СОЦИОЛОГИЯ ЖУРНАЛИСТИКИ

Метель Юрий Андреевич, кандидат экономических наук, главный экономист Отделения по Ставропольскому краю Южного главного управления Центрального банка Российской Федерации, г. Ставрополь, Россия; e-mail: akademik.st.2018@mail.ru, ORCID 0009-0001-5334-0020

Куницына Наталья Николаевна, доктор экономических наук, профессор, заведующая кафедрой финансов и кредита ФГАОУ ВО «Северо-Кавказский федеральный университет», г. Ставрополь, Россия; e-mail: nkunitcyna@ncfu.ru, ORCID 0000-0001-9336-8100

ОЦЕНКА ВЛИЯНИЯ НОВОСТНОГО ФОНА В СМИ НА УРОВЕНЬ ИНФЛЯЦИИ В РОССИИ[1]

В условиях нарастания информационных потоков, влияния медиа на общественные настроения новостной фон, как ключевой элемент медиаинформации, напрямую влияет на экономические решения, а анализ его динамики в реальном времени может существенно повысить точность экономических прогнозов и моделей. В статье представлены результаты изучения воздействия новостного фона, содержащегося в публикациях СМИ, на уровень цен и инфляции. Методический инструментарий, предложенный авторами, позволяет не только анализировать инфляционные процессы, но и прогнозировать их динамику с учетом медийного контекста. Обработка содержания 147 тыс. статей позволила отобрать 3,5 тыс. новостей об инфляции и ценах, разделив их на «проинфляционные» и «дезинфляционные». Для формирования тренировочной и тестовой выборок реализовано подключение к нейронной сети "GigaChat". Моделирование инфляционной тональности размеченных новостей осуществлено посредством предобработки методом TF-IDF и применения логистической регрессии, наивного байесовского классификатора, Random Forest и CatBoost, а также путем дообучения (fine-tuning) нейронных сетей на архитектуре BERT. Методическую основу исследования составили методы парсинга html-сайтов, обработки естественного языка (NLP), архитектуры глубоких нейронных сетей. В результате построен индекс инфляционного фона, содержащийся в публикациях российских СМИ, с высокой точностью описывающий «шоковые» события и позволяющий прогнозировать уровень инфляции. Полученные данные могут быть использованы для более точной настройки экономических моделей и создания инструментов прогнозирования, что подчеркивает значимость медийного контента в оценке экономической ситуации.

***Ключевые слова**: новостной фон, инфляция, индекс потребительских цен, текстовый анализ, NLP, Transformers, BERT.*

Как цитировать: *Метель Ю. А., Куницына Н. Н.* Оценка влияния новостного фона в СМИ на уровень инфляции в России // Вестник Московского университета. Серия 10. Журналистика. 2025. № 2. С. 3–32. DOI: 10.55959/msu.vestnik.journ.2.2025.332

Yuri A. Metel, PhD in Economics, Stavropol Territory Branch of the Southern Main Directorate of the Central Bank of the Russian Federation, Stavropol, Russia; e-mail: akademik.st.2018@mail.ru, ORCID 0009-0001-5334-0020

Natalia N. Kunitsyna, Doctor of Economics, Professor, Head of the Department "Finance and Credit", North Caucasus Federal University, Stavropol, Russia; e-mail: nkunitcyna@ncfu.ru, ORCID 0000-0001-9336-8100

ASSESSMENT OF THE IMPACT OF MEDIA NEWS BACKGROUNDS ON INFLATION IN RUSSIA

Given the growing number and variety of information flows, the influence of the media on public sentiment, the news background, as a key element of media information, directly affects economic decisions, and the analysis of its dynamics in real time can significantly increase the accuracy of economic forecasts and models. The article studied the impact of the news background contained in media publications on the prices and inflation level. The methodological tools proposed by the authors make it possible not only to analyze inflationary processes, but also to predict their dynamics taking into account the media context. Processing 147 thousand news articles by using a pre-trained neural network from Microsoft made it possible to select 3.5 thousand news about inflation and prices, dividing them into "proinflationary" and "disinflationary". To form training and test samples, a connection to the GigaChat neural network is implemented. The inflation tonality of the marked news was simulated by pre-processing using the TF-IDF method and using logistic regression, Naive Bayesian classifier, Random Forest and CatBoost, as well as by additional training (fine-tuning) of neural networks based on the BERT architecture, pre-trained in Russian. The methodological basis was made up of methods for parsing html sites, natural language processing (NLP), deep neural network architecture, and regression algorithms. As a result, we have the inflation background index contained in the Russian media publications, with high accuracy describing the "shock" events in the economy. The findings can be used to fine-tune economic models and create forecasting tools, highlighting the importance of media content in assessing the economic situation.

***Key words**: news background, inflation, consumer price index, text analysis, NLP, Transformers, BERT.*

To cite this article: Metel Yu. A., Kunitsyna N. N. (2025) Otsenka vliyaniya novostnogo fona v SMI na uroven' inflyatsii v Rossii [Assessment of the impact of media news backgrounds on inflation in Russia]. *Vestnik Moskovskogo Universiteta. Seriya 10. Zhurnalistika 2*: 3–32. DOI: 10.55959/msu.vestnik.journ.2.2025.332

DOI: 10.55959/msu.vestnik.journ.2.2025.332

Введение

Повсеместное распространение интернета расширяет доступ к данным, сайтам информационных агентств, новостным лентам, социальным сетям. При этом значительная часть информации представлена в неструктурированном виде, что существенно затрудняет ее интерпретацию и дальнейшее использование. Однако по мере развития искусственного интеллекта появились методы обработки естественного языка: речь идет уже не о восприятии текста на основе отдельных, не связанных между собой слов, а о понимании его в контексте. Обработка естественного языка (*Natural Language Processing, NLP*) ложится на искусственный интеллект и фокусируется на возможности компьютеров читать, понимать и извлекать смысл из человеческих слов[2].

При этом анализ текстов, с точки зрения их влияния на предпочтения потребителей товаров и услуг, представлен в меньших объемах. Обычно для этого используются «классические» подходы на основе оценки частоты упоминаний слов в массивах текстовых документов. Более того, целенаправленных мероприятий по оцифровке новостного фона с позиции его влияния на уровень цен в российских СМИ и вовсе не проводилось, несмотря на то что крупные издательства традиционно публикуют большой объем аналитического материала, связанного с ростом цен на товары и услуги и уровнем инфляционных ожиданий. С практической точки зрения, подобные публикации важны, поскольку содержат результаты уже проведенного группой экспертов анализа, т. е. полезные паттерны.

Так, СМИ во многом способствуют созданию общественного мнения, влияя на восприятие гражданами уровня инфляции и экономической нестабильности. Часто новости, связанные с ростом цен на товары и услуги, а также колебаниями валютного курса, становятся основными индикаторами экономической ситуации для широкой аудитории. Публикации в российских СМИ формируют инфляционные ожидания, что, в свою очередь, может сказываться на поведении потребителей и предпринимателей.

Кроме того, материалы российских медиа служат важным источником информации для исследования влияния информационного фона на экономику. Инфляционные процессы часто интерпретируются через призму внешних факторов, таких как санкции, изменение цен на нефть или глобальные экономические кризисы, а также внутренних проблем, как, например, рост государст-

венного долга или дефекты в управлении экономикой. Публикации, освещающие эти темы, могут влиять на общественные ожидания, создавая позитивный или негативный настрой, который отражается на экономических решениях – от потребительских расходов до инвестиционных стратегий.

Таким образом, тщательное изучение новостных материалов, посвященных инфляции, позволяет не только понять текущую ситуацию в экономике, но и предсказать возможные тренды. Подобный подход помогает не только в научных исследованиях, но и в практической работе государственных органов, финансовых аналитиков и экономистов, стремящихся минимизировать риски и управлять инфляционными процессами с учетом общественного восприятия и настроений.

В рамках исследования был проведен углубленный анализ способов обработки текстовой информации на основе новостных сообщений о темпах роста цен и уровне инфляции, а также разработаны методические рекомендации по оценке инфляционного фона в публикациях отечественных СМИ с позиции его практической значимости при прогнозировании динамики потребительских цен.

Для реализации программы исследования потребовалось сформировать массив эмпирических данных, провести текстовый анализ посредством дообучения (*fine-tuning*) предобученных нейронных сетей наряду с использованием классических алгоритмов *NLP*, рассчитать индекс новостного фона в публикациях российских СМИ и на этой основе осуществить моделирование инфляции.

Обзор литературы

Значительное внимание в России и мире традиционно уделяется новостям об экономической ситуации. Актуальным становится проведение исследований по оценке влияния прессы на макроэкономические показатели. Подходы к анализу и обработке текста в большинстве случаев имеют схожий характер. Поэтому область *NLP* в условиях развития технологий вызывает повышенный интерес.

Так, К. В. Яковлева (2018: 29) для оценки динамики экономической активности в России предложила методику расчета высокочастотного индикатора, использовав более 59 тыс. статей ряда крупных федеральных СМИ. Все тексты прошли предварительную предобработку: токенизицию и лемматизацию[3]. Семантиче-

ское моделирование осуществлено с использованием метода опорных векторов (*Support Vector Machine, SVM*). Тренировочная выборка размечена авторами вручную.

Аналогичный подход при расчете новостного индекса применяет и Банк России (Селезнев, Шибитов, Яковлева, Власов и др., 2021). Методика регулятора также основана на определении частотности вхождения слов в тексты новостей. На первом этапе используется модель, автоматически выделяющая темы, из которых по ключевым словам выделяется одна, максимально покрывающая тематику. Затем для каждой новости на основе обученной модели выявляется релевантность и тональность путем определения баланса слов с положительной и отрицательной коннотацией. Для этого применяются методы латентного распределения Дирихле (*LDA*) и модифицированный стохастический алгоритм, сглаживающий неравномерное распределение количества новостей по источникам.

