DOI: 10.37791/2687-0649-2024-19-2-10-22

Повышение эффективности методов подбора персонала на основе глубоких нейронных сетей

Л.А. Комарова[1*]***, А.Д. Черемухин***[2]

[1]*Финансовый университет при Правительстве Российской Федерации, Москва, Россия*

[2]*Нижегородский государственный инженерно-экономический университет, Княгинино, Россия*

[*]*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*[1*], *A. Cheremuhin*[2]**
[1]*Financial University under the Government of Russian Federation, Moscow, Russia*
[2]*Nizhniy 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 эффективности 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 (эмбеддингов) [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 accuracy, 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; разработана 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[1], 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).

[1] 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 ***Model parameters***	**BERT**	**Roberta**	**DistilBERT**	**ALBERT**
Параметры обучения, млн	110	125	66	60
Слои / скрытые слои / головы	12/768/12	12/768/12	6/768/12	12/768/12
Данные для обучения	Book corp + Eng. Wikipedia	Bert + CCNews + OpenWebText + 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

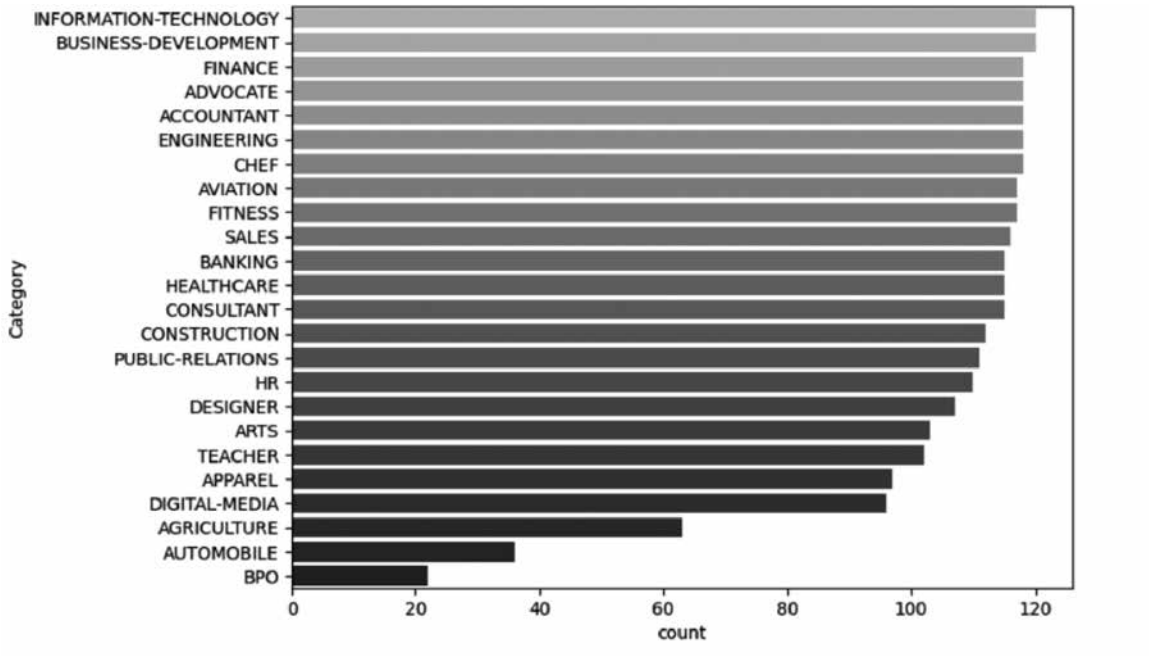

Рис. 1. Распределение резюме по категориям
Fig. 1. CV distribution by categories

	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 Dedica...	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 модели проводился подбор гиперпараметров с использованием библиотеки Optuna [1]. Для решения задачи классификации текстов в таких моделях машинного обучения в качестве признаков целесообразно использовать эмбеддинги слов. Одним из подходов получения эмбеддингов 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 входных 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[1] с подключенным Tensor Processing Units (TPUs) и использованием

[1] URL: 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, AUTOMOBILE, 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 превзошли 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	**0,88**	**0,88**	**0,88**
ChatGPT3.5	0,7	0,7	0,7

