МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра математической кибернетики и компьютерных наук

РАЗРАБОТКА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ УРОВНЯ КОМПЕТЕНЦИЙ ПО РЕЗЮМЕ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные технологии
факультета КНиИТ
Прохорова Максима Александровича

Научный руководитель
доцент, к. ф.-м. н. _____ М. И. Сафрончик

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2026

Актуальность работы. В условиях современного рынка труда, характеризующегося высокой динамикой, дефицитом квалифицированных кадров в ряде отраслей и постоянно растущими требованиями к профессиональным компетенциям, эффективный подбор персонала становится одним из ключевых факторов конкурентоспособности компаний. Традиционный процесс рекрутинга, основанный на ручном анализе резюме, характеризуется существенными ограничениями, включающими высокую трудоёмкость, значительные временные затраты и неизбежную субъективность принимаемых решений. Рекрутер вынужден просматривать десятки и сотни резюме, сопоставляя описанный опыт кандидатов с требованиями вакансий, при этом оценка уровня владения профессиональными компетенциями может существенно варьироваться в зависимости от опыта, квалификации и даже текущего состояния конкретного специалиста. Дополнительным усложняющим фактором является несовпадение формулировок навыков в резюме кандидатов с терминологией, принятой в компании-работодателе.

Решение данной задачи позволяет не только кратно сократить время первичного скрининга резюме, но и обеспечить повышение объективности и стандартизации принимаемых кадровых решений за счёт применения единых критериев оценки для всех кандидатов. Данная работа выполнена в рамках программы «Стартап как диплом», что предполагает не только теоретическую проработку вопроса, но и создание готового к практическому внедрению программного продукта, ориентированного на использование в HR-отделах компаний различного масштаба.

Цель и задачи работы. Целью работы является разработка системы автоматической классификации уровней профессиональных компетенций на основе анализа текстов резюме с использованием предобученной языковой модели RuBERT. Для достижения поставленной цели в рамках работы необходимо решить следующие задачи. Во-первых, провести систематический анализ существующих методов анализа текстовых данных, архитектур нейронных сетей, применяемых для решения задач классификации текста, а также стратегий дообучения предобученных языковых моделей. Во-вторых, выполнить обзор и обоснованный выбор инструментальных средств для реализации системы, включая библиотеки глубокого обучения PyTorch и Transformers. В-третьих, создать сбалансированный датасет для обучения

модели, включающий различные профессиональные компетенции с тремя уровнями владения. В-четвёртых, реализовать и обучить модель классификации уровней компетенций на базе архитектуры RuBERT-base-cased. В-пятых, провести экспериментальное исследование влияния различных стратегий заморозки слоёв модели на итоговое качество классификации. В-шестых, реализовать специализированную функцию потерь, учитывающую упорядоченность классов. В-седьмых, оценить качество разработанной модели с использованием стандартных метрик классификации и провести сравнительный анализ с базовыми подходами, включая логистическую регрессию с векторизацией TF-IDF.

Структура выпускной квалификационной работы. Выпускная квалификационная работа состоит из введения, двух глав, заключения, списка использованных источников и приложений. В первой главе рассматриваются теоретические аспекты анализа текстовых данных, включая обзор методов векторного представления текста, архитектур нейронных сетей для обработки естественного языка, механизма внимания и его роли в архитектуре Transformer, а также принципов построения и дообучения двунаправленных энкодерных моделей семейства BERT. Особое внимание уделяется русскоязычной адаптации BERT — модели RuBERT, её архитектурным особенностям, модификациям и области применения. Во второй главе описывается практическая реализация системы классификации уровней компетенций: приводится описание исходного датасета и процедуры его подготовки, включая токенизацию и формирование входных данных, детально рассматриваются проведённые эксперименты по исследованию стратегий заморозки слоёв модели, реализация функции потерь DistanceAwareLoss и процедура обучения с использованием оптимизатора AdamW. Также во второй главе представлены результаты оценки качества модели с использованием метрик accuracy, precision, recall и F1-меры, выполнен анализ ошибок классификации и проведено сравнение разработанного подхода с базовыми моделями — логистической регрессией с векторизацией TF-IDF и базовым RuBERT без дообучения. В заключении сформулированы основные выводы по результатам выполненной работы, обобщены полученные экспериментальные результаты и намечены направления для дальнейшего развития системы. Список использованных источников включает работы

по обработке естественного языка, глубокому обучению и смежным областям. Приложения содержат листинги ключевых программных модулей и дополнительные иллюстративные материалы.