А. Г. Евстигнеева и Д. А. Карпов (2023) исследовали влияние новостей на формирование инфляционных ожиданий населения. После токенизации и лемматизации слов в текстах новостей авторы применили численные метрики: *TF-IDF*[4] и *Count Vectorizer*. Тематическая классификация проведена на основе латентного распределения Дирихле (*LDA*), метода опорных векторов (*SVM*), случайного леса (*Random Forest*) и градиентного бустинга (*XGBoost*). Для обучения авторы сформировали выборку из 100 тыс. новостей по 25 источникам, представленным в равных долях. При этом для выделения негативных новостей специалисты также применили *LDA*: этот алгоритм требует на вход список экспертно выбранных якорных токенов, описывающих тему. В качестве токенов авторы использовали соответствующие варианты ответов из опроса ВЦИОМ и их синонимы.

Е. А. Федорова и Л. Е. Хрустова (2023: 2427) исследовали влияние новостей о *COVID-19*, публикуемых в официальных источниках и социальных сетях, на российский фондовый рынок. Для оценки тональности авторы использовали «мешок слов»[5] с применением словаря *EcSentiThemeLex* в среде *R*. Авторами сформированы подвыборки слов, которые относятся исключительно к негативному или позитивному фону.

Коллектив авторов МГУ имени М. В. Ломоносова (Гаврилов, Иванов, Клачкова, Королев и др., 2022: 98) моделировал влияние новостных потоков на волатильность доходности ценных бумаг

на российском фондовом рынке с помощью моделей авторегрессионной гетероскедастичности, учитывающих поступающие на рынок тематические новостные потоки. Для обработки текстовой информации они также использовали «мешок слов», удалив из набора текстов стоп-слова и проведя лемматизацию. Важность слов также определялась методом *TF-IDF*.

Достоинством рассмотренных подходов является простота реализации на практике и «мягкие» требования к вычислительным мощностям компьютера. Главным же недостатком видится тот факт, что подобные модели основаны на подсчете лишь частоты упоминаний слов в массивах текстов. При этом совершенно не учитывается контекст, который может диаметрально изменить тональность текста.

В этой связи популярность в последние 5 лет получили подходы к анализу текста по контексту, что стало возможным благодаря развитию нейросетевых моделей решения задач *NLP*, а именно трансформеров[6]. Основное преимущество трансформеров заключается в их высокой эффективности в условиях параллелизации.

Большинство имеющихся моделей являются мультиязычными, т. е. работают на нескольких языках. В части из них хорошо представлен и русский язык. Это, например, *mBERT* (Devlin, Chang, Lee, Toutanova, 2019: 4177), *XLM-R* (Conneau, Wu, Zettlemoyer, Stoyanov, 2020: 6028), *RemBERT* (Chung, Fevry, Tsai, Johnson et al., 2021), *mGPT* (Shliazhko, Fenogenova, Tikhonova, Mikhailov et al., 2022). При этом в последние годы стали развиваться моноязыковые модели (Williams, Nangia, Bowman, 2017), в том числе и для русского языка, однако на сегодняшний день их немного.

Одну из первых моноязыковых моделей для русского языка разработал коллектив авторов из МФТИ[7]. Чуть позже аналогичный проект в открытом доступе разместил Сбер. Оба подхода основаны на архитектуре *BERT*[8] и обучены на текстах русскоязычной *Wikipedia*, новостных статьях ряда российских СМИ, массиве книг, текстов социальных сетей и др.

Важно отметить, что в ходе исследования способов анализа русскоязычного текста выявлена существенная проблема – сложность формирования размеченной обучающей выборки. В интернете уже присутствуют размеченные наборы текстов, например корпус новостей из «Лента.ру», размещенный на портале *Kaggle*[9].

Однако применимы они преимущественно для задач, связанных с определением тематики текста, поскольку оценивают новости в целом, разделяя их на «позитивные», «негативные» или «нейтральные». В случае с анализом новостей об инфляции и ценах такой подход не оправдан, так как наличие слов «снижение»/«уменьшение» обычно оценивается негативно, тогда как в случае с инфляцией – это, скорее, позитивная тональность.

При этом тональность СМИ практически не используется при моделировании инфляционных процессов. Для прогнозирования годовых темпов инфляции инструментами *Machine learning* У. Джункеев (2024: 59) использовал модели градиентного бустинга, решающих деревьев, сверточную и рекурентную нейронные сети (*CNN, RNN*). В качестве факторов автор использовал уровень безработицы, номинальную заработную плату, объем платных услуг населению, оборот розничной торговли, объемы кредитования частного сектора, денежные агрегаты, доходность гособлигаций. Тренировочная выборка включала данные за период с января 2012 по июль 2019 г., тестовая – с августа 2019 по апрель 2023 г. Аналогичного подхода придерживались Е. Павлов (2020: 64) и И. Байбуза (2018: 50).

Для целей настоящего исследования мы посчитали нецелесообразным применение нейронных сетей ввиду недостаточного количества точек для качественного обучения модели, даже на нескольких эпохах. На основе исследования (Choi, Bahadori, Schuetz, Stewart et al., 2016: 306) можно сделать вывод, что для достижения хорошей производительности *RNN* требуется минимум 1000–2000 наблюдений: их число зависит от сложности задачи и специфики данных; на практике зачастую выстраиваются более длинные ряды – от 5000 наблюдений.

В этой связи в настоящем исследовании предпринята попытка применения современных алгоритмов обработки и классификации естественного языка на основе *fine-tuning* моделей трансформеров с дообучением на самостоятельно размеченном наборе данных в сравнении с «классическими» подходами *NLP*. Размеченные новости использованы при прогнозировании индекса потребительских цен (ИПЦ) классическими подходами (линейными методами) и методами машинного обучения (градиентным бустингом и решающими деревьями), набирающими популярность в последние годы.

Методика

С методологических позиций исследование предполагало решение двух блоков задач: классификацию новостей и расчет индекса инфляционного фона. Для этого нами были реализованы следующие этапы:

1. Сбор и фильтрация новостей, соотносящихся с темой инфляции/цен.

После формирования коллекции текстов в выгруженном массиве информации необходимо выделить новости о ценах и инфляционных процессах. В рамках настоящего исследования использован готовый алгоритм классификации, основанный на применении обученной языковой модели, не требующий дополнительной разметки данных для обучения и существенно сокращающий временные и трудовые ресурсы. Модель под названием *mDeBERTa-v3-base-xnli-multilingual-nli-2mil7* представила компания *Microsoft* (He, Gao, Chen, 2021). Эта мультиязычная модель позволяет в автоматическом режиме группировать входной текст по сформированным пользователями категориям. Обучающая выборка содержит более 2,7 млн пар гипотез-предпосылок на 27 языках.

2. Формирование обучающей выборки.

Оценка новостей с «дезинфляционной» и «проинфляционной» точек зрения является довольно специфической задачей. В исследовательской среде отсутствуют размеченные текстовые данные, классифицированные с позиции изменения цен. В этой связи нами проведена собственная разметка данных для последующего моделирования[10]. Для этих целей использована нейронная сеть *GigaChat*.

3. Семантическое моделирование «классическими» подходами.

Использованные на данном этапе методы представлены в таблице 1.

Таблица 1

«Классические» методические подходы к семантическому моделированию

Метод	Описание
Логистическая регрессия	$p = \frac{1}{1+e^{-z}}$ (1) где p – вероятность выбора класса объекта; z – стандартное уравнение регрессии.
Наивные методы Байеса (Naive Bayes, GaussianNB)[11]	$P(y\|x_i) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$ (2) p – вероятность выбора класса объекта; x – переменные модели; Параметры σ_y и μ_y оцениваются с использованием максимального правдоподобия.
Метод случайного леса (Random Forest)	Предусматривает построение модели N деревьев, в которой для каждого n = 1, ..., N можно сгенерировать выборку X_n с помощью бутстрэпа и построить решающее дерево b_n по выборке X_n. Дерево строится до тех пор, пока в каждом листе не более nmin объектов или пока не достигается определенная высота дерева. Итоговый классификатор $a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x)$ позволяет выбрать решение «голосованием по большинству».
Метод градиентного бустинга CatBoost (Gradient Boosting (Machine), GBM)[12]	Проводится инициализация GBM константным значением: $\hat{f}(x) = \widehat{f_0}, \widehat{f_0} = \gamma, \gamma \in R$: (3) $\widehat{f_0} = argmin \sum_{i=1}^{n} L(y_i, \gamma)$ x – переменные модели; y – класс объекта; После чего для каждой итерации $t = 1, ..., M$ повторяется ряд действий: 1) определение псевдо-остатков r_t $r_{it} = -\left[\frac{\delta L(y_i, f(x_i))}{\delta f(x_i)}\right]_{f(x)=\hat{f}(x)}, i = 1, ..., n$ (4) 2) построение нового базового алгоритма $h_t(x)$ на псевдо-остатках $\{(x_i, r_{it})\}_{i=1,...,n}$; 3) поиск оптимального коэффициента ρ_t при $h_t(x)$ относительно исходной функции потерь: $\rho_t = argmin \sum_{i=1}^{n} L(y_i, \hat{f}(x) + \rho * h(x_i, \theta))$ (5) p - вероятность выбора класса объекта; $h(x_i, \Theta)$ – гиперапарметры модели, например глубина дерева и т. д.; L – функция потерь; 4) сохранение $\widehat{f_t}(x) = \rho_{t*} h_t(x)$. После этого формируется итоговая GBM-модель $\hat{f}(x)$: $\hat{f}(x) = \sum_{i=0}^{M} \widehat{f_i}(x)$ (6)

4. *Fine-tuning* предобученных моделей.

В этом разделе рассмотрена предобученная модель *BERT*[13] – большая языковая модель, разработанная коллективом авторов из *Google* (Devlin, Chang, Lee, Toutanova, 2019: 4178). Она предназначена для задач обработки естественного языка. Данная архитектура лежит в основе русскоязычных моделей Сбера и МФТИ.