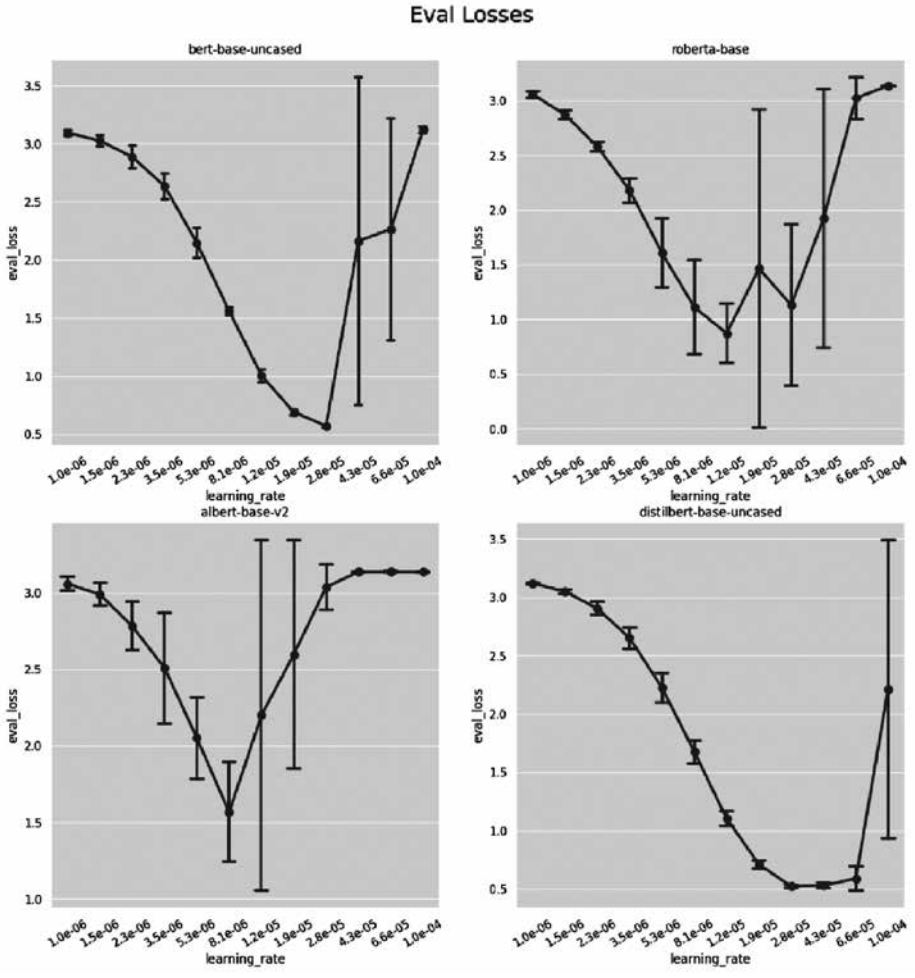

Рис. 3. Графики потерь для четырех базовых BERT-моделей
Fig. 3. Plots of base BERT models

Поскольку модели BERT и DistilBERT дают близкие метрики качества (±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%.

