



**ОТЗЫВЫ И РЕЦЕНЗИИ**

**В.П. Мешалкин**  
Новое издание учебника  
«Конкуренция в предпринимательстве»  
для вузов . . . . . 5

**ИТ-МЕНЕДЖМЕНТ**

**Управление эффективностью**  
**Л.А. Комарова, А.Д. Черемухин**  
Повышение эффективности методов  
подбора персонала на основе  
глубоких нейронных сетей . . . . . 10

**ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА**

**Модели и методики**  
**Т.Т. Каримбаев, С.А. Медведев, Е.Т. Захарова**  
Использование алгоритма оптимизации  
DIRECT и кусочно-линейных функций  
для построения нелинейных регрессионных  
моделей по гетероскедастичным данным . . . . . 23

**М.С. Клячин**  
Построение экономико-математической модели  
для повышения экономической эффективности  
проведения массовых мероприятий  
за счет скользящего планирования . . . . . 42

**Эффективные алгоритмы**

**А.В. Бойков, В.М. Никоноров,  
Д.А. Лучин, И.В. Ильин**  
Применение нейросетевого подхода для обнаружения  
дефектов текстильных материалов  
классов «слет» и «затек воды» . . . . . 55

**С.Г. Небаба, А.А. Захарова**  
Алгоритмы получения, комплексования  
и обработки слабоформализованных гетерогенных  
данных аэрофотосъемки . . . . . 67

**Программная инженерия**

**А.Ю. Пучков, Я.А. Федулов, С.В. Незамаев**  
Гибридная интеллектуальная система машинного  
обучения для моделирования процессов обработки  
фосфатного рудного сырья . . . . . 83

**Д.И. Читалов**  
Разработка модуля для управления  
моделированием системы двух сжимаемых  
жидких фаз с одной дисперсной фазой  
на базе платформы OpenFOAM . . . . . 106

**ЛАБОРАТОРИЯ**

**Моделирование процессов и систем**

**Ф.В. Краснов**  
Повышение полноты и точности поиска товаров  
на торговых интернет-площадках . . . . . 118

Правила оформления материалов . . . . . 137

DOI: 10.37791/2687-0649-2024-19-2-5-9

# Новое издание учебника «Конкуренция в предпринимательстве» для вузов

*Рубин Ю.Б.* Конкуренция в предпринимательстве. – 9-е изд., перераб. и доп. – Москва: Университет «Синергия», 2023. – 896 с. (Университетская серия). DOI: 10.37791/978-5-4257-0546-4-2023-1-896.

**Аннотация.** Представлена рецензия на новое издание учебника «Конкуренция в предпринимательстве» для вузов, вышедшего в издательстве Университета «Синергия». Отмечены основные преимущества учебника: наличие авторского взгляда на такие вопросы, как процесс соперничества, его объекты, предметы, акторы, результаты, ресурсы, типы и характер конкурентного взаимодействия, виды конкурентных действий, методы операционного взаимодействия сторон соперничества, конкурентная устойчивость и конкурентоспособность, конкурентные позиции, конкурентный статус, конкурентные стратегии, тактические конкурентные операции и комбинации, конкурентные ситуации, конкурентные уловки и головоломки, управление конкурентными действиями, конкурентный анализ. В условиях цифровой экономики и применения интеллектуальных технологий управления использование рассматриваемого издания в учебном процессе позволит сформировать у студентов вузов знания и умения, необходимые для ведения успешной предпринимательской деятельности в современных условиях хозяйствования. Описанный в учебнике инструментарий может эффективно использоваться на стратегическом, операционном и ситуационном уровнях участия в конкуренции с учетом потенциала его реализации в составе корпоративных информационных систем организации.

**Ключевые слова:** цифровая экономика, конкуренция, предпринимательство, информационные системы

**Для цитирования:** Мешалкин В.П. Новое издание учебника «Конкуренция в предпринимательстве» для вузов // Прикладная информатика. 2024. Т. 19. № 2. С. 5–9. DOI: 10.37791/2687-0649-2024-19-2-5-9

# New edition of the textbook “Competition in Entrepreneurship” for higher educational institutions

Rubin Yu. Competition in Entrepreneurship. 9th edition, rev. and add. Moscow, Synergy University, 2023, 896 p. (University series). DOI: 10.37791/978-5-4257-0546-4-2023-1-896.

**Abstract.** A review for the new edition of the textbook “Competition in Entrepreneurship” for higher educational institutions, published by the University publishing house “Synergy”, is presented. The following main advantages of the textbook are noted: the presence of the author’s view on such issues as the process of competitiveness, its objects, subjects, actors, results, resources, types and character of competitive interaction, types of competitive actions, methods of operational interaction between competitive parties, competitive stability and competitive ability, competitive positions, competitive status, competitive strategies, tactical competitive operations and combinations, competitive situations, competitive tricks and puzzles, management of competitive actions, competitive analysis. In the context of digital economy and the application of intelligent management technologies, the use of the considered publication in the educational process will allow university students to develop the knowledge and skills necessary to conduct successful business activities in modern economic conditions. The tools described in the textbook can be effectively used at the strategic, operational and situational levels of participation in competition, considering the potential for its implementation as part of the organization corporate information systems.