Реализацию *BERT* разработчики представили в два шага: *pre-training* и *fine-tuning*. Во время *pre-training* модель обучается на неразмеченных данных на задачах предварительной обработки. При этом для *fine-tuning BERT* инициализируется с предобученными параметрами. Все предобученные параметры дообучаются на основе уже размеченных данных, подготовленных специально для решения конкретной задачи *NLP*.

Модель *BERT* представляет собой двунаправленную модель *Transformer*. Но, в отличие от классических моделей такого типа, имеет только *Encoder*. Данные, подающиеся в *Encoder*, представляют собой последовательность токенов, которые сначала конвертируются в векторы, а затем обрабатываются нейронной сетью. Для векторного представления слов авторы используют эмбеддинг *WordPiece* (Wu et al., 2016) с размером словаря в 30 тыс. токенов.

Перед вводом текстовой последовательности в *BERT*, как правило, 15% слов в каждой последовательности заменяется токеном [*MASK*] (Sinha, Jia, Hupkes, Pineau et al., 2021; Sun, Zheng, Hao, Qiu, 2022). Затем модель пытается предсказать исходное значение замаскированных слов на основе контекста, предоставляемого другими, не замаскированными словами в последовательности. Этот метод позволяет проводить двунаправленное обучение в моделях, для которых ранее это было невозможно.

Идея *fine-tuning* строится на том, что матрицы весов, полученные на решении общих задач, помогают лучше решать новые, более узкие задания (Howard, Ruder, 2018: 332). Иными словами, это процесс адаптации предобученной модели под конкретную задачу. Вместо обучения с нуля, нами использована готовая модель, в которой частично или полностью заморожены слои и проведено дообучение на новых данных. *Fine-tuning* в *BERT* осуществляется путем дополнительной тренировки модели на наборе данных, размеченных для узкоспециализированной задачи. Это позволяет адаптировать модель к определенным языковым особенностям реализации проектов, например семантической окраске текста (Sun, Zheng, Hao, Qiu, 2022).

5. Формирование метрик качества моделей классификации текстов.

Оценка качества прогноза на тестовой выборке осуществляется метрикам *F1-score* и *ROC-AUC*. *F1-score* можно интерпретировать как среднее гармоническое значение точности и запоминания, при котором показатель *F1* достигает своего наилучшего значения при 1, а наихудшего – при 0. Формула для определения показателя *F1* имеет следующий вид (Christopher, Raghavan, Schütze, 2008):

$$F1_score = \frac{2*TP}{2*TP+FP+FN}, \quad (7)$$

где *TP* – истинно позитивные ответы;

FP – ошибочно позитивные ответы;

FN – ошибочно негативные ответы;

TN – истинно негативные ответы.

Идея метрики *ROC-AUC* под кривой ошибок (*Receiver Operating Characteristic curve Area Under Curve*) состоит в том, что на построенном графике каждая отмеченная точка соответствует выбору определенного порога (Fawcett, 2006: 868), а площадь[14] *ROC-AUC* показывает качество построенного алгоритма. При этом важность представляет и угол наклона кривой ошибок: она должна стремиться к максимальным значениям *TPR* и минимальным *FPR*, т. е. к точке с координатами (0, 1) [15].

С целью определения степени и направления влияния используемых в модели слов (фрагментов текста) на эндогенную переменную нами применен алгоритм Шепли (*Shapley Values*), основанный на концепции коалиционных игр (Valouet, Al-Memar, Fourie, Bobdiwala et al., 2022: 106520) и позволяющий оценить вклад каждого параметра в принятие решения:

$$\phi_i(p) = \sum_{S \subseteq N/\{i\}} \frac{|S|!(n-|S|-1)!}{n!} (p(S \cup \{i\} - p(S)) \quad (8)$$

где *S* – произвольный набор факторов без *i*-ого фактора;

n – количество факторов;

$p(S \cup \{i\})$ – предсказание модели с *i*-тым фактором;

$p(S)$ – предсказание модели без *i*-того фактора.

Классификация новостей на «проинфляционные» и «дезинфляционные»

Поскольку формирование коллекции новостных статей и ее преобразование в необходимую форму – трудоемкий процесс, с целью его автоматизации нами применена процедура *Web Scrapping*. В качестве источника новостей выбраны разделы «Экономика», «Политика», «Бизнес», «Финансы» и «Потребительский рынок» российской ежедневной общественно-политической газеты с усиленным деловым блоком «Коммерсантъ»[16]. Несмотря на то что независимость и редакционная политика СМИ время от времени вызывают дискуссии, в рамках исследования будем считать обработанные нами текстовые материалы непредвзятыми.

Следующим этапом стал отбор новостей об инфляции и ценах на промежутке с января 2014 по январь 2024 г. Для этого использована предобученная нейронная сеть *mDeBERTa-v3-base-xnli-multilingual-nli-2mil7* от *Microsoft*. Для классификации на вход модели переданы 2 класса: «инфляция и цены» и «не инфляция и цены».

В результате применения нейронной сети от *Microsoft* из 150 тыс. статей отобраны 14,3 тыс. текстов об инфляции. В итоге получена таблица из новостей, классифицируемых на «проинфляционные» (метка «0») и «дезинфляционные» (метка «1»). Отметим, что для устранения дисбаланса классов размеченные выборки с проинфляционными и дезинфляционными новостями были скорректированы до примерно одинакового размера – 1643 ед. и 1621 ед. соответственно.

Первоначально размеченная коллекция текстов разделена на тренировочную и тестовую выборки в пропорции 80% к 20%. При этом тестовая не была задействована в обучении моделей и предназначена исключительно для оценки их качества.

На первом этапе моделирования применены классические алгоритмы, для чего текст предварительно обработан: 1) токенизирован; 2) удалены стоп-слова (частицы, предлоги и т. д.); 3) лемматизирован[17]. После преобразований текст имеет следующий вид (см. табл. 2).

Таблица 2

Предобработанный текст новостных статей для применения классических моделей (фрагмент)

Оригинальный текст	Текст после предобработки
Ввод жилья по итогам января-апреля составил 36 млн кв. м, сообщил Росстат, – таким образом, показатель в годовом выражении сократился на 4,1% (за аналогичный период 2022 года было сдано в эксплуатацию 37,6 млн кв. м) ...м	ввод жилье итог январяапрель составлять 36 млн кв м сообщать росстат образ показатель годовой выражение сокращаться 41 аналогичный период 2022 сдавать эксплуатация 376 млн кв м...
Укрепление рубля способствовало сокращению цен на часть ассортимента российских продовольственных сетей. Так, лидер рынка X5 Retail Group (сети «Пятерочка», «Перекресток», «Карусель») отметил за год снижение стоимости на 15-20% своих товаров ...	укрепление рубль способствовать сокращение цена часть ассортимент российский продовольственный сеть лидер рынок x5 retail group сеть пятерочка перекресток карусель отмечать снижение стоимость 15 20 свой товар...
...	...

Далее предобработанные новости были оцифрованы. Для этого применен метод *TF-IDF*. Он вычисляет важность каждого слова в документе относительно количества его употреблений в данном документе и во всей коллекции текстов[18].

Все модели реализованы на языке программирования *Python* с использованием библиотек *CatBoost* и *Sklearn*. Оценка качества модели в целом, без привязки к конкретному порогу, проведена путем анализа *F1-score* и *ROC-AUC* площади под кривой ошибок. Матрицы ошибок для каждой из шести моделей представлены на рисунке 1, показатели *ROC-AUC* – на рисунке 2.

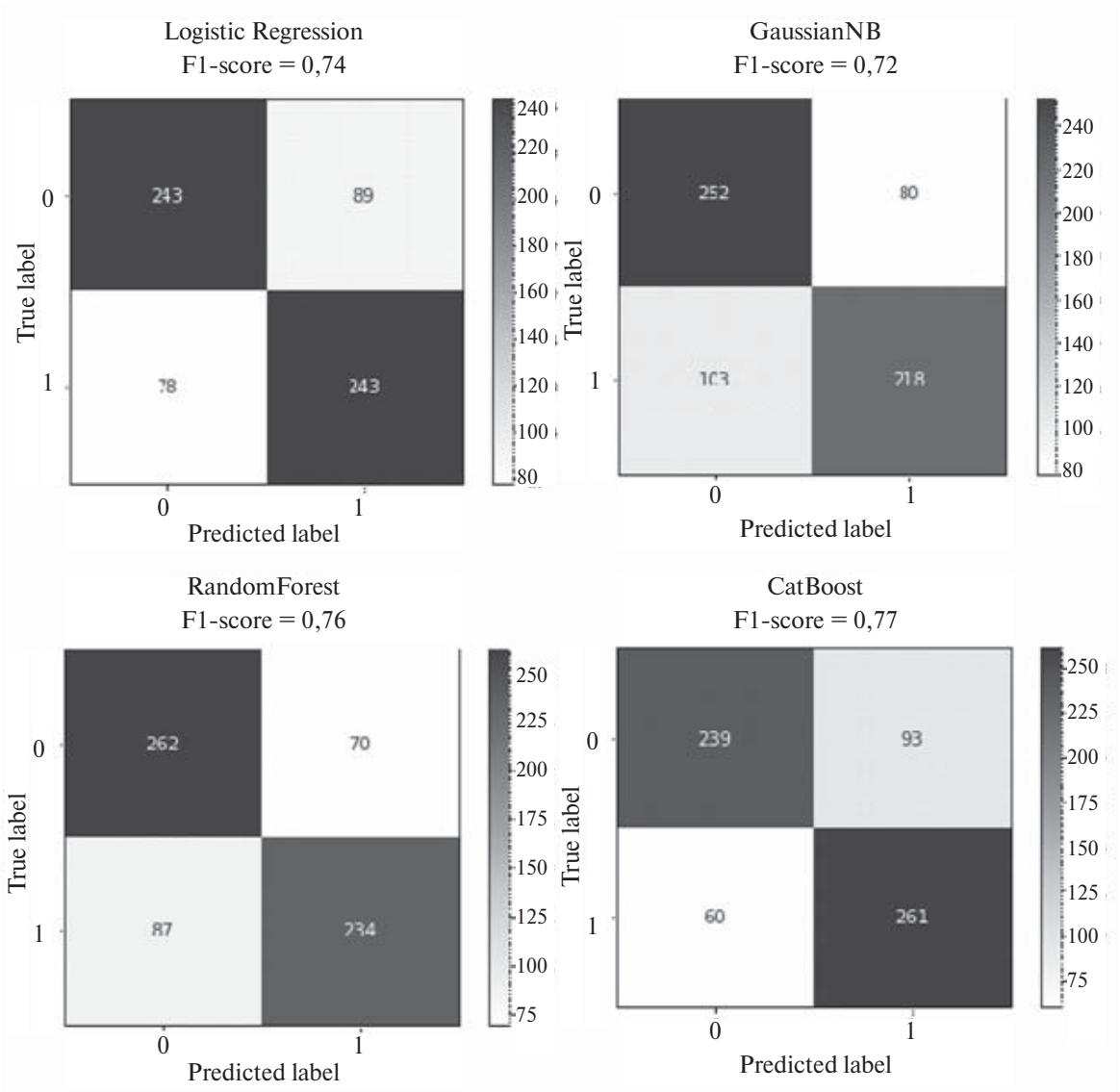


Рисунок 1. Матрицы ошибок по результатам тестирования классических моделей классификации текстов[19]

