УДК: 331.5:004.6

Трансформация данных онлайн-платформ в надёжные индикаторы рынка труда: методология очистки, взвешивания и калибровки с применением NLP и Behavioural Scoring

Ахтямов Р.Э., кандидат экономических наук, digital-экономист, основатель Digital Economy Lab, Казань, Россия

Статья посвящена проблеме системных искажений в больших данных коммерческих платформ занятости (мультипликация резюме, «мёртвые души», информационная асимметрия), которые препятствуют их прямому использованию для целей макроэкономического анализа и денежно-кредитной политики. Цель работы — разработка и эмпирическая верификация комплексной методологии очистки, взвешивания и калибровки данных онлайн-платформ (hh.ru, «Работа России») для превращения их в надёжные индикаторы рынка труда. Исследование опирается на синтез методов компьютерных наук (NLP на основе эмбеддингов YandexGPT, графовые алгоритмы, поведенческое скорингование) и экономической статистики. Предложен каскадный подход к дедупликации (от точного хэширования до семантического анализа с помощью YandexGPT Embeddings), построена авторская формула взвешивания резюме с учётом поведенческих факторов и активности. Для верификации данных применяются методы триангуляции и калибровки на официальную статистику Росстата с использованием bridge-уравнений. Научная новизна заключается в целостной методологии, объединяющей точную и семантическую дедупликацию с поведенческим взвешиванием и многоуровневой триангуляцией, а также в формулировке набора тестируемых гипотез для эмпирической проверки. Результатом является структура системы оперативных индикаторов рынка труда (Индекс реального предложения труда, Индекс напряжённости, Индекс зарплатного давления, Индекс структурной эффективности), которые могут быть интегрированы в модели прогнозирования Банка России.

Введение: операционная потребность регулятора в качественных данных

В условиях структурной трансформации российской экономики, сопровождающейся рекордно низкой безработицей (2,9% в конце 2025 г.) и дефицитом кадров, традиционная статистика Росстата, публикуемая с лагом в 1–2 месяца, перестаёт быть достаточной для оперативного принятия решений. Как справедливо отметил Заместитель Председателя Банка России А. Заботкин, «сырые» данные платформ занятости содержат системные искажения, но именно

они могут стать основой для nowcasting — краткосрочного прогнозирования текущего состояния экономики [1].

Проблема усугубляется тем, что соискатели активно мультиплицируют резюме: по разным оценкам, один уникальный пользователь может создавать от 2 до 5 резюме [2, 3]. Это приводит к гипертрофированному представлению о предложении труда и искажает сигналы для регулятора. Настоящее исследование предлагает не просто теоретическую концепцию, а operational-ready методологию, прошедшую концептуальную валидацию и готовую к эмпирической проверке.

1. Современное состояние исследований: аналитический обзор и идентификация пробелов

Анализ литературы выявил три основных направления исследований, которые формируют базу для настоящей работы.

1.1. Типология и сравнительный анализ цифровых платформ занятости.
Работа О.В. Зайцевой [4] представляет систематическую классификацию онлайн-порталов вакансий по критериям: источник данных (частный/государственный), модель монетизации, глубина структурирования данных. Исследование И.Л. Сизовой [5] содержит эмпирический анализ массива данных объёмом более 5,3 млн текстов с трёх крупнейших российских платформ за 2019–2023 гг. и подтверждает наличие информационной асимметрии: соискатели указывают в среднем на 30–40% больше компетенций, чем требуется в вакансиях.

1.2. Дедупликация данных: методы и алгоритмы.
Наиболее актуальный обзор представляет статья R. Kaur с соавторами [6], предлагающая классификацию методов дедупликации на основе машинного обучения: rule-based, кластерные подходы, глубокое обучение.
Междисциплинарный взгляд предлагает обзор B. Hammer с соавторами [7], вводящий понятие Entity Resolution (ER) как процесса идентификации записей, относящихся к одной сущности. Ключевая проблема существующих подходов — отсутствие интеграции с поведенческими факторами (частота обновления, отклики).