**Keywords:** digital economy, competitiveness, entrepreneurship, information systems

**For citation:** Meshalkin V. New edition of the textbook “Competition in Entrepreneurship” for higher educational institutions. *Prikladnaya informatika*=Journal of Applied Informatics, 2024, vol.19, no.2, pp.5-9 (in Russian). DOI: 10.37791/2687-0649-2024-19-2-5-9

Начало XXI века сфокусировало внимание российской высшей школы и отечественной науки на проблемах, которые хотя и осознавались ранее, но не воспринимались как критически значимые для развития страны. Становится вполне очевидным, что без обеспечения конкурентоспособности национальной экономики в условиях ее цифровой трансформации, эво-

люции российского бизнеса в направлении широкого использования интеллектуальных технологий управления (в этом плане следует рассматривать и усилия, связанные с имплементацией подходов в области использования «во благо» искусственного интеллекта) на технологический суверенитет и ощутимые преимущества во всемирной деловой среде рассчитывать трудно.

На этом фоне ускорения и обострения конкуренции в глобальном экономическом пространстве стоит обратить внимание на учебник «Конкуренция в предпринимательстве», вышедший в издательстве Университета «Синергия» в 2023 году, – явно незаурядное явление на просторах публикационной активности и учебно-методического творчества.

Книг о стратегиях, тактике, профессиональных навыках и технологиях участия в специфике конкуренции, которые к тому же были бы написаны языком, понятным студентам, у нас не так много. Читателей ждет глубокое погружение в конкурентную деятельность, которая составляет, вероятно, решающее направление достижения ощутимых успехов в современном предпринимательстве. Автор последовательно разбирает такие вопросы, как процесс соперничества, его объекты, предметы, акторы, результаты, ресурсы, типы и характер конкурентного взаимодействия, виды конкурентных действий, методы операционного взаимодействия сторон соперничества, конкурентную устойчивость и конкурентоспособность, конкурентные позиции, конкурентный статус, конкурентные стратегии, тактические конкурентные операции и комбинации, конкурентные ситуации, конкурентные уловки и головоломки, управление конкурентными действиями, конкурентный анализ. Пожалуй, впервые на русском языке данные вопросы рассматриваются столь подробно, разносторонне и системно, да к тому же в формате учебника для вузов.

Примечательна в этой связи фигура самого автора. Юрий Борисович Рубин – профессор, доктор экономических наук, член-корреспондент Российской академии

образования, основоположник функциональной теории предпринимательства, хорошо известен как автор нестандартных, зачастую полемически заостренных публикаций в научных журналах и в СМИ в целом. В начале 2000-х годов в качестве одного из руководителей УМО по информатике вошел в число инициаторов формирования ныне чрезвычайно востребованной специальности (в настоящее время – направления) «Прикладная информатика» в системе высшего образования РФ. Был одним из зачинателей применения дистанционных образовательных технологий, создав в 2000 году первый в России интернет-портал полнообъемного дистанционного обучения E-Education.ru и в 2001 году первый в стране комплект электронных учебников на CD-носителях по специальностям «Информационные системы в экономике», «Менеджмент» и «Антикризисное управление». С 2009 по 2015 год Ю.Б. Рубин входил в Совет директоров Европейского фонда обеспечения качества электронного обучения (EFQUEL). В 2000–2020-е годы в центре научных исследований Ю.Б. Рубина оказались также вопросы формирования системы гарантий качества образования и независимой оценки качества российского образования, создания предпринимательских университетов, теории и практики электронного обучения, обеспечения конкурентоспособности программ высшего образования и взаимных гарантий участников сферы образования.

Ядром столь разнообразной тематики всегда были вопросы обеспечения конкурентоспособности деятельности на основе применения субъектами актуальных методов и инструментов участия в конкуренции. Параллельно совершенствовались

и представления автора о конкуренции как о специфическом объекте междисциплинарных исследований, которые публиковались в монографиях, научных статьях, учебниках и в конце концов подвели его к формулированию конструктивной теории конкуренции, которая, на мой взгляд, является вкладом в управленческую науку, имеющим несомненное теоретическое и прикладное значение.

Учебник «Конкуренция в предпринимательстве» выходит уже 9-м изданием и является дополнением учебно-методической базы курса «Управление собственным бизнесом» – первого и пока единственного на русском языке по данной тематике. Его содержание обусловлено особенностями понимания предпринимательства как специфического вида профессиональной деятельности людей, требующего от акторов наличия необходимых для его успешного выполнения набора обязательных знаний и навыков, в том числе в области участия в конкуренции. Ю.Б. Рубин вводит проблематику в контекст изучения процесса становления в России системы обучения предпринимательству на разных уровнях получения образования. Не случайно он с 2015 года возглавляет Рабочую группу СПК ФР по созданию профессионального стандарта предпринимателя, с 2017 года – Научный совет по развитию предпринимательского образования при Отделении общего среднего образования Российской академии образования, с 2021 года – Федеральную инновационную площадку «Модель инновационной системы непрерывного предпринимательского образования в образовательной организации высшего образования Российской Федерации», а с 2022 года входит в Рабочую

группу по реализации мер, направленных на развитие молодежного предпринимательства Совета Министерства науки и высшего образования РФ по реализации государственной молодежной политики.