Функциональная структура системы. Разрабатываемый программный комплекс представляет собой интеллектуальную систему анализа текстовой информации, реализующую задачу классификации уровня владения профессиональной компетенцией на основе текстового описания опыта кандидата, извлечённого из резюме. Пользователь инициирует процесс анализа через веб-интерфейс или программный интерфейс, предоставляя системе текстовое описание опыта кандидата в рамках конкретной профессиональной компетенции. Система выполняет автоматическую обработку входного текста и формирует заключение об уровне владения навыком по трёхбалльной шкале, где каждому уровню соответствует определённая категория: базовое владение (Junior), уверенное владение (Middle) и экспертный уровень (Senior). В отличие от классических алгоритмов фильтрации по ключевым словам, которые не учитывают контекст и семантическую близость формулировок, используемый в работе подход на основе архитектуры Transformer позволяет анализировать текст с учётом связей между словами на различных расстояниях, что обеспечивает высокую точность и гибкость анализа.

Архитектура системы. При проектировании системы был выбран модульный монолитный подход с возможностью последующего развёртывания в виде отдельного микросервиса в составе более крупной платформы. Такой выбор обусловлен относительной простотой развёртывания и эксплуатации монолитного приложения при сохранении возможности горизонтального масштабирования наиболее вычислительно нагруженных компонентов. В рамках разрабатываемого программного комплекса выделены следующие основные компоненты. Модуль подготовки данных осуществляет токенизацию входного текста с использованием штатного токенизатора модели RuBERT, формирование входных тензоров требуемой размерности, а также применение необходимых преобразований, включая паддинг и создание маски внимания. Модуль загрузки предобученной модели обеспечивает инициализацию модели RuBERT-base-cased с предварительно обученными весами, а также добавление классификационной головы для адапта-

ции к задаче трёхклассовой классификации. Модуль обучения и валидации реализует процедуру дообучения модели на специализированном датасете, включая вычисление функции потерь, обратное распространение ошибки и обновление параметров оптимизатора. Модуль инференса предназначен для получения предсказаний на новых данных в режиме реального времени и может быть использован как составная часть более широкой системы анализа резюме.

Выбор базовой модели. Выбор архитектуры RuBERT-base-cased в качестве основы для разрабатываемой системы обоснован рядом факторов. Во-первых, модель предварительно обучена на крупных корпусах русскоязычных данных, включающих новостные статьи, художественную литературу, веб-тексты и другие источники, что обеспечивает глубокое понимание морфологии, синтаксиса и семантики русского языка. Во-вторых, архитектура RuBERT-base-cased включает 12 слоёв трансформерного кодировщика, 12 голов внимания и содержит 180 миллионов параметров, что представляет собой оптимальный баланс между выразительной способностью модели и вычислительными затратами на её обучение и инференс. В-третьих, использование cased-версии обусловлено тем, что регистр символов может нести важную смысловую нагрузку при описании профессиональных навыков, например, при упоминании конкретных языков программирования, названий технологий или фирменных наименований продуктов. В-четвёртых, модель доступна через экосистему Hugging Face Transformers, что существенно упрощает её загрузку, интеграцию и тонкую настройку. Альтернативные варианты, такие как более лёгкая версия RuBERT-tiny, содержащая всего 3 слоя и около 30 миллионов параметров, не обеспечивают необходимого качества анализа из-за ограниченной способности улавливать сложные контекстные зависимости.

Инструменты реализации. Для реализации разработанной системы выбран фреймворк глубокого обучения PyTorch, что обусловлено его широким распространением как в исследовательской, так и в промышленной практике. PyTorch предоставляет динамический вычислительный граф, что упрощает отладку и позволяет гибко изменять архитектуру модели в процессе экспериментов. Кроме того, PyTorch обладает развитой экосистемой инструментов и библиотек, включая Hugging Face Transformers, которая

предоставляет унифицированный высокоуровневый API для загрузки предобученных моделей, их тонкой настройки и выполнения инференса. Использование библиотеки Transformers позволяет существенно сократить объём кода, необходимого для реализации пайплайна обработки, и сосредоточиться на экспериментальной части исследования. Для оптимизации процесса обучения используется библиотека оптимизаторов, включающая реализацию AdamW — адаптивного метода оптимизации с коррекцией моментов и весовой регуляризацией, который хорошо зарекомендовал себя при дообучении трансформерных моделей.