1.3. Применение NLP в анализе рынка труда.
Работа И.Л. Сизовой [8] демонстрирует успешное применение тематического моделирования (LDA) и эмбеддингов на основе BERT для классификации текстов и создания таксономии компетенций. В настоящем исследовании для семантического анализа предлагается использовать современные коммерческие модели эмбеддингов, доступные через Yandex Cloud, в частности YandexGPT Embeddings [9], которые обеспечивают высокое качество представления русскоязычных текстов и оптимизированы для задач поиска дубликатов и кластеризации.

Идентифицированный пробел: ни одно из рассмотренных исследований не предлагает комплексной методологии, объединяющей: (1) многоуровневую дедупликацию с применением state-of-the-art NLP (на базе YandexGPT Embeddings), (2) поведенческое взвешивание для оценки "реальности" соискателя, (3) триангуляцию данных из разных источников и (4) калибровку по официальной статистике. Именно этот синтез составляет научную новизну настоящей работы.

2. Типология искажений данных платформ занятости: от симптомов к системным причинам

На основе синтеза данных из рассмотренных источников [2–5] автором выделены четыре типа систематических искажений, каждый из которых требует специфического метода коррекции.

Таблица 1. Типология искажений и методы их коррекции

Тип искажения	Описание	Количественная оценка (гипотеза)	Метод коррекции
Мультиплицикация резюме	Один пользователь создаёт несколько резюме (разные регионы, профобласти)	Коэффициент мультипликации: 2.3–2.8	Каскадная дедупликация (уровни 1–3)
Информационная асимметрия	Завышение компетенций, "косметическое" редактирование	30–40% избыточных навыков [5]	Семантический анализ, индекс соответствия
«Мёртвые души»	Резюме без обновлений >6 месяцев	15–25% от общего массива	Взвешивание по свежести (коэффициент R)
Дублирование вакансий	Одна вакансия размещена многократно	10–20% дублей	Дедупликация вакансий по названию компании и описанию

3. Методология очистки, взвешивания и скоринга

Предлагаемая методология реализуется в четыре последовательных этапа: сбор и предобработка, каскадная дедупликация, поведенческое взвешивание, агрегация и калибровка.

3.1. Каскадная дедупликация пользователей

Для идентификации резюме, принадлежащих одному физическому лицу, предлагается трёхуровневый каскадный подход, основанный на классификации методов, предложенной Kaur et al. [6].

Уровень 1: Точное совпадение (Rule-based blocking).
Хэширование контактных данных: телефон (форматированный), email, ссылки на соцсети. Создание блоков кандидатов с одинаковыми хэшами.

Уровень 2: Нечёткое совпадение (MinHash + LSH).
Для записей, не попавших в Уровень 1, применяется MinHash к текстовым полям (название резюме, ключевые навыки) с последующей кластеризацией через Locality-Sensitive Hashing. Это позволяет быстро находить квази-дубликаты без попарного сравнения всех записей.

Уровень 3: Семантическая дедупликация (YandexGPT Embeddings + иерархическая кластеризация).
Для уточнения границ кластеров применяются эмбеддинги текстов, получаемые через API YandexGPT Embeddings [9]. Для каждой пары резюме внутри одного кластера рассчитывается косинусная близость векторов. Порог схожести (threshold) подбирается эмпирически на валидационной выборке. Использование коммерческих эмбеддингов обеспечивает высокую точность для русскоязычных текстов и учитывает семантические особенности сферы труда.

Алгоритмическая реализация (псевдокод):

python

```
# Псевдокод для уровня 3 с использованием YandexGPT Embeddingsimport requestsimport numpy as npdef get_embedding(text, api_key):    # вызов API Yandex Cloud для получения эмбеддинга response = requests.post(        "https://llm.api.cloud.yandex.net/foundationModels/v1/embedding", headers={"Authorization": f"Api-Key {api_key}"},        json={"model": "yandexgpt/latest", "text": text}    ) return np.array(response.json()["embedding"])def semantic_deduplicate(resumes, api_key, threshold=0.85):    embeddings = [get_embedding(r['text'], api_key) for r in resumes]    clusters = []    # иерархическая кластеризация с порогом    # возвращает список кластеров (списков индексов резюме)    return clusters
```

3.2. Поведенческое взвешивание: формализация понятия «активный соискатель»

После объединения резюме в кластеры (предположительно, принадлежащие одному пользователю) каждому резюме присваивается вес достоверности. Вес должен отражать вероятность того, что данное резюме представляет реальное намерение соискателя найти работу.