Успешное профессиональное предпринимательство – прежде всего конкурентоспособное предпринимательство, а успешный предприниматель – конкурентоспособный предприниматель. Таким, уверен автор, его делает обладание компетенциями и профессионально важными качествами в области совершения конкурентных действий и управления ими на стратегическом, операционном и ситуационном уровнях участия в конкуренции. Как, против кого, совместно с кем совершать эти действия в конкурентной среде, как избегать или смягчать неизбежные риски, форсировать технические параметры конкурентных действий, составлять из них тактические конкурентные операции, осуществлять операционные маневры, лавировать, применять тактические хитрости, планировать, организовывать, контролировать совершение конкурентных действий, а также мотивировать самих себя, воспитывая конкурентный дух и конкурентный образ мышления, – на все эти непростые вопросы учебник содержит актуальные, порой парадоксальные ответы.

Чтобы быть конкурентоспособным, качественно и функционально выполнять конкурентные действия, субъект современного предпринимательства, полагает автор, должен хорошо разбираться в их многообразии. Восприятие конкуренции в предпринимательстве как совокупности адресуемых соперникам конкурентных действий делает термины «конкурентные действия», «конкурентная деятельность»,

«участие в конкуренции» определяющими в понятийном аппарате учебника. На этой основе автор производит упорядочение терминологического аппарата теории конкуренции, находя достойное применение таким категориям, как «конкурентные операции», «конкурентные стратегии», «конкурентная устойчивость бизнеса», «конкурентоспособность», «конкурентные преимущества и недостатки», «конкурентные ситуации», «конкурентные результаты», «конкурентная ценность», «компетентность в вопросах участия в конкуренции», «личностная конкурентность», «конкурентный потенциал», «конкурентный статус», «конкурентные функции», «конкурентные позиции».

Подробно рассматриваются и специфические вопросы управления конкурентными действиями на всех уровнях внутрифирменного управления – пожалуй, впервые в русскоязычном сегменте литературы о менеджменте.

В издании представлены учебные кейсы и тестовые задания, позволяющие

студентам успешно усваивать богатый материал.

Учебник будет способствовать решению важнейшей задачи – подготовки высококвалифицированных кадров по экономико-управленческим и инженерно-техническим направлениям.

В общем российский читатель получил интересное и полезное учебное издание, которое способно помочь овладеть премудростями современного предпринимательства в условиях цифровой экономики, научиться профессионально действовать и создавать стартапы в трудно предсказуемой конкурентной среде, используя потенциал информационных систем управления организацией.

*Директор  
Международного института  
логистики ресурсосбережения  
и технологической инноватики  
РХТУ им. Д. И. Менделеева,  
академик РАН  
В. П. Мешалкин*

## Сведения об авторе

*Мешалкин Валерий Павлович*, ORCID 0000-0001-6956-6705, докт. техн. наук, профессор, академик РАН, директор Международного института логистики ресурсосбережения и технологической инноватики, Российский химико-технологический университет имени Д.И. Менделеева, Москва, Россия, clogist@muctr.ru

Статья поступила 27.03.2024, рассмотрена 01.04.2024, принята 04.04.2024

## About the author

*Valery P. Meshalkin*, ORCID 0000-0001-6956-6705, Dr. Sci. (Eng.), Professor, Academician of Russian Academy of Sciences (RAS), Director of International Institute of Logistics and Resource Technology Innovation, D.I. Mendeleev University of Chemical Technology of Russia, Moscow, Russia, clogist@muctr.ru

Received 27.03.2024, reviewed 01.04.2024, accepted 04.04.2024

# Повышение эффективности методов подбора персонала на основе глубоких нейронных сетей

Л.А. Комарова<sup>1</sup>, А.Д. Черемухин<sup>2</sup>

<sup>1</sup>Финансовый университет при Правительстве Российской Федерации, Москва, Россия

<sup>2</sup>Нижегородский государственный инженерно-экономический университет, Княгинино, Россия  
<sup>\*</sup>229388@edu.fa.ru

**Аннотация.** Индустрия подбора персонала находится на переломном этапе: интеграция искусственного интеллекта уже оказала свое влияние на традиционные процессы найма и может произвести революцию. В этой статье представлен подход к классификации резюме по категориям должностей, использующий поиск семантического сходства для усовершенствования механизма подбора кандидатов в рекрутинге. Предложенный метод отличается от традиционных систем, основанных на ключевых словах, и представляет собой структуру глубокого обучения, которая понимает и обрабатывает сложную семантику документов, связанных с работой. Целью исследования является разработка метода классификации текстов резюме, имеющих сложную организационную структуру. Данное исследование решает сразу несколько задач: повышение точности классификации резюме и нахождение наиболее стабильной модели для решения задачи классификации резюме. Авторы провели сравнение стандартных методов машинного обучения с нейросетевыми и показали эффективность последних. Результаты указывают на улучшение метрик качества по сравнению с традиционными ML-моделями, предлагая подход, который может использоваться для прескрининга при подборе персонала с помощью искусственного интеллекта, который выбирает подходящих кандидатов из других кандидатов на вакансию. Также авторами была обнаружена проблема нестабильности результатов при дообучении больших языковых моделей, когда модель даже при одинаковых значениях гиперпараметров дает разные результаты. Был проведен ряд экспериментов, чтобы лучше понять это явление, с вариацией двух параметров – learning rate и seed. В результате обнаружены существенное увеличение производительности при определенном пороге параметром и возможность количественно определить, какие из найденных моделей работают лучше.