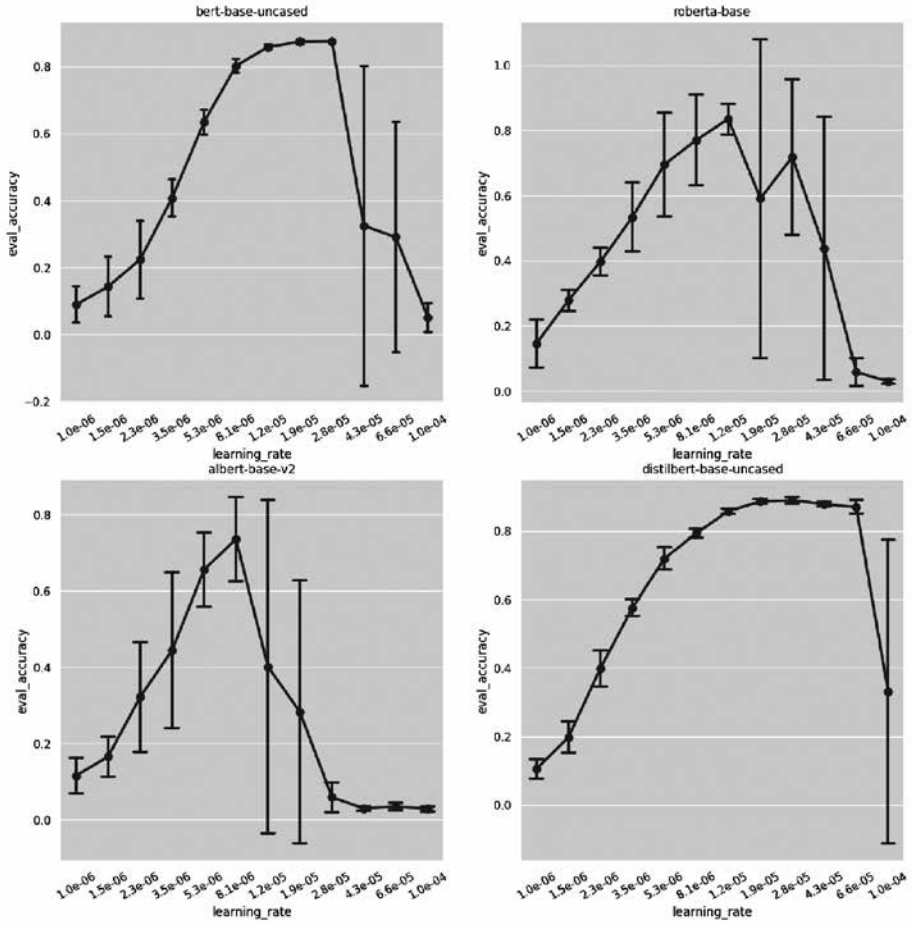

Рис. 4. Среднее значение eval_accuracy по моделям BERT

Fig. 4. Mean eval_accuracy for BERT models

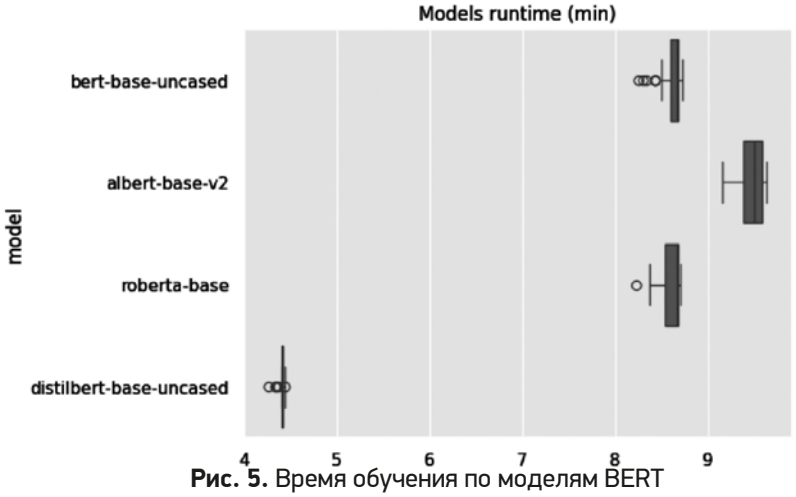

Рис. 5. Время обучения по моделям BERT

Fig. 5. Runtime for BERT models

Заключение

Это исследование является вводным в решение задачи эффективного подбора кандидатов на позицию и позволило выявить эффективные подходы в продвижении области классификации резюме за счет использования возможностей искусственного интеллекта, в частности за счет использования нейронных сетей на основе BERT-архитектуры. Данное исследование подтверждает эффективность системы глубокого обучения для решения поставленной задачи, предлагая новый путь для прескрининга при подборе персонала с помощью ИИ, который может более эффективно отбирать подходящих кандидатов. Дальнейшая работа авторами планируется проводиться уже на базах данных с русскоязычными резюме.