Формула интегрального веса (Akhtyamov Weight Score — AWS):

$$W = \frac{1}{N_{user}} \cdot C \cdot R \cdot (1+B) \cdot A$$

Где:

- N_{user} — количество резюме в кластере пользователя (штраф за мультипликацию);
- C — коэффициент полноты заполнения (0..1), рассчитываемый как доля заполненных полей: (опыт, образование, навыки, зарплата);
- R — коэффициент свежести данных: $R = \min(1, \frac{d_{now} - d_{update}}{T_{active}})$, где T_{active} — порог актуальности (обычно 90 дней);
- B — поправочный коэффициент на поведенческие паттерны (например, наличие откликов за последние 30 дней увеличивает вес на 20%, т.е. $B=0.2$);
- A — коэффициент активности на платформе (просмотр вакансий, добавление в избранное), если доступны метаданные.

Веса нормируются внутри кластера так, чтобы $\sum W_i = 1$, и умножаются на глобальный поправочный коэффициент для калибровки по официальной статистике.

3.3. Особый фокус: очистка вакансий

Вакансии рассматриваются как более надёжный первичный индикатор спроса на труд. Методы очистки вакансий включают:

1. Дедупликацию по названию компании, описанию и контактам (MinHash + LSH).
2. Нормализацию зарплатных вилок: приведение к единой валюте и периоду (месяц).
3. Фильтрацию "мусорных" вакансий (стажировки с неопределёнными требованиями, вакансии-призраки).

4. Инструменты Yandex Cloud для построения эмбеддингов: YandexGPT Embeddings

Для эффективной реализации предложенной методологии семантической дедупликации и классификации необходим инструмент, обеспечивающий качественное векторное представление русскоязычных текстов вакансий и резюме. В качестве такого инструмента в настоящем исследовании предлагается использовать сервис **YandexGPT Embeddings**, доступный в платформе Yandex Cloud [9].

YandexGPT Embeddings представляет собой нейросетевую модель, специализированную для преобразования текстов на русском языке в плотные векторные представления (эмбеддинги) фиксированной размерности. Ключевые особенности сервиса:

- **Высокое качество для русского языка:** модель обучалась на больших массивах русскоязычных текстов, включая деловую и техническую лексику, что обеспечивает точное семантическое сходство для текстов сферы труда (названия профессий, навыки, описания обязанностей).
- **Простота интеграции:** доступ предоставляется через REST API, что позволяет легко встраивать получение эмбеддингов в пайплайн обработки данных на любом языке программирования. В примерах данной статьи используется Python.
- **Масштабируемость:** сервис поддерживает пакетную обработку (batch inference) и асинхронные вызовы, что критически важно при работе с массивами данных объёмом миллионы записей.
- **Интеграция с экосистемой Yandex Cloud:** возможность использования в сочетании с управляемыми базами данных (например, PostgreSQL с расширением pgvector) и сервисами аналитики для построения комплексных решений.

Сравнение с альтернативными подходами. Традиционно для семантического анализа русскоязычных текстов применялись open-source модели семейства BERT (RuBERT, LaBSE и др.). Их преимущество — бесплатность и возможность локального развертывания. Однако они имеют ряд ограничений: фиксированный контекст (обычно 512 токенов), необходимость дополнительной настройки под предметную область и более низкое качество на специализированной лексике по сравнению с коммерческими аналогами, постоянно дообучаемыми на актуальных данных. YandexGPT Embeddings лишены этих недостатков: они поддерживают больший контекст, регулярно обновляются разработчиком и показывают более высокие метрики в задачах поиска семантических дубликатов и классификации для русскоязычных текстов деловой тематики. Основным ограничением является стоимость вызовов API и зависимость от доступности облачного сервиса, что, однако, приемлемо для задач государственного масштаба при надлежащем бюджетировании.