**Ключевые слова:** HR, ИИ, BERT, ALBERT, классификация, процесс рекрутинга

**Для цитирования:** Комарова Л.А., Черемухин А.Д. Повышение эффективности методов подбора персонала на основе глубоких нейронных сетей // Прикладная информатика. 2024. Т. 19. № 2. С. 10–22. DOI: 10.37791/2687-0649-2024-19-2-10-22

# Increasing the efficiency of recruitment based on deep neural networks

L. Komarova<sup>1</sup>\*, A. Cheremuhin<sup>2</sup>

<sup>1</sup>Financial University under the Government of Russian Federation, Moscow, Russia

<sup>2</sup>Nizhny Novgorod State Engineering-Economic University, Knyaginino, Russia  
\*229388@edu.fa.ru

**Abstract.** The recruitment industry is at an inflection point: the integration of artificial intelligence has already made its impact on traditional recruitment processes and has the potential to revolutionize it. This article presents an approach to classify resumes into job categories, using semantic similarity search to improve the candidate selection mechanism in recruiting. Our method differs from traditional keyword-based systems and is a deep learning framework that understands and processes the complex semantics of work-related documents. The purpose of the study is to develop a method for classifying resume texts with a complex organizational structure. This study solves several problems at once: increasing the accuracy of resume classification and finding the most stable model for solving the problem of resume classification. We compared standard machine learning methods with neural network ones and showed the effectiveness of the latter. The results indicate an improvement over traditional ML models, suggesting an approach that can be used for pre-screening artificial intelligence recruiting that selects suitable candidates from other applicants. Further, we discovered problems with instability of results when retraining large language models, when the model, even with the same values of the hyperparameters, gives different results. To better understand this phenomenon, we conducted a series of experiments with the main BERT models, varying two parameters – learning rate and seed. As a result, we find a significant increase in performance at a certain threshold parameter, and we quantify which of the found models perform better.

**Keywords:** HR, AI, BERT, ALBERT, classification, hiring process

**For citation:** Komarova L., Cheremuhin A. Increasing the efficiency of recruitment based on deep neural networks. *Prikladnaya informatika*=Journal of Applied Informatics, 2024, vol.19, no.2, pp.10-22 (in Russian). DOI: 10.37791/2687-0649-2024-19-2-10-22

## Введение

Наступление эры цифровой трансформации привело к переполнению всех сфер человеческой деятельности огромным объемом данных, что повлекло за собой многочисленные дискуссии о потенциале применения больших данных и необходимости пересмотра подходов к управлению разными аспектами деятельности организаций.

В условиях наблюдающегося дефицита высококлассных специалистов почти во всех отраслях экономики вопрос эф-

фективности подбора персонала начинает непосредственно влиять на общую эффективность работы организации. Это повлекло за собой множество попыток автоматизировать данный процесс, в том числе и с помощью методов машинного обучения. Один из вариантов такой автоматизации – разработка алгоритма, оценивающего соответствие вакансий и резюме. В данной статье представлен подход к классификации резюме по категориям работы с использованием техник поиска семантической схожести, ее целью является оценка точности решения задачи

оценки соответствия вакансий и резюме с применением методов NLP (области исследований, направленных на то, чтобы научить компьютер понимать, интерпретировать и генерировать человеческую речь) и машинного обучения.

## Материалы и методы

В ходе анализа научных работ было выявлено несколько подходов к решению проблемы классификации резюме по соответствующим предметным областям.

Первый подход заключался в применении базовых алгоритмов машинного обучения [7]. Конкретные методы включали случайный лес (Random Forest), метод опорных векторов (SVM) и наивный байесовский классификатор (Naive Bayes) [11] с использованием алгоритма оценки важности слов (TF-IDF) для получения их векторных представлений (эмбедингов) [13, 21].

Данные подходы использовались для классификации извлеченных навыков из резюме по соответствующим классам-должностям. Наивысший результат по метрикам был зафиксирован в [16] при использовании классификатора «случайный лес» (accuracy – 0,7, precision – 0,68, recall – 0,683, F1 – 0,678).

Второй подход – использование графовых нейронных сетей для классификации резюме [15, 19]. Они использовались для определения связей слов в корпусе, сосредоточив внимание на сущностях с иерархическим обучением графа, для улучшения методов классификации текста и были реализованы с использованием библиотеки глубокого обучения PyTorch.

В [24] ученые сравнивали эффективность нескольких моделей: Bi-LSTM, GCN (Graph Convolutional Networks), dotGAT (Graph Attention Neural Networks with Dot-

Product Attention), MGAT (Multi-headed Attention Neural Networks). Результаты исследования показали, что наилучшее качество продемонстрировала модель MGAT с дообучением на большом датасете с резюме (accuracy – 0,7552, macro-F1 – 0,7548).

Несколько по-другому подошли к использованию графовых нейронных сетей ученые в [25], используя в качестве основной модели HGCN (Heterogeneous Graph Convolutional Network). Эта модель предназначена для анализа данных резюме и предоставления рекомендаций (подходящих вакансий) на основании метаданных в графе резюме для определения синтаксических и семантических отношений между резюме и вакансией. Она объединяет узлы графа и информацию из соседних узлов для анализа данных, а добавление метаданных повышает производительность за счет предоставления дополнительной информации из резюме. Данный подход показал свою эффективность (accuracy – 0,859, precision – 0,9, macro-F1 – 0,82, AUC – 0,8475) относительно GAT и GCN.