Список литературы

1. *Akiba T., Sano S., Yanase T., Ohta T., Koyama M.* Optuna: A next-generation hyperparameter optimization framework // arXiv:1907.10902. 25 Jul 2019. URL: https://arxiv.org/pdf/1907.10902.pdf (дата обращения: 27.03.2024).
2. *Bhuyan S. S., Mahanta S. K., Pakray P., Favre B*. Textual entailment as an evaluation metric for abstractive text summarization // Natural Language Processing Journal. 2023. Vol. 4. Article 100028. DOI: 10.1016/j.nlp.2023.100028.
3. *Bocharova M. Y*., *Malakhov E. V., Mezhuyev V. I*. VacancySBERT: The approach for representation of titles and skills for semantic similarity search in the recruitment domain // Applied Aspects of Information Technology. 2023. Vol. 6. No. 1. P. 52–59. DOI: 10.15276/aait.06.2023.4.
4. *Briskilal J., Subalalitha C. N*. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa // Information Processing & Management. 2022. Vol. 59. No. 1. Article 102756. DOI: 10.1016/j.ipm.2021.102756.
5. *Cai P., Chen X., Jin P., Wang H., Li T.* Distributional discrepancy: A metric for unconditional text generation // Knowledge-Based Systems. 2021. Vol. 217. Article 106850. DOI: 10.1016/j.knosys.2021.106850.
6. *De Mauro A., Greco M., Grimaldi M., Ritala P.* Human resources for Big Data professions: A systematic classification of job roles and required skill sets // International Processing & Management. 2018. Vol. 54. No. 5. P. 807–817. DOI: 10.1016/j.ipm.2017.05.004.
7. *De Mauro A., Greco M., Grimaldi M., Nobili G.* Beyond data scientists: A review of big data skills and job families // International Forum on Knowledge Asset Dynamics (IFKAD 2016), 2016. URL: https://www.researchgate.net/publication/305109030_Beyond_Data_Scientists_a_Review_of_Big_Data_Skills_and_Job_Families (дата обращения: 27.03.2024).
8. *Deng Y., Eden M. R., Cremaschi S*. Metrics for evaluating machine learning models prediction accuracy and uncertainty // Computer Aided Chemical Engineering. 2023. Vol. 52. P. 1325–1330. DOI: 10.1016/B978-0-443-15274-0.50211-0.
9. *Devlin J., Chang M.-W., Lee K., Toutanova K*. BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv:1810.04805. 24 May 2019. URL: https://arxiv.org/pdf/1810.04805.pdf (дата обращения: 27.03.2024).
10. *Ding M., Zhou C., Yang H., Tang J.* CogLTX: Applying bert to long texts. – In: Advances in Neural Information Processing Systems. 2020. No. 33. P. 12792–12804.
11. *El Mohadab M., Bouikhalene B., Safi S.* Automatic CV processing for scientific research using data mining algorithm // Journal of King Saud University – Computer and Information Sciences. 2020. Vol. 32. No. 5. P. 561–567. DOI: 10.1016/j.jksuci.2018.07.002.
12. *Giray S.* Prompt engineering with ChatGPT: A guide for academic writers // Annals of Biomedical Engineering. 2023. Vol. 51. No. 12. P. 2629–2633. DOI: 10.10017/s10439-023-03272-4.
13. *Kim S.-W., Gil J.-M.* Research paper classification systems based on TF-IDF and LDA schemes // Human-centric Computing and Information Sciences. 2019. No. 9. Article 30. DOI: 10.1186/s13673-019-0192-7.
14. *Kmail A., Maree M., Belkhatir M*. MatchingSem: online recruitment system based on multiple semantic resources // Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'15). 2015. P. 2654–2659. DOI: 10.1109/FSKD.2015.7382376.

15. *Nasser S., Sreejith C., Irshad M*. Convolutional neural network with word embedding based approach for resume classification // Proceedings of the 2018 International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR). 2018. P. 1–6. DOI: 10.1109/ICETIETR.2018.8529097.
16. *Pal R., Shaikh S., Satpute S., Bhagwat S.* Resume classification using various machine learning algorithms // ITM Web of Conferences. 2022. No. 44. Article 03011. DOI: 10.1051 20224403011.
17. *Roy P. K., Chahar S*. N-gram feature based resume classification using machine learning. – In: Communications in Computer and Information Science. 2022. Vol. 1579. P. 239–251. DOI: 10.1007/978-3-031-10766-5_18.
18. *Roy P.K., Chowdhary S.S., Bhatia R*. A machine learning approach for automation of resume recommendation system // Procedia Computer Science. 2020. Vol. 167. P. 2318–2327. DOI: 10.1016/j.procs.2020.03.284.
19. *Sayfullina L., Malmi E., Liao Y., Jung A*. Domain adaptation for resume classification using convolutional neural networks // arXiv:1707.05576. 18 July 2017. URL: https://arxiv.org/pdf/1707.05576.pdf (дата обращения: 27.03.2024).
20. *Sharma N*. Job Skills extraction with LSTM and Word Embeddings. – Sydney: University of Technology Sydney (UTS), 2019. – 5 p.
21. *Solbiati A., Heffernan K., Damaskinos G., Poddar S., Modi S., Cali J.* Unsupervised topic segmentation of meetings with BERT Embeddings // arXiv:2106.12978. 24 Jun 2021. URL: https://arxiv.org/pdf/2106.12978.pdf (дата обращения: 27.03.2024).
22. *Sun X., Li X., Li J., Wu F., Guo S., Zhang T., Wang G.* Text classification via large language models // Findings of the Association for Computational Linguistics: EMNLP 2023. 2023. P. 8990–9005. DOI: 10.18653/v1/2023.findings-emnlp.603.
23. *Ternikov A. A*. Skill-based clustering algorithm for online job advertisements // Izvestiya of Saratov University. Mathematics. Mechanics. Informatics. 2022. Vol. 22. No. 2. P. 250–265. DOI: 10.18500/1816-9791-2022-22-2-250-265.
24. *Thinh T.-T.-Q., Chung Y.-C., Kuo R. J.* A domain adaptation approach for resume classification using graph attention networks and natural language processing // Knowledge-Based Systems. 2023. Vol. 266. Article 110364. DOI: 10.1016/j.knosys.2023.110364.
25. *Wang H., Yang W., Li J., Ou J., Song Y., Chen Y.-W.* An improved heterogeneous graph convolutional network for job recommendation // Engineering Applications of Artificial Intelligence. 2023. Vol. 126. Article 107147. DOI: 10.1016/j.engappai.2023.107147.