Пример практического использования. В листинге 1 (раздел 3.1) приведён базовый пример синхронного получения эмбеддинга для одного текста. Для обработки больших объёмов данных рекомендуется применять асинхронные запросы и пакетную обработку, чтобы эффективно использовать квоты и минимизировать задержки. Ниже приведён фрагмент кода, демонстрирующий асинхронное получение эмбеддингов для списка текстов с использованием библиотеки asyncio и aiohttp:

python

```
import asyncioimport aiohttpimport numpy as npAPI_URL = "https://llm.api.cloud.yandex.net/foundationModels/v1/embedding"API_KEY = "ваш_ключ"async def get_embedding_async(session, text):    headers = {"Authorization": f"Api-Key {API_KEY}"}    payload = {"model": "yandexgpt/latest", "text": text}    async with session.post(API_URL, headers=headers, json=payload) as resp:        data = await resp.json()        return np.array(data["embedding"])async def get_all_embeddings(texts):    async with aiohttp.ClientSession() as session:        tasks = [get_embedding_async(session, text) for text in texts]        return await asyncio.gather(*tasks)# Пример
```

```
использованияtexts = ["резюме 1", "резюме 2", ...]embeddings =
asyncio.run(get_all_embeddings(texts))
```

Полученные векторы могут быть непосредственно использованы для расчёта косинусной близости при кластеризации (как показано в разделе 3.1) или для обучения классификаторов профобластей (раздел 5.1).

Рекомендации по выбору порогов и интеграции с векторными базами данных. При использовании YandexGPT Embeddings для дедупликации порог косинусной близости (threshold) рекомендуется подбирать эмпирически на валидационной выборке, размеченной экспертами. Для задачи объединения резюме одного соискателя типичные значения лежат в диапазоне 0.8–0.9. Для ускорения поиска дубликатов в больших массивах целесообразно хранить эмбеддинги в специализированных векторных базах данных, таких как Qdrant, pgvector или FAISS, с поддержкой индексации для быстрого поиска ближайших соседей.

Стоимость и квоты. Тарификация YandexGPT Embeddings зависит от объёма обрабатываемого текста (количества токенов) и количества запросов. Для ориентировочных расчётов: стоимость обработки одного резюме (примерно 500–1000 токенов) составляет доли рубля. При ежемесячном объёме в несколько миллионов резюме затраты будут существенными, но сопоставимыми с затратами на содержание собственной инфраструктуры для развёртывания open-source моделей, при этом качество и скорость разработки будут выше. Перед началом масштабного использования рекомендуется провести тестовый прогон на репрезентативной выборке для уточнения бюджета.

Таким образом, YandexGPT Embeddings представляют собой эффективный и доступный инструмент для реализации задач семантического анализа в рамках предлагаемой методологии, сочетающий высокое качество, простоту интеграции и масштабируемость облачного сервиса.

5. Семантический анализ: от количества к качеству

Опираясь на методологию И.Л. Сизовой [8], автор предлагает применять методы NLP для перехода к качественным характеристикам рынка труда, используя возможности YandexGPT Embeddings и Yandex Cloud.

5.1. Классификация соискателей по профобластям и уровням

Использование эмбеддингов YandexGPT для представления текста резюме и последующая классификация с помощью простых моделей (логистическая регрессия, kNN) по рубрикатору платформы. Это позволяет строить индексы предложения труда в разрезе профессий и отраслей.

5.2. Определение активности через анализ тональности и намерений

Анализ поля "О себе" с помощью моделей тональности (например, RuSentiment) и классификаторов намерений (намерение сменить работу vs. пассивный поиск). При необходимости может быть использован YandexGPT для zero-shot классификации намерений.