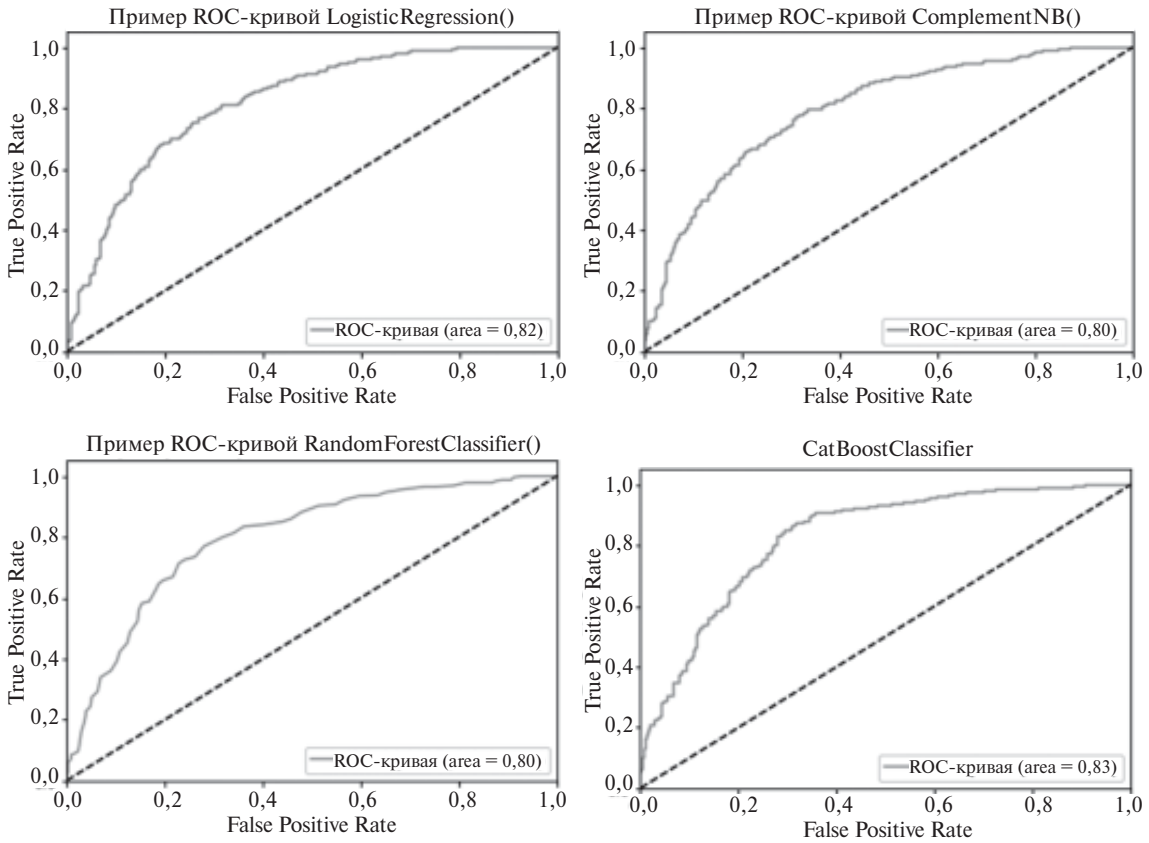


Рисунок 2. Кривые ошибок ROC классических моделей[20]

Так, согласно проведенным расчетам, наилучшие результаты из классических моделей показывает *CatBoost* с *F1-score,* равным 0,77, *ROC-AUC* – 0,83. На рисунке 3 представлены слова, которые увеличивают вероятность классификации новости в категории «проинфляционные» и «дезинфляционные» для лучшей модели из классического набора. Для этого нами применен алгоритм *Shapley Value*.

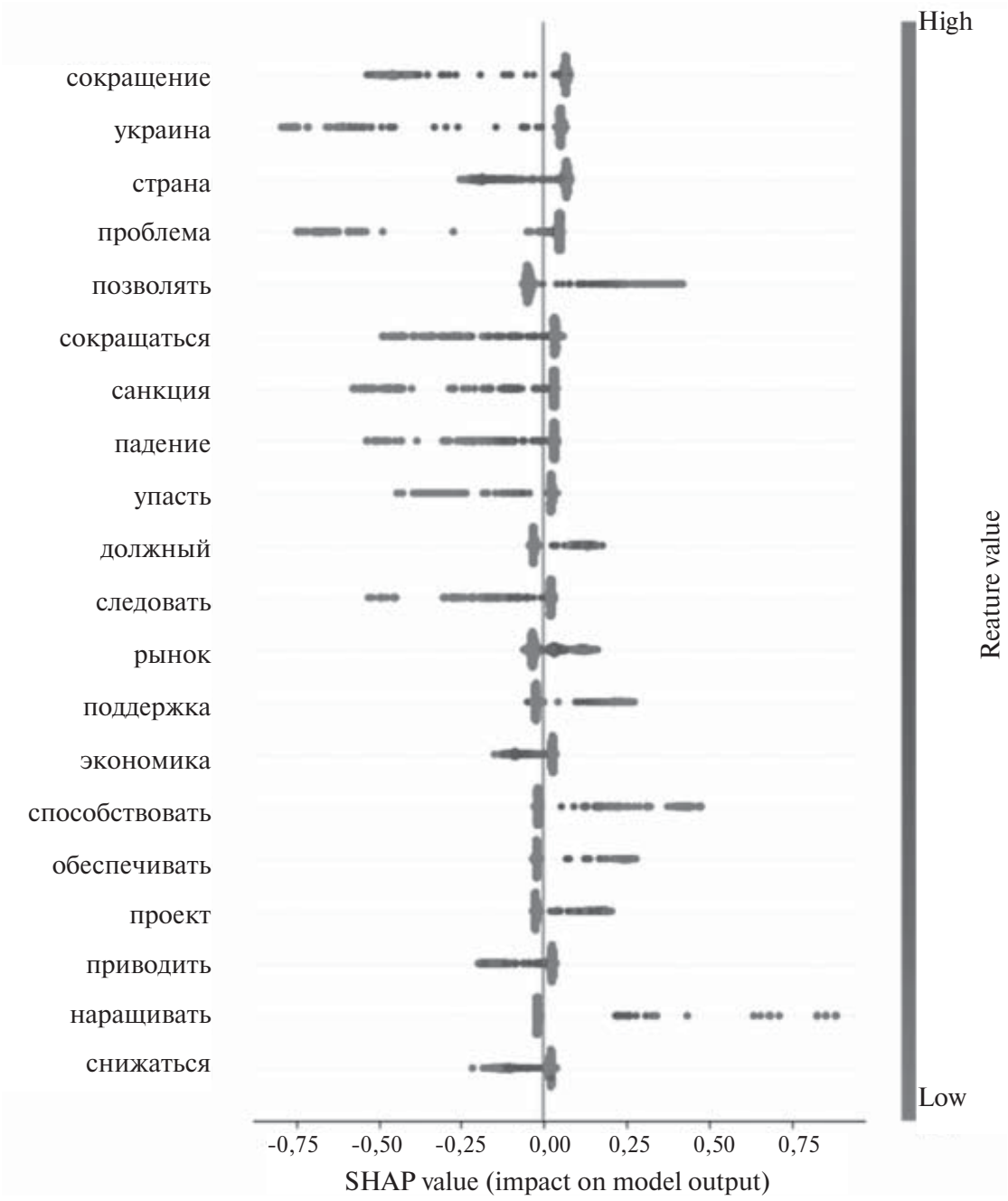


Рисунок 3. Визуализация значимых факторов моделей по значению Shapley Value

Так, текст, содержащий слова «сокращение», «Украина», «страна», «проблема», «сокращаться», «санкция», «падение», «упасть», «следовать», «экономика», «проводить», «снижаться», увеличивают вероятность того, что он является проинфляционным. При этом такие слова, как «позволять», «должный», «поддержка», «обеспечивать», «способствовать», «проект», «рынок», «наращивать», больше свойственны дезинфляционным текстам.

Как говорилось ранее, в исследовании дообучены 2 модели, предобученные на русскоязычных текстах: модель от МФТИ *DeepPavlov/rubert-base-cased* и модель от Сбера *ai-forever/ruBert-base*.

Все модели реализованы на языке программирования *Python* с использованием библиотек *PyTorch* и *Transformers*.