Сведения об авторах

Комарова Любовь Александровна, ORCID 0000-0001-5277-8234, аспирант, кафедра анализа данных и машинного обучения, Финансовый университет при Правительстве Российской Федерации, Москва, Россия, 229338@edu.fa.ru

Черемухин Артем Дмитриевич, ORCID 0000-0003-4076-5916, канд. экон. наук, доцент кафедры математики и вычислительной техники, Нижегородский государственный инженерно-экономический университет, Княгинино, Россия, ngieu.cheremuhin@yandex.ru

Статья поступила 18.12.2023, рассмотрена 16.01.2024, принята 15.02.2024

References

1. Akiba T., Sano S., Yanase T., Ohta T., Koyama M. Optuna: A next-generation hyperparameter optimization framework. arXiv:1907.10902, 25 Jul 2019. Available at: https://arxiv.org/pdf/1907.10902.pdf (accessed 27.03.2024).
2. Bhuyan S. S., Mahanta S. K., Pakray P., Favre B. Textual entailment as an evaluation metric for abstractive text summarization. Natural Language Processing Journal, 2023, vol.4, article 100028. DOI: 10.1016/j.nlp.2023.100028.
3. Bocharova M. Y., Malakhov E. V., Mezhuyev V. I. VacancySBERT: The approach for representation of titles and skills for semantic similarity search in the recruitment domain. Applied Aspects of Information Technology, 2023, vol.6, no.1, pp.52-59. DOI: 10.15276/aait.06.2023.4.