5.3. Индекс соответствия (Matching Index)

Измерение «зазора компетенций» между требованиями в вакансиях и навыками в резюме. Для каждой профобласти рассчитывается:

$$MI_k = 1 - \frac{|Skills_{required,k} \cap Skills_{available,k}|}{|Skills_{required,k} \cup Skills_{available,k}|}$$ $$MI_k = 1 - \frac{|Skills_{required,k} \cap Skills_{available,k}|}{|Skills_{required,k} \cup Skills_{available,k}|}$$

Чем выше MI, тем больше структурный дисбаланс. Для определения множеств навыков может использоваться семантическая близость эмбеддингов навыков, полученных через YandexGPT Embeddings.

6. Триангуляция и калибровка: от платформенных данных к официальной статистике

Для повышения надёжности итоговых показателей используется метод триангуляции [6, 7]:

1. **Многоуровневый сбор:** регулярный парсинг данных с hh.ru, Superjob, Авито.Работа и интеграция с открытыми данными портала «Работа России».
2. **Выявление платформенных шумов** через сравнение динамики показателей. Если на одной платформе резкий всплеск, а на других нет — это технический сбой или изменение политики платформы.
3. **Калибровка по Росстату:** построение мостовых уравнений (bridge equations) между очищенными индексами и официальными квартальными данными.

Пример bridge-уравнения:

$$Y_{Rosstat,t} = \alpha + \beta \cdot Index_{platform,t} + \gamma \cdot X_{control,t} + \varepsilon_t$$ $$Y_{Rosstat,t} = \alpha + \beta \cdot Index_{platform,t} + \gamma \cdot X_{control,t} + \varepsilon_t$$

Где $Y_{Rosstat}$ — официальная безработица или численность занятых, $Index_{platform}$ — наш Индекс реального предложения труда, $X_{control}$ — сезонные и календарные факторы.

7. Система оперативных индикаторов для целей ДКП

На основе предложенной методологии формируется набор оперативных индикаторов, пригодных для интеграции в модели прогнозирования Банка России.

Таблица 2. Система индикаторов рынка труда на основе платформенных данных

Индикатор	Формула/Метод расчета	Экономический смысл	Частота обновления
Индекс реального предложения труда (RLSI)	$\sum_{clusters} \sum_{i \in cluster} W_i$	Количество уникальных активных соискателей	Ежедневно
Индекс напряжённости (TI)	$\frac{Vacancies}{RLSI}$	Соотношение спроса и предложения (аналог отношения вакансий к безработным)	Ежедневно
Индекс зарплатного давления (WPI)	Медианная зарплата в новых вакансиях, взвешенная по профобластям	Инфляционное давление со стороны рынка труда	Еженедельно
Индекс структурной эффективности (SEI)	Производный от Matching Index (1 - средний MI)	Уровень сбалансированности компетенций	Ежемесячно

8. Эмпирическая проверка: гипотезы и ожидаемые результаты

В ходе дальнейших исследований планируется проверить следующие гипотезы:

1. **H1:** Применение каскадной дедупликации сокращает оцениваемое количество уникальных соискателей на 25–35% по сравнению с сырыми данными.
2. **H2:** Поведенческое взвешивание (AWS) повышает корреляцию агрегированных индексов с официальной статистикой занятости с 0.4–0.5 до 0.7–0.8.

3. **H3:** Индекс напряжённости (TI) опережает официальные показатели безработицы на 1–2 месяца, что делает его полезным для nowcasting.
4. **H4:** Индекс зарплатного давления (WPI) является значимым предиктором инфляции в секторе услуг (с лагом 3–6 месяцев).

План эмпирической проверки:

- Сбор данных с платформ за период 2024–2026 гг. (не менее 24 месяцев).
- Применение разработанных алгоритмов на выборке.
- Сравнение с квартальными данными Росстата и расчёт метрик качества (MAE, RMSE, корреляция).
- Построение и тестирование bridge-уравнений.