Важно отметить, что дообучаются такие модели гораздо дольше, чем обучаются классические алгоритмы. Но низкая скорость обучения сопровождается более высокими прогнозными свойствами. Так, *F1-score* в модели МФТИ составил 0,81, в модели Сбера – 0,85, что выше показателей *CatBoost* (0,77).

Как показывают результаты применения модели на тестовой выборке, более качественные прогнозы позволяет сделать модель *ai-forever/ruBert-base*. Важным отличием применения предобученных нейросетей является оценка текста по контексту, а не по отдельным словам, как в случае с классическими моделями. Алгоритм *Shap Value* также позволяет выделять части текста, по которым модель делала предсказание. Прогноз составлен относительно проинфляционных новостей (метка «0»). Выделение частей текста разными цветами позволяет отметить «проинфляционные» и «дезинфляционные» слова.На рисунке 4 представлена проинфляционная новость.

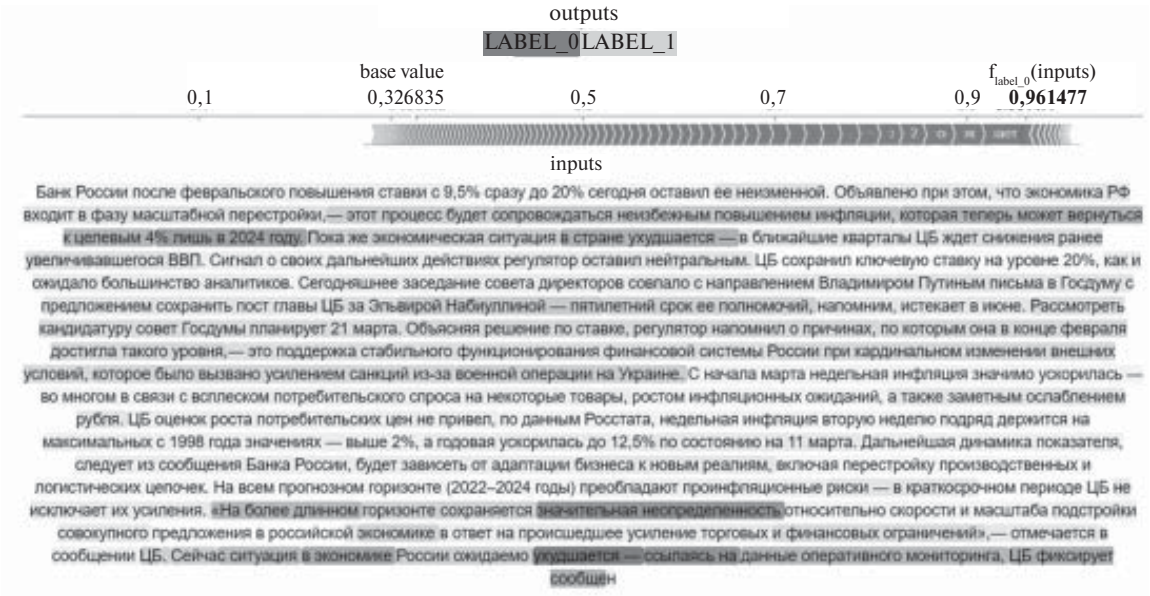


Рисунок 4. Визуализация значимых частей полностью проинфляционного текста по Shapley Value

Так, серым оттенком выделены части текста «<...> этот процесс будет сопровождаться повышением инфляции <...>» и т. д. Особенно негативно моделью оценивается слово «ухудшается».

На рисунке 5 приведена дезинфляционная новость, в которой присутствует проинфляционный контент.

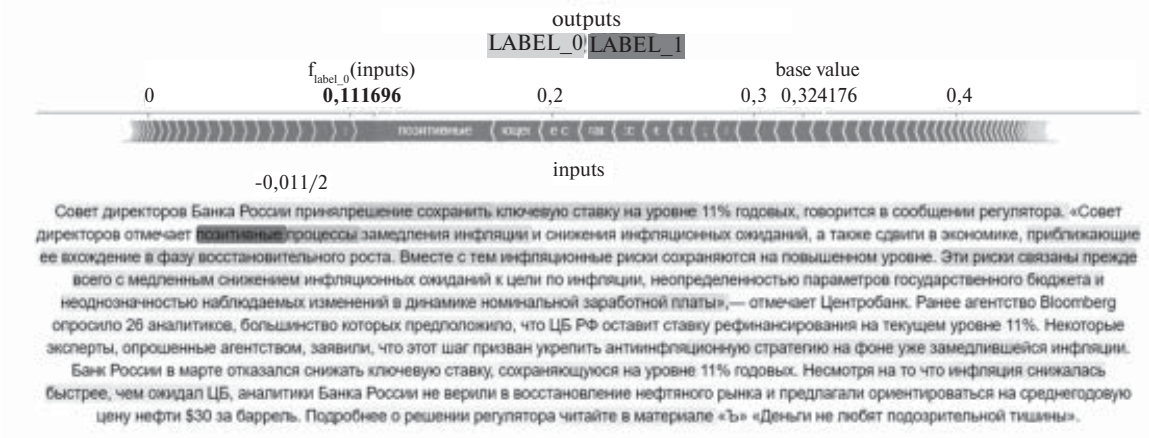


Рисунок 5. Визуализация значимых частей дезинфляционного текста с присутствием проинфляционного фрагмента по Shapley Value

Можно заметить, что текст преимущественно повествует об уменьшении инфляции. Но присутствует и негативный контент. Так, проинфляционным моментом в тексте модель определила сохранение ряда инфляционных рисков, связанных с медленным «снижением инфляционных ожиданий и неопределенностью параметров государственного бюджета».

На рисунке 6 представлен полностью дезинфляционный текст.

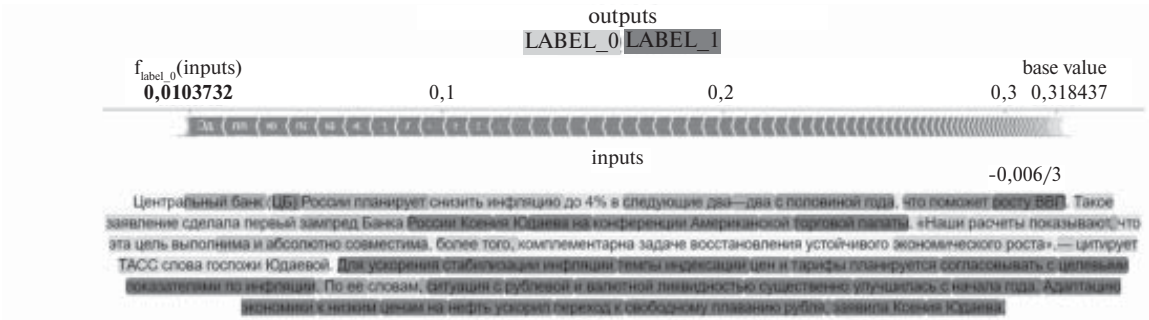


Рисунок 6. Визуализация значимых частей полностью дезинфляционного текста по Shapley Value

В данном случае весь текст окрашен серым цветом, а наиболее ключевые моменты согласно модели – темно-серым.

Расчет индекса инфляционного фона

Для расчета индекса инфляционного фона использована формула 9:

$$Inflation\ Index = \frac{positive - negative}{all_news} \quad (9)$$

где $Inflation\ Index$ – индекс инфляционного фона (далее – *Index*);
positive – количество дезинфляционных новостей;
negative – количество проинфляционных новостей;
all_news – количество всех выгруженных новостей за период.

Важно отметить, что для расчета индекса использованы только «строго проинфляционные» и «строго дезинфляционные» новости, оцененные по модели Сбера *ai-forever/ruBert-base,* т. е. те, в которых вероятность отнесения к определенному классу больше 70%. Остальные определены как нейтральные.

Все новости были сведены к месячному виду. Динамика рассчитанного индекса и индекса с учетом сезонной корректировки (*Index_SA*) представлена на рисунке 7.

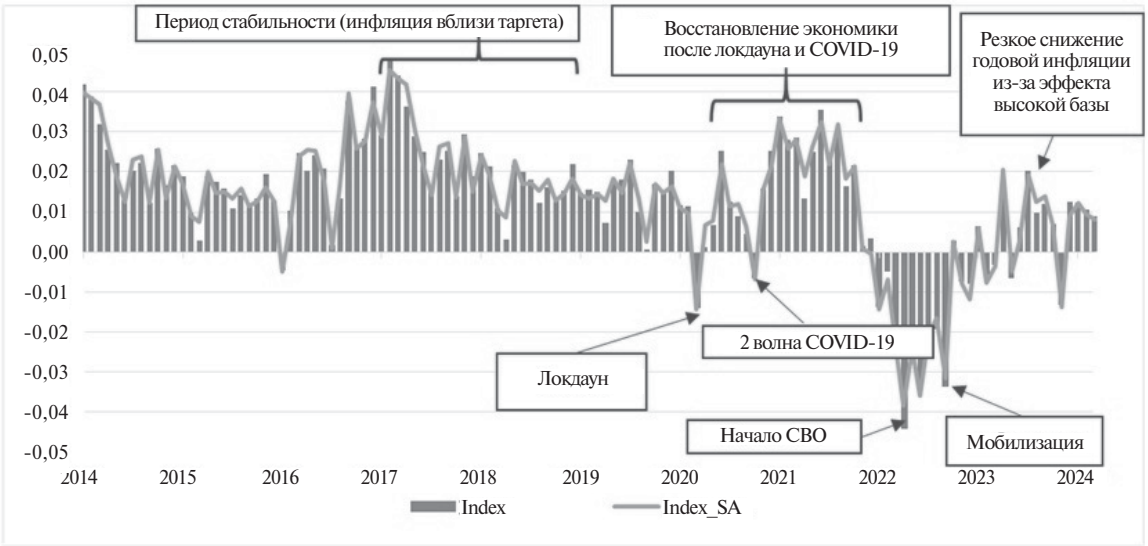


Рисунок 7. Динамика индекса инфляционного фона, 2014–2024 гг.[21]

Так, построенный индекс отражает реакцию общества на события, происходившие в тот или иной период: они указаны в прямоугольниках. Например, с 2017 по 2019 г. можно заметить его положительное значение. В эти годы инфляция наблюдалась вблизи целевого значения, определенного Банком России (4%). В период локдауна в связи с *COVID-19* новостной фон был отрицательным,

цены росли. Затем отмечаем этап восстановления экономики вплоть до начала 2022 г. и последовательный переход к эффекту высокой базы – ситуации, когда низкие темпы роста индикатора объясняются его крайне высоким стартовым показателем.