Третий подход – использование семантики [17, 18] для решения задачи классификации резюме. Исследование [6] выявляет повторяющиеся особенности в семействах должностей при помощи метода LDA [13, 14] и кластеризации (K-means) [21, 23] и использует их для характеристики каждой области. В нем также рассматривается относительная важность различных наборов навыков в этих группах должностей. Данный подход хорошо подходит для структурированных и простых задач категоризации, где объяснимость и простота имеют первостепенное значение.

Проведенный анализ научной литературы [2, 5] позволил заключить, что для оценки эффективности предложенных

моделей чаще всего используются классические метрики оценки качества классификации, такие как ассигасу, precision, macro-F1 и AUC [8].

Однако классическая метрика AUC больше подходит для оценки качества бинарной классификации и может плохо себя показывать на несбалансированных данных, поэтому авторы приняли решение исключить ее из перечня метрик для оценки качества соотнесения резюме и вакансий.

Было построено 6 моделей, реализующих классические методы машинного обучения (на основе случайного леса (Random Forest), метода k-ближайших соседей (KNN), метода опорных векторов (SVC), линейного метода опорных векторов (LinearSVM), модели многоклассовой классификации OneVsRest и алгоритма, использующего стохастический градиентный спуск (SGDClassifier)), и 5 нейросетевых моделей семейства BERT: BERT [10, 20] (глубокая нейронная сеть с миллионами параметров, способная улавливать контекстуальные зависимости между словами; раз-

работана компанией Google в 2018 году), RoBERTa [4] и DistilBERT, Albert из Hugging Face [3] и LLM модель ChatGPT3.5 [12, 22]. Сравнение параметров нейросетевых моделей представлено в таблице 1.

Объектом исследования стал набор данных резюме, находящийся в свободном доступе на сайте Kaggle.com<sup>1</sup>, который содержит не только текст резюме и названия должностей, но и ID-категории, к которой относится данное резюме (24 категории). Всего в наборе содержится 2484 резюме в pdf-формате. Распределение резюме по категориям представлено на рисунке 1.

Первичная обработка данных датасета включала в себя использование библиотеки nltk (удаление стоп-слов, приведение текста к нижнему регистру, фильтрация тегов). Фрагмент полученного датафрейма представлен на рисунке 2 (Feature – столбец с очищенными данными, которые использовались для дальнейшего моделирования).

<sup>1</sup> URL: <https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset/data> (дата обращения: 27.03.2024).

Таблица 1. Базовые BERT-модели

Table 1. Base BERT models

| Параметры моделей<br><i>Model parameters</i> | BERT                       | Roberta                                                  | DistilBERT                                                                | ALBERT                                                                                                               |
|----------------------------------------------|----------------------------|----------------------------------------------------------|---------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| Параметры обучения, млн                      | 110                        | 125                                                      | 66                                                                        | 60                                                                                                                   |
| Слой / скрытые слои / головы                 | 12/768/12                  | 12/768/12                                                | 6/768/12                                                                  | 12/768/12                                                                                                            |
| Данные для обучения                          | Book corp + Eng. Wikipedia | Bert + CCNews + Open-WebText + Stories                   | Book corp + Eng. Wikipedia                                                | Book corp + Eng. Wikipedia                                                                                           |
| Pre-training (особенности)                   | Bidirectional transformer  | Обучение на большем объеме данных (в 10 раз больше BERT) | Bert + методы дистилляции знаний, чтобы уменьшить размер модели (~на 40%) | Bert + два метода уменьшения параметров, чтобы уменьшить использование памяти и увеличить скорость во время обучения |

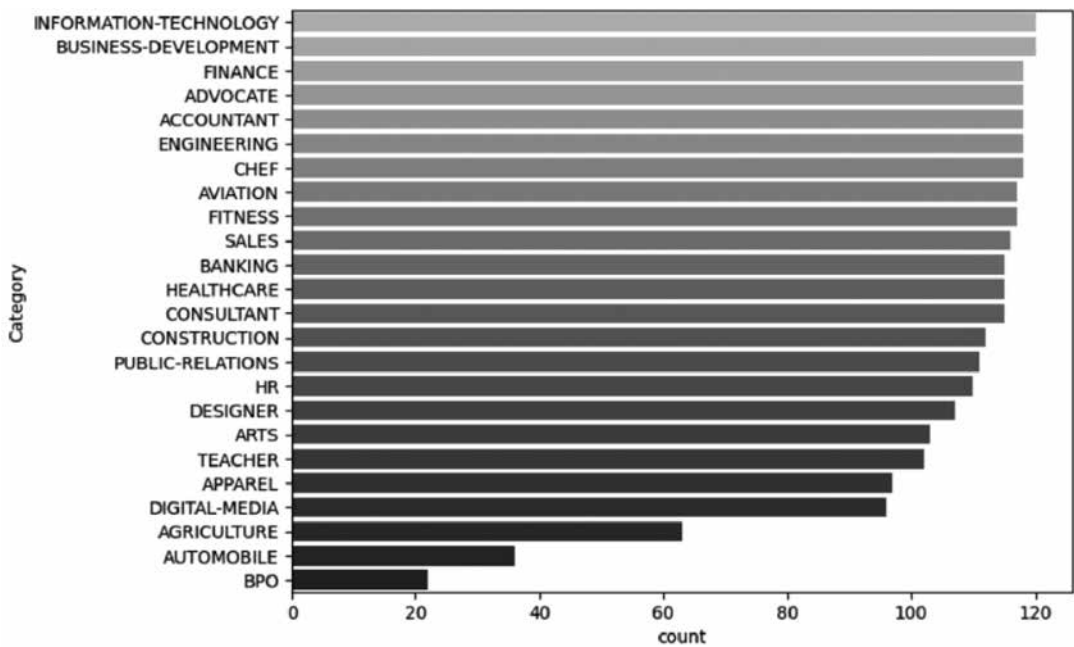


Рис. 1. Распределение резюме по категориям  
Fig. 1. CV distribution by categories