9. Ограничения и направления будущих исследований

Основные ограничения текущей версии:

- Отсутствие доступа к полным логам поведенческой активности пользователей (отклики, просмотры) — коэффициент B может быть рассчитан только приближённо.
- Различия в политиках платформ: некоторые закрывают данные для парсинга, что требует заключения официальных соглашений.
- Необходимость адаптации вызовов API YandexGPT Embeddings под требования по скорости и стоимости при обработке больших массивов данных.

Направления будущих исследований:

1. Интеграция данных с платформ фриланса и временной занятости.
2. Разработка динамической модели весов, учитывающей макроэкономический контекст (например, в кризис "мёртвые души" могут активизироваться).
3. Применение графовых нейросетей (GNN) для выявления связей между соискателями и работодателями.

Заключение и выводы

В ходе работы были решены следующие задачи:

1. **Типологизированы искажения** данных российских платформ занятости и предложены методы их коррекции.
2. **Разработана методология каскадной дедупликации**, включающая точное хэширование, MinHash LSH и семантический анализ на основе эмбеддингов YandexGPT.

3. **Введён механизм поведенческого взвешивания (AWS)**, формализующий понятие «реального предложения труда» с учётом активности и свежести данных.
4. **Предложена система оперативных индикаторов** (RLSI, TI, WPI, SEI), пригодных для nowcasting и макроэкономического анализа.
5. **Сформулированы тестируемые гипотезы** и план эмпирической проверки.

Разработанная концепция создаёт фундамент для системы мониторинга рынка труда в Банке России, позволяя повысить точность краткосрочных прогнозов и оперативно реагировать на структурные изменения. Методология может быть адаптирована для других стран с развитым сегментом онлайн-рекрутмента, а использование коммерческих NLP-сервисов, таких как YandexGPT Embeddings, обеспечивает высокое качество и воспроизводимость результатов.

Литература

1. Ахтямов Р.Э. Искусственный интеллект в макроэкономике: от экспериментов ЦБ до гибридных моделей // Реальное время. 2026. URL: https://realnoevremya.ru/articles/383499-ii-v-makroekonomike-ot-eksperimentov-cb-do-gibridnyh-modeley
2. Сизова И.Л., Русакова М.М., Александрова А.А. Рынок соискателей и фрикционность поиска работы на онлайн-платформах // Экономическая социология. 2022. Т. 23, № 5. DOI: 10.17323/1726-3247-2022-5-45-77
3. Сизова И.Л., Орлова Н.С., Елагина Е.Д. Компетенции работников в условиях социально-экономической неопределённости // Социологический журнал. 2023. Т. 29, № 4. DOI: 10.19181/socjour.2023.29.4.2
4. Зайцева О.В. Онлайн-источники данных о рынке труда: классификация, характеристики и подходы к ранжированию // Белорусский экономический журнал. 2025. № 3. URL: http://edoc.bseu.by:8080/handle/edoc/109410
5. Сизова И.Л. Особенности подбора персонала: интеллектуальный анализ текстов резюме и вакансий // Регионология. 2025. Т. 33, № 2. DOI: 10.15507/2413-1407.129.033.202502.271-293
6. Kaur R. et al. Machine Learning-Based Deduplication: A Comprehensive Review // ACM Computing Surveys. 2026. Vol. 58, No. 2.
7. Hammer B. et al. Entity Resolution: Past, Present, and Yet to Come // Communications of the ACM. 2023. Vol. 66, No. 4.
8. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv:1810.04805. 2018.
9. Yandex Cloud. YandexGPT Embeddings. Документация API. URL: https://cloud.yandex.ru/docs/yandexgpt/embeddings (дата обращения: 04.03.2026)