В дальнейшем исследовании использована трендовая компонента, для ее вычленения применен фильтр Ходрика-Прескотта из модуля *Statsmodels*. Она обладает меньшей волатильностью и лучше коррелирует с темпами годовой инфляции.

Уровень корреляции между трендом индекса новостного фона с лагом в 1 месяц и инфляцией в годовом выражении составляет 0,71, что свидетельствует о ее высоком уровне. Следовательно, данный показатель можно использовать для прогноза инфляционных процессов.

Результаты моделирования

Практическое применение предложенного подхода к расчету индекса инфляционного фона для прогнозирования уровня инфляции потребовало построения шести моделей: трех моделей, включающих индекс в качестве регрессора, и трех моделей без него – линейной модели множественной регрессии *Linear Regression*, градиентного бустинга *CatBoostRegressor*, модели случайного леса *Random Forest Regressor.*

Отметим, что *CatBoost* и *Random Forest* имеют ряд гиперпараметров. Они управляют процессом обучения и могут значительно влиять на качество модели. Это, например, количество итераций (деревьев), глубина деревьев, шаг обучения (скорость, с которой модель обновляет веса) и т. д. Для автоматического перебора комбинаций гиперпараметров моделей мы использовали *GridSearchCV* из модуля *SkLearn* с кросс-валидацией на 5 фолдах (частях выборки). В качестве метрики выбрана *MAE*.

В процессе подбора экзогенных переменных мы учитывали тот факт, что Росстат публикует данные макроэкономической статистики (индексы промышленного производства, оборот розничной торговли и др.) с задержкой минимум в 1 месяц, что существенно затрудняет их использование при моделировании инфляции. Для этого необходим лаг минимум в 2 месяца (для фактических значений), что далеко не всегда обосновано с точки зрения экономической логики. В этой связи для прогнозирования нами использованы оперативные данные, публикуемые раньше официального

ИПЦ: курс рубля по отношению к доллару с лагом в 1 месяц[22] (публикуется ежедневно Банком России); ценовые ожидания предприятий с лагом в 1 месяц (публикуются Банком России за месяц до объявления ИПЦ и содержат предвидения представителей бизнеса на несколько месяцев вперед); инфляционные ожидания населения с лагом в 1 месяц (публикуются Банком России за месяц до объявления ИПЦ и содержат предсказания населения на несколько месяцев вперед), а также предложенный нами индекс инфляционного фона с лагом в 1 месяц, который может рассчитываться ежедневно.

Выборка разделена на тренировочную и тестовую. В тренировочную включены данные за период с января 2014 по сентябрь 2022 г., в тестовую – с октября 2022 по март 2024 г.

Затем описывающие переменные в каждой выборке распределены в две группы: категориальные (номинативные) и непрерывные, после чего проведена их стандартизация – масштабирование путем преобразования статистического распределения в формат со средним значением, равным нулю, и стандартным отклонением, равным единице по формуле:

$$z = \frac{x-u}{s} \tag{10}$$

где z – стандартизированное значение;
x – текущее значение показателя;
u – среднее значение выборки;
s – стандартное отклонение.

В качестве метрик использованы *RMSE* (*Root Mean Squared Error*) и *MAE* (*Mean absolute error*), рассчитанные по формулам 11 и 12:

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \widehat{y_k})^2} \tag{11}$$

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|y_k - \widehat{y_k}|^2 \tag{12}$$

где n – количество точек;
y_k – фактическое значение ИПЦ;
$\widehat{y_k}$ – прогнозное значение ИПЦ.

Целью прогноза для использования его результатов Банком России является получение предсказательных данных на горизонте в 1 месяц[23]. В исследовании мы представили тестовую выборку, увеличенную до 18 месяцев. Выбор периода обусловлен включением в него месяцев с высоким уровнем волатильности инфля-

ции, что позволяет оценить работоспособность алгоритмов моделирования на «шоковых» периодах. Уравнение регрессии для индекса потребительских цен в модели линейной множественной регрессии выглядит следующим образом:

$$ИПЦ = 5{,}68 - 1{,}31 * Sentiment_index + 2{,}62 * Price_expectation + 1{,}55 * Inflation_expectations + 1{,}72 * Doll \quad (13)$$

где *ИПЦ* – индекс потребительских цен;

Sentiment_index – индекс инфляционного фона СМИ;

Price_expectations – индекс ценовых ожиданий бизнеса;

Inflation_expectations – индекс инфляционных ожиданий населения;

Doll – курс рубля к доллару.

На рисунке 8 приведены результаты прогнозирования на горизонте в 18 месяцев.

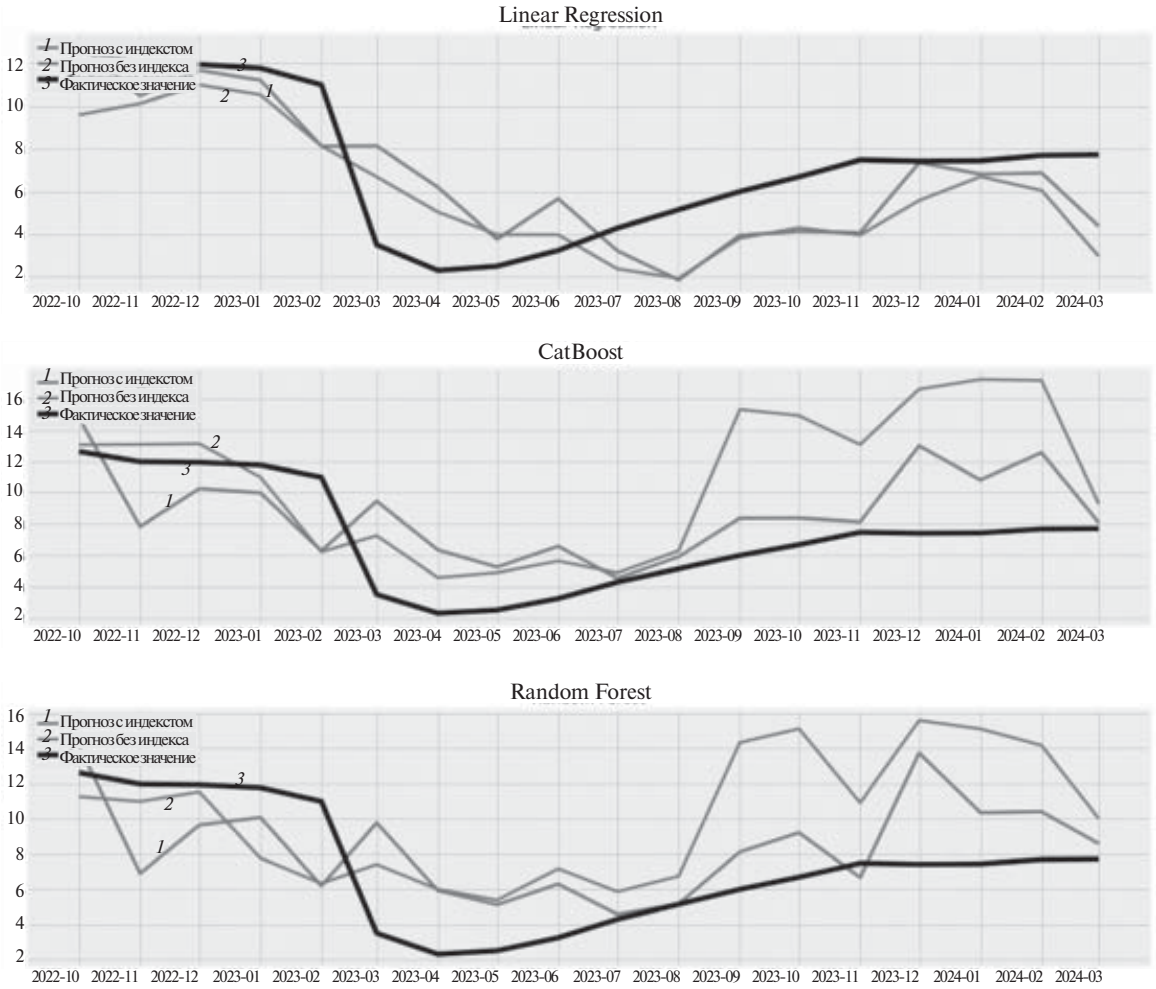


Рисунок 8. Результаты использования индекса инфляционного фона для прогнозирования ИПЦ

В целом проведенный эксперимент позволяет сделать вывод, что добавление тренда индекса инфляционного фона в многофакторные модели способствует повышению их прогнозных свойств. Об этом свидетельствуют полученные метрики качества моделей на тестовой выборке (таблица 3). Так, добавление рассчитанного индекса позволило снизить как *MAE*, так и *RMSE* во всех тестируемых моделях в среднем на 1,23 и 0,98 соответственно. Тест Диболда-Мариано подтверждает, что разница в ошибках прогнозирования существенна, т. е. модели с добавлением нового показателя формируют лучший прогноз. Полученные результаты могут лечь в основу совершенствования методов прогнозирования макроэкономических показателей.