| ID | Resume_str | Category                                 | Feature                                              |
|----|------------|------------------------------------------|------------------------------------------------------|
| 0  | 16852973   | HR ADMINISTRATOR/MARKETING ASSOCIATE\... | HR hr administrator marketing associate hr admini... |
| 1  | 22323967   | HR SPECIALIST, US HR OPERATIONS ...      | HR hr specialist hr operations summary media prof... |
| 2  | 33176873   | HR DIRECTOR Summary Over 2...            | HR hr director summary years experience recruitin... |
| 3  | 27018550   | HR SPECIALIST Summary Dedic...           | HR hr specialist summary dedicated driven dynamic... |
| 4  | 17812897   | HR MANAGER Skill Highlights ...          | HR hr manager skill highlights hr skills hr depar... |

Рис. 2. Срез данных  
Fig. 2. Data snapshot

Для построения моделей, согласно общепринятому подходу, данные были разделены на 3 части: train (80%), test (10%), validation (10%). Обучающая часть данных (1984 текста) была использована для обучения моделей классификации резюме, а метрики были рассчитаны на тестовой части (248 текстов).

При обучении и оценке качества классических ML-моделей использовалась библиотека sklearn, при этом для каждой мо-

дели проводился подбор гиперпараметров с использованием библиотеки Optuna [1]. Для решения задачи классификации текстов в таких моделях машинного обучения в качестве признаков целесообразно использовать эмбединги слов. Одним из подходов получения эмбедингов из текстов является использование алгоритмов, таких как TF-IDF, word2vec, GloVe или fast text. Эти алгоритмы берут большой объем текста и изучают векторные

представления для каждого слова, так что слова, часто встречающиеся в одном и том же контексте, сопоставляются с близлежащими точками в векторном пространстве. В данной статье использован подход TF-IDF (Term Frequency-Inverse Document Frequency) для обучения классических ML-моделей в решении нашей задачи, так как этот метод извлекает важность слов в документе на основе их частоты в документе и обратной частоты встречаемости в коллекции документов. Он относительно прост в реализации и работает быстро, а также применим при отсутствии большой обучающей выборки.

При обучении нейросетевых моделей была использована библиотека Transformers и проведен fine-tuning на 3 эпохах. Авторы взяли рекомендованные для BERT [9] гиперпараметры, расширив диапазон значений. Так, для каждой модели был проведен цикл обучения по 11 значениям learning rate

и 3 seed. В параметрах обучения для последовательностей токенов более 512 был добавлен параметр truncate, который обрезает входные данные до установленной длины.

Модель ChatGPT3.5 предполагает использование API, через который в модель подается промпт с запросом (листинг 1).

Для проведения эксперимента была использована библиотека LangChain, ее методы позволяют передать соответствующий промпт, параметры и задать формат выходных данных. Соответствующий код для проведения классификации с помощью модели ChatGPT3.5 представлен в листинге 2.

## Результаты

Эксперимент был проведен в среде Google Colab<sup>1</sup> с подключенным Tensor Processing Units (TPUs) и использованием

<sup>1</sup> URL: [https://colab.research.google.com/drive/1LR80Edjc7kcjEF1vivD64cf1HiZEhfe\\_?usp=sharing](https://colab.research.google.com/drive/1LR80Edjc7kcjEF1vivD64cf1HiZEhfe_?usp=sharing) (дата обращения: 27.03.2024).

### Листинг 1. Промпт для модели ChatGPT

#### Listing 1. ChatGPT prompt

```
template_string = """You are experienced AI HR assistant. \
Take resume below and classify it in one of the following classes: HR, DESIGNER, INFORMATION-TECHNOLOGY, TEACHER, ADVOCATE, BUSINESS-DEVELOPMENT, \
HEALTHCARE, FITNESS, AGRICULTURE, BPO, SALES, CONSULTANT, DIGITAL-MEDIA, AUTOBIOLE, CHEF, FINANCE, APPAREL, ENGINEERING, ACCOUNTANT, CONSTRUCTION, PUBLIC-RELATIONS, \
BANKING, ARTS, AVIATION

Take the resume description below delimited by triple backticks and use it to create the label of class for the resume.

resume: ```{resume}```

Think step by step before giving an answer

{format_instructions}
"""
```

### Листинг 2. Langchain-метод для подачи запроса LLM

#### Listing 2. Langchain method to ask LLM

```
from langchain.prompts import PromptTemplate, ChatPromptTemplate, HumanMessagePromptTemplate
from langchain.chat_models import ChatOpenAI

llm = ChatOpenAI(model_name="gpt-3.5-turbo-0613", temperature=0.0)

prompt = ChatPromptTemplate(
    messages=[
        HumanMessagePromptTemplate.from_template(template_string)
    ],
    input_variables=["resume"],
    partial_variables={"format_instructions": format_instructions},
    output_parser=output_parser
)
```

NVIDIA V100 (32 GB) TPUs. Анализ результатов представлен в таблице 2.