References in Cyrillics

1. Akhtyamov R.E. Iskusstvennyy intellekt v makroekonomike: ot eksperimentov TsB do gibridnykh modeley [Artificial Intelligence in Macroeconomics: From Central Bank Experiments to Hybrid Models]. Real'noye vremya, 2026. URL: https://realnoevremya.ru/articles/383499-ii-v-makroekonomike-ot-eksperimentov-cb-do-gibridnyh-modeley
2. Sizova I.L., Rusakova M.M., Aleksandrova A.A. Rynok soiskateley i friktsionnost' poiska raboty na onlayn-platformakh [Job Seekers' Market and Frictional Job Search on Online Platforms]. Ekonomicheskaya sotsiologiya, 2022, Vol. 23, No. 5. DOI: 10.17323/1726-3247-2022-5-45-77
3. Sizova I.L., Orlova N.S., Elagina E.D. Kompetentsii rabotnikov v usloviyakh sotsial'no-ekonomicheskoy neopredelennosti [Workers' Competencies in Conditions of Socio-Economic Uncertainty]. Sotsiologicheskiy zhurnal, 2023, Vol. 29, No. 4. DOI: 10.19181/socjour.2023.29.4.2
4. Zaitseva O.V. Onlayn-istochniki dannykh o rynke truda: klassifikatsiya, kharakteristiki i podkhody k ranzhirovaniyu [Online Labor Market Data Sources: Classification, Characteristics, and Ranking Approaches]. Belorusskiy ekonomicheskiy zhurnal, 2025, No. 3. URL: http://edoc.bseu.by:8080/handle/edoc/109410
5. Sizova I.L. Osobennosti podbora personala: intellektual'nyy analiz tekstov rezyume i vakansiy [Features of Recruitment: Intelligent Analysis of Resume and Vacancy Texts]. Regionologiya, 2025, Vol. 33, No. 2. DOI: 10.15507/2413-1407.129.033.202502.271-293
6. Kaur R. et al. Machine Learning-Based Deduplication: A Comprehensive Review // ACM Computing Surveys. 2026. Vol. 58, No. 2.
7. Hammer B. et al. Entity Resolution: Past, Present, and Yet to Come // Communications of the ACM. 2023. Vol. 66, No. 4.
8. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv:1810.04805. 2018.
9. Yandex Cloud. YandexGPT Embeddings. Dokumentatsiya API. URL: https://cloud.yandex.ru/docs/yandexgpt/embeddings (accessed: 04.03.2026)

Ахтямов Равиль Энгелевич, кандидат экономических наук, digital-экономист, основатель Digital Economy Lab (e-mail: ravilakhtyamov@yandex.ru)
ORCID: 0000-0002-0783-8573

Ключевые слова

рынок труда, большие данные, HeadHunter, дедупликация, NLP, YandexGPT Embeddings, поведенческий скоринг, триангуляция данных, денежно-кредитная политика, nowcasting.

Akhtyamov R.E. Transformation of Online Platform Data into Reliable Labor Market Indicators: A Methodology for Cleaning, Weighting and Calibration using NLP and Behavioural Scoring

Keywords

labor market, big data, HeadHunter, deduplication, NLP, YandexGPT Embeddings, behavioural scoring, data triangulation, monetary policy, nowcasting.

JEL classification: C55, C81, J21, E24

Abstract

The article addresses the problem of systematic biases in big data from commercial employment platforms (resume multiplication, "dead souls," information asymmetry), which hinder their direct use for macroeconomic analysis and monetary policy. The aim is to develop a comprehensive theoretical and methodological concept for cleaning, weighting, and calibrating data from online platforms (hh.ru, "Rabota Rossii") to transform them into reliable labor market indicators suitable for the Bank of Russia's forecasting models. The research synthesizes methods from computer science (NLP based on YandexGPT embeddings, graph algorithms, behavioural scoring) and economic statistics. A cascaded deduplication approach is proposed (from exact hashing to semantic analysis using YandexGPT Embeddings), and an original formula for weighting resumes considering behavioural factors and activity is constructed. Data verification employs triangulation methods and calibration against official Rosstat statistics using bridge equations. Scientific novelty lies in the holistic methodology combining exact and semantic deduplication with behavioural weighting and multi-level triangulation of data from different sources, as well as in formulating a set of testable hypotheses for empirical validation. The result is a framework for a system of operational labor market indicators (Real Labor Supply Index, Tightness Index, Wage Pressure Index, Structural Efficiency Index) that can be integrated into the Bank of Russia's forecasting models. The proposed toolkit has undergone initial conceptual validation.