УДК 004.85

РАЗРАБОТКА СИСТЕМЫ АНАЛИЗА ТЕКСТОВ ВАКАНСИЙ ОНЛАЙН-БИРЖ ТРУДА НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Сотниченко З. Э. (ИТМО)
Научный руководитель – доцент Болдырева Е. А.
(ИТМО)

Введение. Современный рынок труда характеризуется огромным объемом вакансий, публикуемых ежедневно на онлайн-биржах труда, что делает ручной анализ этих данных практически невозможным. Автоматическая классификация вакансий по профессиональным категориям с помощью методов машинного обучения позволяет получать оперативные трудовые индикаторы и выявлять тенденции рынка в режиме реального времени [1]. Целью исследования являлось создание и сравнительный анализ моделей машинного обучения для многоклассовой классификации текстов вакансий, а также разработка аналитической системы для мониторинга динамики и прогнозирования тенденций рынка труда.

Основная часть. Исследование охватило полный цикл работы с данными: от сбора и предобработки до обучения моделей и разработки веб-приложения.

Сбор и подготовка данных. Сформирован датасет из 64 621 уникальной вакансии, собранной из трех источников: HeadHunter (94,9%), SuperJob (3,6%) и TrudVsem (1,5%). Данные прошли очистку, дедупликацию и NLP-предобработку: токенизацию, лемматизацию, удаление стоп-слов. Выделено 14 профессиональных категорий, включая ИТ, продажи, медицину, логистику, финансы и другие. Векторизация выполнена методом TF-IDF [2] с параметрами: 50 000 признаков, униграммы и биграммы, min_df=2, max_df=0,95.

Обучение и сравнение моделей. Обучено шесть моделей на выборке 80/20 со стратифицированным разбиением и 5-fold кросс-валидацией. Лучший результат показала LightGBM [3]: accuracy 89,66%, precision 90,91%, F1-score 89,94%. Второе место заняла LinearSVC (89,18%), затем XGBoost [4] (87,88%), логистическая регрессия (83,37%), случайный лес (76,90%) и Naive Bayes (74,12%). Разрыв между лучшей и худшей моделями составил более 15%, что подтвердило критическую важность выбора алгоритма для данной задачи.

Веб-прототип аналитической системы. На базе фреймворка Streamlit разработано веб-приложение с модульной архитектурой из девяти компонентов: дашборд с обзорной статистикой, интерактивная классификация вакансий, оценка ML-моделей с матрицами ошибок, анализ динамики навыков, прогнозирование emerging-профессий, прогнозы временных рядов (Prophet, ARIMA), выявление закономерностей и поиск вакансий. Пользовательское тестирование с участием 12 респондентов из числа HR-специалистов и аналитиков показало среднюю оценку 4,3 из 5 и подтвердило экономию времени на аналитических задачах в 40-60%.

Анализ динамики и прогнозирование. На расширенном датасете из 173 562 вакансий (июль 2025 – январь 2026) проведён анализ временных рядов. Среднее количество публикаций составило 952,7 вакансий/день с коэффициентом вариации 30,2%. Выявлены emerging-профессии: Data Science/ML Engineering (рост доли +53%), DevOps/Cloud Engineering (+28%), Cybersecurity (+39%). Определены компетенции со снижающимся спросом: базовые офисные навыки (-25%), рутинная бухгалтерия (-21%), ручное тестирование ПО (-25%).

Выводы. Разработанная система подтвердила эффективность применения методов машинного обучения для анализа рынка труда. Модель LightGBM достигла точности 89,66% при классификации по 14 категориям, что достаточно для промышленного

применения. Методы градиентного бустинга превзошли классические алгоритмы для текстов с высокой семантической вариативностью. Интеграция моделей в веб-систему с интерактивной визуализацией обеспечила доступность результатов для HR-специалистов и аналитиков. Анализ временных рядов выявил устойчивые тренды роста спроса на специалистов в области Data Science, DevOps и кибербезопасности при одновременном снижении потребности в рутинных офисных и бухгалтерских компетенциях.

Список использованных источников:

1. Pedregosa F. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830. URL: https://jmlr.org/papers/v12/pedregosa11a.html (дата обращения: 10.02.2026).
2. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management. – 1988. – Vol. 24, No. 5. – P. 513–523. URL: https://doi.org/10.1016/0306-4573(88)90021-0 (дата обращения: 10.02.2026).
3. Ke G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Advances in Neural Information Processing Systems 30 (NIPS 2017). – Long Beach, CA, USA, 2017. – P. 3146–3154. URL: https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree (дата обращения: 10.02.2026).
4. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – San Francisco, CA, USA, 2016. – P. 785–794. URL: https://doi.org/10.1145/2939672.2939785 (дата обращения: 10.02.2026).