Таблица 3

Результаты моделирования индекса потребительских цен

Модель	**Без индекса инфляционного фона**		**С индексом инфляционного фона**		**Тест Диболда-Мариано (p-value)**
	RMSE	**MAE**	**RMSE**	**MAE**	
Linear Regression	2,466	2,236	2,371	1,935	0,018
CatBoost	5,359	4,111	3,303	2,805	0,005
Random Forest	4,851	4,101	3,302	2,754	0,004

В качестве возможных направлений дальнейшего совершенствования проведенного исследования в практических целях видится расширение перечня новостных источников через использование *RSS*-каналов[24]. Они стабильно обновляются крупнейшими федеральными СМИ и содержат сводку новостей за 3–7 последних дней в удобном для парсинга формате *XML*, но часто перезаписываются (раз в 3–7 дней). Данный подход позволит охватывать более широкую новостную ленту и точнее оценивать инфляционный фон в медийном пространстве. Также перспективным видится добавление комментариев пользователей из социальных сетей *Telegram*, ВКонтакте, Яндекс.Дзен и др. Однако для реализации такого подхода необходимо решить глобальную задачу – фильтрацию и отсеивание токсичных текстов (спама), чего нет в сообщениях официальных СМИ.

Заключение

В работе представлена методика расчета индекса инфляционного фона, информация о котором взята из российских СМИ. Проведенные расчеты показали, что автоматизированная оценка текстовой информации новостных агентств является важным аналитическим инструментом и может быть использована в процессе моделирования не только уровня инфляции, но и других макроэкономических показателей.

Реализованные подходы позволяют в автоматическом режиме формировать коллекцию новостных статей, отбирать тексты, тематически относящиеся к инфляции и ценам, классифицировать их на дезинфляционные и проинфляционные. Для этих целей предложено использование предобученных моделей-трансформеров. Их апробация показала значительное превосходство с точки зрения точности в сравнении с «классическими» алгоритмами.

Построенный индекс инфляционного фона с высокой долей вероятности сигнализирует о значимых «шоках» в экономике страны. Кроме того, его включение в модели прогнозирования инфляции способствовало улучшению прогнозных свойств каждой из них по двум метрикам качества (*RMSE* и *MAE*). Это позволяет сделать вывод о том, что новостные данные обладают достаточной прогнозной силой. Их учет позволит в перспективе оперативно реагировать на текущую экономическую ситуацию и принимать соответствующие решения.

Примечания

[1] Настоящая статья отражает личную позицию авторов. Содержание и результаты данного исследования не следует рассматривать, в том числе цитировать, в каких-либо изданиях, как официальную позицию Банка России или указание на официальную политику или решение регулятора.

[2] Ваш путеводитель по миру *NLP* (обработке естественного языка) // Хабр. 2023. Режим доступа: https://habr.com/ru/companies/otus/articles/705482/ (дата обращения: 03.03.2024).

[3] Токенизация – процесс разбиения текста на более мелкие части, такие как слова или предложения. Лемматизация – процесс приведения словоформы к лемме – ее нормальной (словарной) форме.

[4] *TF-IDF* – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса.

[5] «Мешок слов» – это упрощенное представление текста, которое используется в обработке естественных языков и информационном поиске. В этой модели текст (одно предложение или весь документ) представляется в виде мешка (муль-

тимножества) его слов без какого-либо учета грамматики и порядка слов, но с сохранением информации об их количестве. Мешок слов обычно используется в методах классификации документов, где частотность вхождения слова используется как признак для обучения классификатора.

[6] Трансформер – архитектура глубоких нейронных сетей, основанная на механизме внимания (*Attention*). Впервые модель трансформера была предложена в статье *Attention is All You Need* от разработчиков *Google* в 2017 г. (Vaswani, Shazeer, Parmar, Uszkoreit et al., 2017)

[7] К*уратов Ю. М., Игнатов Ф. В., Корнев Д. П., Чепурова А. А.* Прорывы #DeepPavlov в 2019 году: обзор и итоги года // Хабр. 2019. Режим доступа: https://habr.com/ru/companies/mipt/articles/488818/ (дата обращения: 03.03.2024).

[8] *Аверкиев С.* Это не чат, это GigaChat. Русскоязычная ChatGTP от Сбера // Хабр. 2023. Режим доступа: https://habr.com/ru/companies/sberbank/articles/730108/

[9] *Амиргалиев Б., Тен Х.* Анализ настроений на русском языке // Кэггл. 2017. Режим доступа: https://kaggle.com/competitions/sentiment-анализ-in-russian (дата обращения: 03.03.2024).

[10] В крупных *IT*-компаниях такую задачу решает отдельно выделенная команда специалистов. Услуги разметки данных также оказывают сторонние организации. Но эти способы не подходят для текущего исследования ввиду высокой стоимости или отсутствия необходимого числа разметчиков данных.

[11] Gaussian Naive Bayes (GaussianNB). *Scikit-learn*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html (accessed: 03.03.2024).

[12] *Натекин А.* Открытый курс машинного обучения. Градиентный бустинг // Хабр. 2017. Режим доступа: https://habr.com/ru/companies/ods/articles/327250/ (дата обращения: 03.03.2024).

[13] *BERT – Bidirection Encoder Representations from Transformers* (с англ. – двунаправленная нейронная сеть-кодировщик).

[14] Линии в диапазоне от (0,0) до (1,1) в координатах *True Positive Rate* (*TPR*) и *False Positive Rate* (*FPR*).

[15] *Лабынцев Е.* Метрики в задачах машинного обучения // Хабр. 2017. Режим доступа: https://habr.com/ru/companies/ods/articles/328372/ (дата обращения: 03.03.2024).

[16] Предварительно мы изучили посещаемость федеральных новостных порталов (Лента, РИА Новости, Коммерсантъ, РБК, Известия, Интерфакс, Газета.ру и другие), рассмотрели их контент. Многие новости в них дублируются, новостные страницы имеют примерно одинаковое количество посещений. В этой связи мы остановили свой выбор на издании «Коммерсантъ», которое, помимо прочих равных условий, удобно для парсинга *html*-кода веб страниц.

[17] При использовании «классических» алгоритмов для лемматизации текста использован модуль *nltk* (*from nltk.stem import WordNetLemmatizer*). Для моделей-трансформеров Сбер и МФТИ лемматизация не требовалась.

[18] Извлечение признаков из текстовых данных с использованием *TF-IDF* // Хабр. Режим доступа: https://habr.com/ru/companies/otus/articles/755772/ (дата обращения: 03.03.2024).

[19] *True label* – верные значения классов. *Predicted label* – предсказанные значения классов.

[20] *True Positive Rate* – истинно положительные предсказания модели. *False Positive Rate* – ложно положительные предсказания модели.

[21] *SA* (*seasonal adjustment*) – сезонно-сглаженный ряд.

[22] Лаги были определены на основе теста Грейнджера на причинность.

[23] Исходя из оптимальных лагов по тесту Грейнджера на причинность.

[24] *RSS* (англ. *Rich Site Summary* – обогащенная сводка сайта) – семейство *XML*-форматов, предназначенных для описания лент новостей, анонсов статей, изменений в блогах и т. п.

Библиография

Байбуза И. Прогнозирование инфляции с помощью методов машинного обучения // Деньги и кредит. 2018. № 4. С. 42–59.

Гаврилов В., Иванов М. А., Клачкова О. А., Королев В. Ю. и др. Влияние тематических новостных потоков на компоненты волатильности фондового рынка России // Вестник Института экономики Российской академии наук. 2022. № 2. С. 93–109.

Джункеев У. Прогнозирование инфляции в России на основе градиентного бустинга и нейронных сетей // Деньги и кредит. 2024. № 1. С. 53–76.

Евстигнеева А. Г., Карпов Д. А. Влияние негативных новостей на восприятие инфляции населением // Доклады об экономических исследованиях. 2023. № 111. Режим доступа: https://cbr.ru/Content/Document/File/144918/wp_111.pdf (дата обращения: 03.03.2024).

Павлов Е. Прогнозирование инфляции в России с помощью нейронных сетей // Деньги и кредит. 2020. №1. С. 57–73.

Селезнев С., Шибитов Д., Яковлева К., Власов С. и др. Новая методология новостного индекса // Доклады об экономических исследованиях. 2021. Режим доступа: https://cbr.ru/Collection/Collection/File/35503/index_2107.pdf (дата обращения: 10.03.2024).

Федорова Е. А., Хрустова Л. Е. Влияние новостей о COVID-19 на динамику российского фондового рынка // Финансы и кредит. 2023. Т. 29. № 11. С. 2412–2438.

Яковлева К. В. Оценка экономической активности на основе текстового анализа // Деньги и кредит. 2018. Т. 77. № 4. С. 26–41.

Choi E., Bahadori M. T., Schuetz A., Stewart W. F., Sun J. (2016) Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR Workshop and Conference Proceedings* 56: 301–318. Available at: https://doi.org/10.48550/arXiv.1511.05942

Christopher D. M., Raghavan P., Schütze H. (2008) *Introduction to information retrieval*. Cambridge: Cambridge University Press. 482 p.

Chung H. W., Fevry T., Tsai H., Johnson M., Ruder S. (2021) Rethinking embedding coupling in pre-trained language models. *International Conference on Learning Representations*. Available at: https://arxiv.org/abs/2010.12821 (accessed: 10.03.2024).

Conneau A., Wu S., Li H., Zettlemoyer L., Stoyanov V. (2020) Emerging Cross-lingual Structure in Pretrained Language Models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Pp. 6022–6034. DOI: 10.18653/v1/2020.acl-main.536

Devlin J., Chang M., Lee K., Toutanova K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Volume 1 (Long and Short Papers). P. 4171–4186. DOI: 10.18653/v1/N19-1423.

Fawcett T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8): 861–874.

He P., Gao J., Chen W. (2021) DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *Arxiv*. Available at: https://arxiv.org/pdf/2111.09543.pdf (accessed: 17.03.2024).

Howard J., Ruder S. (2018) Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics*. Pp. 328–339. DOI: 10.18653/v1/P18-1031.

Shliazhko O., Fenogenova A., Tikhonova M., Mikhailov V., Kozlova A., Shavrina T. (2022) mGPT: Few-Shot Learners Go Multilingual. *Arxiv*. Available at: https://arxiv.org/pdf/2204.07580.pdf (accessed: 17.03.2024).

