyberleninka.ru/article/n/analiz-tekushchego-sostoyaniya-importozameshcheniya-v-sfere-programmnogo-obespecheniya-v-rossii> (data obrashcheniya: 16.09.2024).
6. **Leonova L.B.** Korrelyatsionno-regressionnyy analiz kak instrument prognozirovaniya vliyaniya funktsionirovaniya sotsial'no-ekonomicheskogo klastera v sfere ZHKKH na ekonomiku regiona. *Vestnik UGNTU. Nauka, obrazovaniye, ekonomika. Seriya: Ekonomika*. 2020. no. 33. URL: <https://cyberleninka.ru/article/n/korrelyatsionno-regressionnyy-analiz-kak-instrument-prognozirovaniya-vliyaniya-funktsionirovaniya-sotsialno-ekonomicheskogo> (date of access: 06.07.2024). (in Russian)
7. **Maravelakis P.** The use of statistics in social sciences. *Journal of Humanities and Applied Social Sciences*. 2019. vol. 1, no. 2, pp. 87-97. DOI: 10.1108/JHASS-08-2019-0038.
8. **Vorobyev A.A., Nozdracheva E.D.** Metodika prognozirovaniya rezul'tatov vyborov s ispol'zovaniyem kombinirovaniya metodov analiticheskogo i trendovogo modelirovaniya. *Monitoring obshchestvennogo mneniya: ekonomicheskiye i sotsial'nyye peremeny*. 2022. no 2, pp. 24-41. DOI: <https://doi.org/10.14515/monitoring.2022.2.1921>. (in Russian)
9. **Felker M.N., Chesnov V.V.** Issledovaniye vliyaniya izmeneniya parametrov modeli ARIMA na kachestvo prognoza dlya korotkikh naborov dannykh. *Vestnik Yuzhno-Ural'skogo gosudarstvennogo universiteta. Seriya: Kompyuternyye tekhnologii, upravleniye, radioelektronika*. 2021, no. 3. URL: <https://cyberleninka.ru/article/n/issledovanie-vliyaniya-izmeneniya-parametrov-modeli-arima-na-kachestvo-prognoza-dlya-korotkikh-naborov-dannykh> (date of access: 30.07.2024). (in Russian)
10. **Heiberger R. H.** Applying Machine Learning in Sociology: How to Predict Gender and Reveal Research Preferences. *Köln Z Soziol*. 2022. vol. 74, no. Suppl 1. pp. 383-406. DOI: <https://doi.org/10.1007/s11577-022-00839-2>.
11. **Garip F., Macy M.W.** Machine Learning in Sociology: Current and Future Applications // *The Oxford Handbook of the Sociology of Machine Learning*. 2023. URL: <https://doi.org/10.1093/oxfordhb/9780197653609.013.11> (date of access: 01.08.2024).
12. **Siami-Namini S., Tavakoli N., Siami Namin A.A.** Comparison of ARIMA and LSTM in Forecasting Time Series. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018. P. 1394-1401. DOI: 10.1109/ICMLA.2018.00227.
13. **Li K., Qian N.** Application of improved LSTM model and visualization analysis in bike-sharing demand prediction. *2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*. 2023, pp. 656-661. DOI: 10.1109/CIPAE60493.2023.00127.
14. **Chen S., Han X., Shen Y., Ye C.** Application of Improved LSTM Algorithm in Macroeconomic Forecasting. *Computational Intelligence and Neuroscience*. 2021. DOI: 10.1155/2021/4471044.
15. **Luque C., Sosa J.** Bayesian analysis for social science research. *Model-Assisted Statistics and Applications*. 2024. DOI: <https://doi.org/10.3233/MAS-231460>.
16. **Kuznetsov A. M.** Proaktivnaya sistema obrabotki informatsii sotsiologicheskikh issledovaniy. Svidetel'stvo o gosudarstvennoy registratsii programmy dlya EVM № 2024617985. A. M. Kuznetsov. Zayavka № 202461581. Data postupleniya: 22 marta 2024 g. Zaregistrovano v Reestre programm dlya EVM: 08 aprelya 2024. (in Russian)
17. **Larson J., & al.** A Comparison of LSTM and BERT for Small Corpus. ArXiv preprint arXiv:2009.05451. 2020. URL: <https://arxiv.org/abs/2009.05451>.