4. Briskilal J., Subalalitha C. N. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. Information Processing & Management, 2022, vol.59, no.1, article 102756. DOI: 10.1016/j.ipm.2021.102756.
5. Cai P., Chen X., Jin P., Wang H., Li T. Distributional discrepancy: A metric for unconditional text generation. Knowledge-Based Systems, 2021, vol.217, article 106850. DOI: 10.1016/j.knosys.2021.106850.
6. De Mauro A., Greco M., Grimaldi M., Ritala P. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. International Processing & Management, 2018, vol.54, no.5, pp.807-817. DOI: 10.1016/j.ipm.2017.05.004.
7. De Mauro A., Greco M., Grimaldi M., Nobili G. Beyond data scientists: A review of big data skills and job families. International Forum on Knowledge Asset Dynamics (IFKAD 2016), 2016. Available at: https://www.researchgate.net/publication/305109030_Beyond_Data_Scientists_a_Review_of_Big_Data_Skills_and_Job_Families (accessed 27.03.2024).
8. Deng Y., Eden M. R., Cremaschi S. Metrics for evaluating machine learning models prediction accuracy and uncertainty. Computer Aided Chemical Engineering, 2023, vol.52, pp.1325-1330. DOI: 10.1016/B978-0-443-15274-0.50211-0.
9. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 24 May 2019. Available at: https://arxiv.org/pdf/1810.04805.pdf (accessed 27.03.2024).
10. Ding M., Zhou C., Yang H., Tang J. CogLTX: Applying bert to long texts. In: Advances in Neural Information Processing Systems, 2020, no.33, pp.12792-12804.
11. El Mohadab M., Bouikhalene B., Safi S. Automatic CV processing for scientific research using data mining algorithm. Journal of King Saud University – Computer and Information Sciences, 2020, vol.32, no.5, pp.561-567. DOI: 10.1016/j.jksuci.2018.07.002.
12. Giray S. Prompt engineering with ChatGPT: A guide for academic writers. Annals of Biomedical Engineering, 2023, vol.51, no.12, pp.2629-2633. DOI: 10.10017/s10439-023-03272-4.
13. Kim S.-W., Gil J.-M. Research paper classification systems based on TF-IDF and LDA schemes. Human-centric Computing and Information Sciences, 2019, no.9, article 30. DOI: 10.1186/s13673-019-0192-7.
14. Kmail A., Maree M., Belkhatir M. MatchingSem: online recruitment system based on multiple semantic resources. Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'15), 2015, pp.2654-2659. DOI: 10.1109/FSKD.2015.7382376.
15. Nasser S., Sreejith C., Irshad M. Convolutional neural network with word embedding based approach for resume classification. Proceedings of the 2018 International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR), 2018, pp.1-6. DOI: 10.1109/ICETIETR.2018.8529097.
16. Pal R., Shaikh S., Satpute S., Bhagwat S. Resume classification using various machine learning algorithms. ITM Web of Conferences, 2022, no.44, article 03011. DOI: 10.1051 20224403011.
17. Roy P. K., Chahar S. N-gram feature based resume classification using machine learning. – In: Communications in Computer and Information Science, 2022, vol.1579, pp.239-251. DOI: 10.1007/978-3-031-10766-5_18.
18. Roy P.K., Chowdhary S.S., Bhatia R. A machine learning approach for automation of resume recommendation system. Procedia Computer Science, 2020, vol.167, pp.2318-2327. DOI: 10.1016/j.procs.2020.03.284.
19. Sayfullina L., Malmi E., Liao Y., Jung A. Domain adaptation for resume classification using convolutional neural networks. arXiv:1707.05576, 18 July 2017. Available at: https://arxiv.org/pdf/1707.05576.pdf (accessed 27.03.2024).
20. Sharma N. Job Skills extraction with LSTM and Word Embeddings. Sydney, University of Technology Sydney (UTS), 2019, 5 p.
21. Solbiati A., Heffernan K., Damaskinos G., Poddar S., Modi S., Cali J. Unsupervised topic segmentation of meetings with BERT Embeddings. arXiv:2106.12978, 24 Jun 2021. Available at: https://arxiv.org/pdf/2106.12978.pdf (accessed 27.03.2024).
22. Sun X., Li X., Li J., Wu F., Guo S., Zhang T., Wang G. Text classification via large language models. Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp.8990-9005. DOI: 10.18653/v1/2023.findings-emnlp.603.

23. Ternikov A. A. Skill-based clustering algorithm for online job advertisements. Izvestiya of Saratov University. Mathematics. Mechanics. Informatics, 2022, vol.22, no.2, pp.250-265. DOI: 10.18500/1816-9791-2022-22-2-250-265.
24. Thinh T.-T.-Q., Chung Y.-C., Kuo R. J. A domain adaptation approach for resume classification using graph attention networks and natural language processing. Knowledge-Based Systems, 2023, vol.266, article 110364. DOI: 10.1016/j.knosys.2023.110364.
25. Wang H., Yang W., Li J., Ou J., Song Y., Chen Y.-W. An improved heterogeneous graph convolutional network for job recommendation. Engineering Applications of Artificial Intelligence, 2023, vol.126, article 107147. DOI: 10.1016/j.engappai.2023.107147.

About the authors

Lyubov A. Komarova, ORCID 0000-0001-5277-8234, Postgraduate, Data Analysis and Machine Learning Department, Financial University under the Government of Russian Federation, Moscow, Russia, 229388@edu.fa.ru

Artem D. Cheremuhin, ORCID 0000-0003-4076-5916, Cand. Sci. (Econ.), Associate Professor at Mathematics and Computer Science Department, Nizhniy Novgorod State Engineering-Economic University, Knyaginino, Russia, ngieu.cheremuhin@yandex.ru

Received 18.12.2023, reviewed 16.01.2024, accepted 15.02.2024