По результатам моделирования можно сделать вывод, что BERT и его производные (ALBERT, DistilBERT) значительно превосходили как традиционные модели машинного обучения (Random Forest и SVM), так и одну из передовых на сегодняшний день LLM-моделей с точки зрения выбранных метрик качества.

Анализ поведения функции потерь для разных нейросетевых моделей, представленный на рисунке 3, позволяет сделать вывод, что у всех моделей есть фаза быстрого снижения и последующего роста. Однако модель DistilBERT имеет постепенное снижение потерь, которое делает потом резкий скачок, что можно связать с архитектурными изменениями, внесенными в DistilBERT, такие как методы сокращения параметров или дистилляции знаний, которые могут влиять на его чувствительность к скорости обучения. Наблюдаемое поведение всех моделей подчеркивает важность тщательного выбора

и настройки гиперпараметров, включая скорость обучения, для каждой архитектуры модели.

Среднее значение показателей точности обучения в зависимости от learning rate с визуализацией стандартного отклонения (рис. 4) отображает зависимость точности обучения от параметра learning rate. Первоначально по мере увеличения learning rate модели могут быстрее сходиться или выходить за пределы локальных минимумов из-за более крупных обновлений параметров модели, что приводит к повышению точности, поскольку модели более эффективно обучаются на данных. После определенного момента слишком сильное увеличение скорости обучения может привести к нестабильности процесса оптимизации. Данное поведение характерно для всех моделей bert-like с меньшей выраженностью для модели DistilBERT, что свидетельствует о ее более стабильном обучении.

Авторы также отметили скорость обучения моделей как важный параметр (рис. 5).

Таблица 2. Результат обучения моделей

Table 2. Modeling results

| Model                  | Accuracy    | Precision   | Micro-F1    |
|------------------------|-------------|-------------|-------------|
| SGDClassifier          | 0,72        | 0,69        | 0,69        |
| KNN                    | 0,55        | 0,55        | 0,57        |
| SVC                    | 0,63        | 0,58        | 0,63        |
| LinearSVC              | 0,62        | 0,62        | 0,62        |
| KNeighborsClassifier   | 0,55        | 0,49        | 0,53        |
| RandomForestClassifier | 0,54        | 0,53        | 0,53        |
| OneVsRest              | 0,58        | 0,56        | 0,62        |
| BERT                   | 0,87        | 0,87        | 0,87        |
| ALBERT                 | 0,887       | 0,887       | 0,887       |
| RoBERTa                | 0,887       | 0,887       | 0,887       |
| DistilBERT             | <b>0,88</b> | <b>0,88</b> | <b>0,88</b> |
| ChatGPT3.5             | 0,7         | 0,7         | 0,7         |

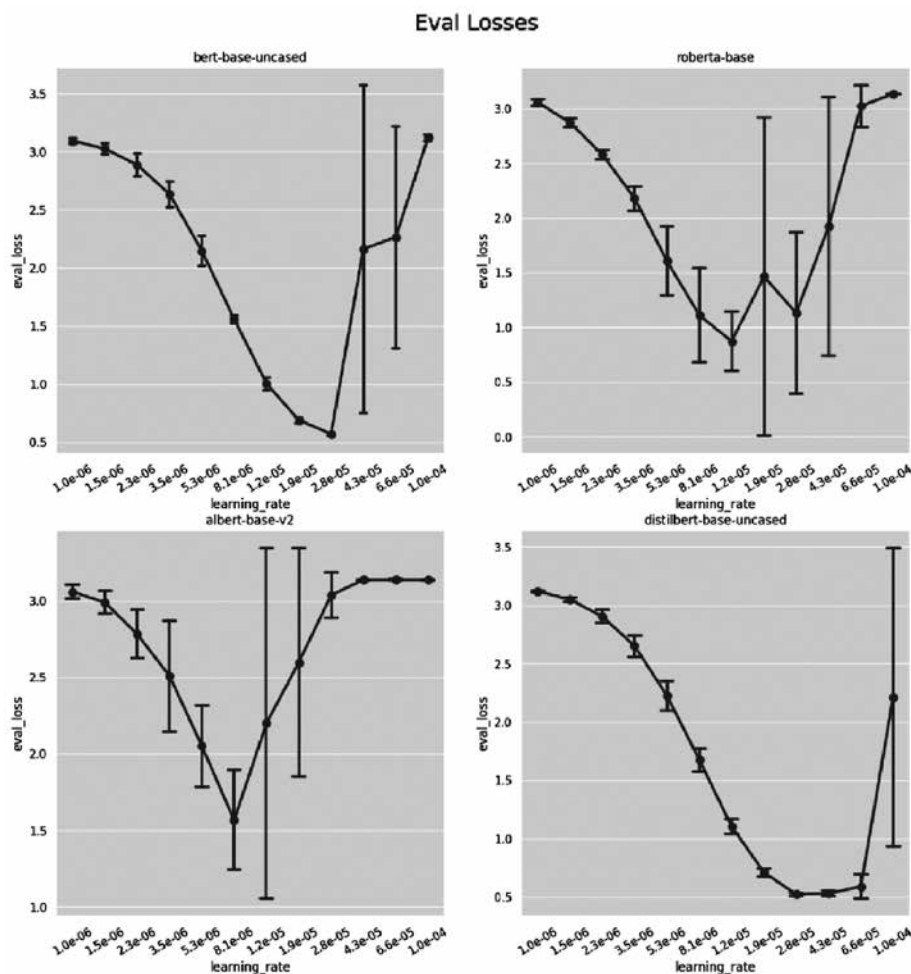


Рис. 3. Графики потерь для четырех базовых BERT-моделей

Fig. 3. Plots of base BERT models

Поскольку модели BERT и DistilBERT дают близкие метрики качества ( $\pm 1\%$ ), то целесообразно использовать в практическом применении модель, которая быстрее может дообучаться.