Sinha K., Jia R., Hupkes D., Pineau J., Williams A., Kiela D. (2021) Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. Available at: https://arxiv.org/pdf/2109.03564.pdf (accessed: 17.03.2024).

Sun Y., Zheng Y., Hao C., Qiu H. (2022) NSP-BERT: A Prompt-based Few-Shot Learner Through an Original Pre-training Task – Next Sentence Prediction. *Arxiv*. Available at: https://arxiv.org/pdf/2104.06644.pdf (accessed: 17.03.2024).

Valouet T., Al-Memar M., Fourie H., Bobdiwala S., Saso S., Pipi M., Stalder C., Bennett P., Timmerman D., Bourte T. and De Moor B. (2022) Gradient boosted trees with individual explanations: An alternative to logistic regression for predicting viability in the first trimester of pregnancy. *Computer Methods and Programs in Biomedicine* 213: 106520.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. (2017) Attention Is All You Need. *Arxiv*. Available at: https://arxiv.org/abs/1706.03762 (accessed: 17.03.2024).

Williams A., Nangia N., Bowman S. R. (2017) A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *Arxiv*. Available at: https://doi.org/10.48550/arXiv.1704.05426 (accessed: 10.03.2024).

Wu Y., Schuster M., Chen Zh., Le Q. V., Norouzi M., Macherey W., Krikun M., Cao Yu., Gao Q., Macherey K., Klingner J., Shah A., Johnson M., Liu X., Kaiser L., Gouws S., Kato Y., Kudo T., Kazawa H., Stevens K., Kurian G., Patil N., Wang W., Young C., Smith J., Riesa J., Rudnick A., Vinyals O., Corrado G., Hughes M., Dean J. (2016) Google's Neural Machine Translation System: Bridging the Gap between Human and Machine. *Arxiv*. Available at: https://arxiv.org/abs/1609.08144 (accessed: 14.03.2024).

Notes

Vash putevoditel' po miru NLP (obrabotke estestvennogo yazyka) [Your NLP (Natural Language Processing) World Guide] Habr. 2023. Available at: https://habr.com/ru/companies/otus/articles/705482/ (accessed: 03.03.2024).

Kuratov Yu. M., Ignatov F. V., Kornev D. P., Chepurova A. A. (2019) Proryvy #DeepPavlov v 2019 godu: obzor i itogi goda [# DeepPavlov breakthroughs in 2019: review and results of the year]. Habr. Available at: https://habr.com/ru/companies/mipt/articles/488818/ (accessed: 03.03.2024).

Averkiev S. (2023) Eto ne chat, eto GigaChat. Russkoyazychnaya ChatGTP ot Sbera [This is not a chat, this is GigaChat. Russian-language ChatGTP from Sberbank] Habr. Available at: https://habr.com/ru/companies/sberbank/articles/730108/ (accessed: 03.03.2024).

Amirgaliev B., Ten H. (2017) Analiz nastroenij na russkom yazyke [Mood analysis in Russian] Keggle. Available at: https://kaggle.com/competitions/sentiment-анализ-in-russian (accessed: 03.03.2024).

Natekin A. Otkrytyj kurs mashinnogo obucheniya. Gradientnyj busting [Open machine learning course. Gradient boosting] (2017). Khabr. Available at: https://habr.com/ru/companies/ods/articles/327250/ (accessed: 03.03.2024).

Labyncev E. (2017) Metriki v zadachah mashinnogo obucheniya [Metrics in machine learning problems] Khabr. Available at: https://habr.com/ru/companies/ods/articles/328372/ (accessed: 03.03.2024).

Izvlechenie priznakov iz tekstovyh dannyh s ispol'zovaniem TF-IDF [Extract characteristics from text data using TF-IDF]. Habr. Available at: https://habr.com/ru/companies/otus/articles/755772/ (accessed: 03.03.2024).

References

Baybuza I. (2018) Prognozirovanie inflyatsii s pomoshch'yu metodov mashinnogo obucheniya [Inflation Forecasting Using Machine Learning Methods]. *Russian Journal of Money and Finance* 77(4): 42–59. (In Russian)

Choi E., Bahadori M. T., Schuetz A., Stewart W. F., Sun J. (2016) Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR Workshop and Conference Proceedings* 56: 301–318. Available at: https://doi.org/10.48550/arXiv.1511.05942

Christopher D. M., Raghavan P., Schütze H. (2008) *Introduction to information retrieval*. Cambridge: Cambridge University Press. 482 p.

Chung H. W., Fevry T., Tsai H., Johnson M., Ruder S. (2021) Rethinking embedding coupling in pre-trained language models. *International Conference on Learning Representations*. Available at: https://arxiv.org/abs/2010.12821 (accessed: 10.03.2024).

Conneau A., Wu S., Li H., Zettlemoyer L., Stoyanov V. (2020) Emerging Cross-lingual Structure in Pretrained Language Models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Pp. 6022–6034. DOI: 10.18653/v1/2020.acl-main.536

Devlin J., Chang M., Lee K., Toutanova K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Volume 1 (Long and Short Papers). P. 4171–4186. DOI: 10.18653/v1/N19-1423.

Dzhunkeev U. (2024) Prognozirovanie inflyatsii v Rossii na osnove gradientnogo bustinga i nejronnykh setey [Forecasting Inflation in Russia Using Gradient Boosting and Neural Networks]. *Russian Journal of Money and Finance* 83 (1): 53–76. (In Russian)

Evstigneeva A., Karpov D. (2023) Vliyanie negativnykh novostey na vospriyatie inflyatsii naseleniem [Impact of negative news on households' perception of inflation]. *Bank of Russia Working Papers* 111. Available at: https://cbr.ru/Content/Document/File/144918/wp_111.pdf (accessed: 03.03.2024). (In Russian)

Fawcett T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8): 861–874.

Fedorova E. A., Hrustova L. E. (2023) Vliyanie novostey o COVID-19 na dinamiku rossiyskogo fondovogo rynka [Impact of news about COVID-19 on the dynamics of the Russian stock market]. *Finance and Credit* 29 (11): 2412–2438. (In Russian).

Gavrilov V., Ivanov M. A., Klachkova O. A., Korolev V. Yu., Roshhina Ya. A. (2022) Vliyanie tematicheskikh novostnykh potokov na komponenty volatil'nosti fondovogo rynka Rossii [Impact of thematic news streams on Russian stock market volatility components]. *The Bulletin of the Institute of Economics of the Russian Academy of Sciences* 2: 93–109. (In Russian)

He P., Gao J., Chen W. (2021) DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *Arxiv*. Available at: https://arxiv.org/pdf/2111.09543.pdf (accessed: 17.03.2024).

Howard J., Ruder S. (2018) Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics*. Pp. 328–339. DOI: 10.18653/v1/P18-1031.

Pavlov E. (2020) Prognozirovanie inflyatsii v Rossii s pomoshch'yu neyronnykh setey [Forecasting Inflation in Russia Using Neural Networks]. *Russian Journal of Money and Finance* 79 (1): 57–73. (In Russian)

Seleznev S., Shibitov D., Yakovleva K., Vlasov S., Chernjad'ev D. (2021) Novaya metodologiya novostnogo indeksa [New News Index Methodology]. *Bank of Russia Working Papers*. Available at: https://cbr.ru/Collection/Collection/File/35503/index_2107.pdf (accessed: 10.03.2024) (In Russian).

Shliazhko O., Fenogenova A., Tikhonova M., Mikhailov V., Kozlova A., Shavrina T. (2022) mGPT: Few-Shot Learners Go Multilingual. *Arxiv*. Available at: https://arxiv.org/pdf/2204.07580.pdf (accessed: 17.03.2024).

Sinha K., Jia R., Hupkes D., Pineau J., Williams A., Kiela D. (2021) Masked Language Modeling and the Distributional Hypothesis: Order Word

Matters Pre-training for Little. Available at: https://arxiv.org/pdf/2109.03564.pdf (accessed: 17.03.2024).

Sun Y., Zheng Y., Hao C., Qiu H. (2022) NSP-BERT: A Prompt-based Few-Shot Learner Through an Original Pre-training Task – Next Sentence Prediction. *Arxiv*. Available at: https://arxiv.org/pdf/2104.06644.pdf (accessed: 17.03.2024).

Valouet T., Al-Memar M., Fourie H., Bobdiwala S., Saso S., Pipi M., Stalder C., Bennett P., Timmerman D., Bourte T. and De Moor B. (2022) Gradient boosted trees with individual explanations: An alternative to logistic regression for predicting viability in the first trimester of pregnancy. *Computer Methods and Programs in Biomedicine* 213: 106520.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. (2017) Attention Is All You Need. *Arxiv*. Available at: https://arxiv.org/abs/1706.03762 (accessed: 17.03.2024).

Williams A., Nangia N., Bowman S.R. (2017) A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *Arxiv*. Available at: https://doi.org/10.48550/arXiv.1704.05426 (accessed: 10.03.2024).

Wu Y., Schuster M., Chen Zh., Le Q. V., Norouzi M., Macherey W., Krikun M., Cao Yu., Gao Q., Macherey K., Klingner J., Shah A., Johnson M., Liu X., Kaiser L., Gouws S., Kato Y., Kudo T., Kazawa H., Stevens K., Kurian G., Patil N., Wang W., Young C., Smith J., Riesa J., Rudnick A., Vinyals O., Corrado G., Hughes M., Dean J. (2016) Google's Neural Machine Translation System: Bridging the Gap between Human and Machine. *Arxiv*. Available at: https://arxiv.org/abs/1609.08144 (accessed: 14.03.2024).

Yakovleva K. (2018) Otsenka ekonomicheskoy aktivnosti na osnove tekstovogo analiza [Assessment of economic activity based on text analysis]. *Russian Journal of Money and Finance* 77(4): 26–41. (In Russian)

Поступила в редакцию
14.01.2025