Таким образом, используя дообученную модель DistilBERT с подобранным гиперпараметром learning rate, авторы добились более эффективного решения прикладной задачи классификации резюме по сравнению со стандартными подходами, используемыми в рекрутинге.

## Обсуждение

Результаты показывают, что максимальная точность, которую достигают базовые модели, 0,72 для SGDClassifier. Модель DistilBERT, которая была нами дообучена с использованием Google Cloud Platform, классифицирует резюме по категориям с точностью 0,88. Оптимальные параметры обучения составили: max\_seq\_length = 512, batch\_size = 6, learning =  $1 \times e^{-9}$ , длительность – 3 эпохи, что позволяет улучшить точность линейных моделей на 24%.

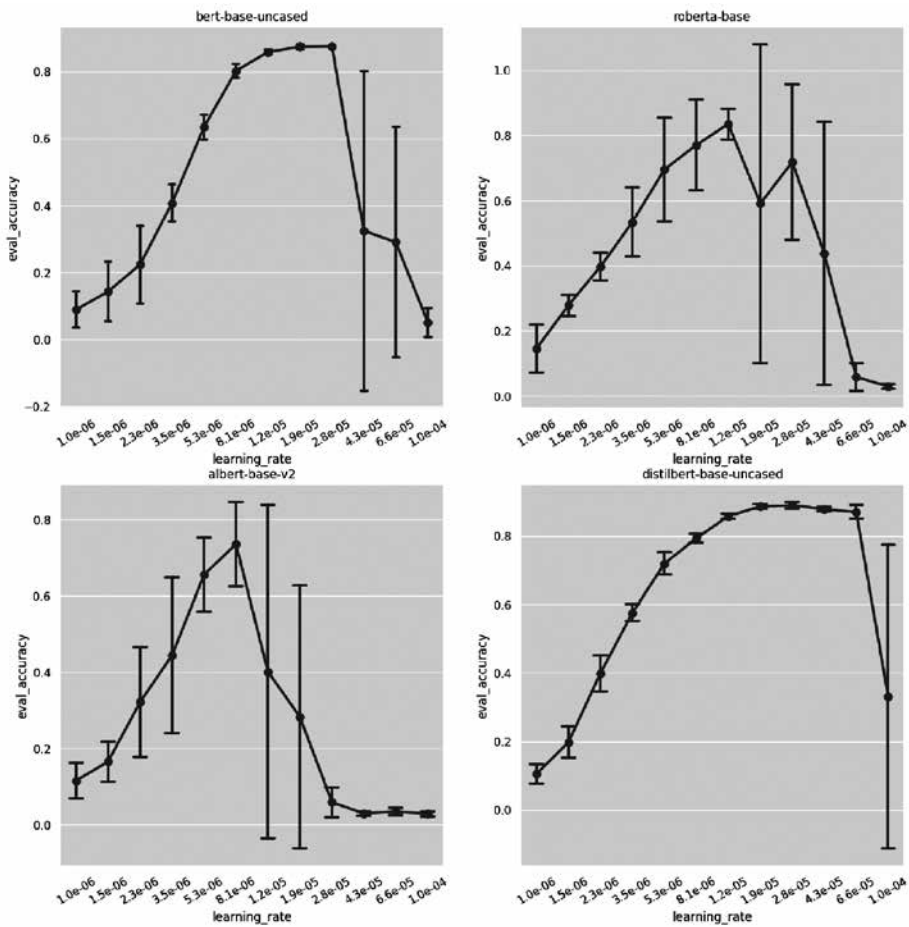


Рис. 4. Среднее значение eval\_accuracy по моделям BERT

Fig. 4. Mean eval\_accuracy for BERT models

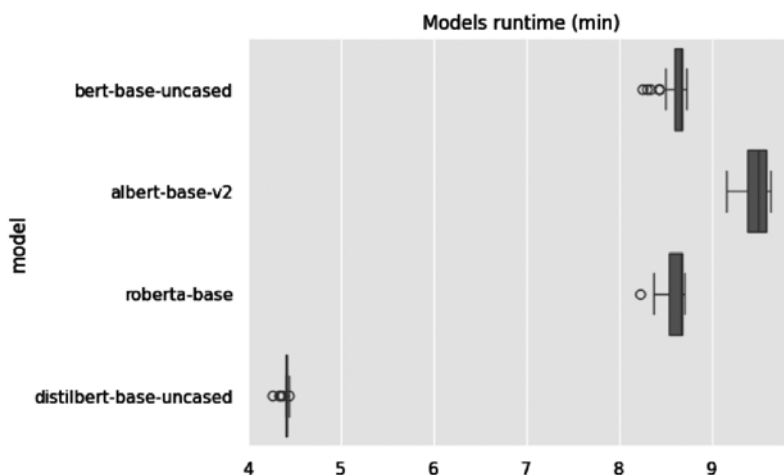


Рис. 5. Время обучения по моделям BERT

Fig. 5. Runtime for BERT models