Проектирование экспериментов по заморозке слоёв. Ключевым архитектурным решением, исследованным в рамках данной работы, является выбор оптимальной стратегии дообучения предобученной модели в условиях ограниченного объёма размеченных данных. Предобученная модель RuBERT содержит универсальные знания о русском языке, накопленные в процессе обучения на миллионах документов. Согласно современным представлениям об устройстве трансформерных моделей, нижние слои кодировщика отвечают за извлечение базовых лингвистических признаков, таких как морфологические характеристики слов и синтаксические связи в предложении. Средние слои формируют семантические представления на уровне фраз и предложений. Верхние слои, наиболее близкие к выходу модели, адаптируются под конкретную целевую задачу и содержат наиболее специфичные признаки. Полное дообучение всех слоёв модели требует значительного объёма размеченных данных и может привести к переобучению при малом размере выборки, поскольку модель начинает запоминать шум и специфические формулировки из обучающего набора вместо обобщения закономерностей. В связи с этим в работе проведено систематическое исследование влияния заморозки различного количества слоёв на итоговое качество классификации.

Схема проведения экспериментов. В рамках экспериментального исследования были рассмотрены семь конфигураций заморозки, охватывающих весь диапазон возможных стратегий от минимальной до максимальной адаптации модели. Первая конфигурация предполагает заморозку всех 12 слоёв трансформерного кодировщика, при этом обучается только классификационная голова, добавленная поверх модели. Эта конфигурация соот-

ветствует подходу, при котором модель используется как фиксированный экстрактор признаков, а задача классификации решается обучаемым классификатором на вершине этих признаков. Вторая конфигурация предусматривает заморозку 10 слоёв, при этом обучаются два верхних слоя (слои 10 и 11) и классификационная голова. Третья, четвёртая, пятая и шестая конфигурации предполагают заморозку 8, 6, 4 и 2 слоёв соответственно, что приводит к последовательному увеличению числа обучаемых слоёв от 4 до 10. Седьмая конфигурация представляет собой полное обучение всех слоёв модели без какой-либо заморозки, что соответствует традиционному подходу к тонкой настройке, применяемому при наличии достаточно больших размеченных датасетов.

Для каждой конфигурации выполнялось дообучение модели в течение 10 эпох с использованием оптимизатора AdamW, представляющего собой модификацию стандартного Adam, в которой корректно реализована весовая регуляризация. Значение скорости обучения выбрано равным 2e-5, что является стандартным и эмпирически обоснованным значением для тонкой настройки трансформерных моделей. Размер мини-пакета выбран равным 8, что обусловлено ограничениями объёма видеопамяти графического процессора при работе с моделью, содержащей 180 миллионов параметров.

Реализация функции потерь. Стандартным выбором для многоклассовой классификации является кросс-энтропийная функция потерь, которая минимизирует расхождение между предсказанным моделью распределением вероятностей по классам и истинным распределением. Однако стандартная кросс-энтропия обладает существенным недостатком применительно к задачам с упорядоченными классами: она одинаково штрафует любую ошибку классификации независимо от того, насколько далеко предсказанный класс находится от истинного. В задаче классификации уровней владения компетенциями классы обладают естественным порядком, который отражает прогрессию в освоении навыка: от базового владения через уверенное к экспертному уровню. Ошибка между Junior и Senior является значительно более грубой, чем ошибка между Junior и Middle, поскольку она предполагает принципиально неверную оценку квалификации кандидата.

Для учёта этого порядка в работе реализована специализированная функция потерь, получившая название DistanceAwareLoss. Данная функция модифицирует стандартную кросс-энтропию путём взвешивания ошибки в соответствии с квадратом расстояния между истинным и предсказанным классами. Чем больше расстояние, тем сильнее штраф, что побуждает модель к минимизации грубых ошибок. Параметр масштабирования штрафа был подобран эмпирически и установлен равным 2.0. При таком значении ошибка между Junior и Senior штрафуется в четыре раза сильнее, чем ошибка между соседними уровнями, что стимулирует модель к консервативным предсказаниям в пограничных случаях.

Подготовка данных. Для обучения и последующей оценки качества разработанной модели в рамках работы был сформирован специализированный датасет, содержащий 24 профессиональные компетенции, каждая из которых представлена тремя уровнями владения. Исходные данные, собранные из различных источников, включая открытые базы резюме и экспертные оценки, были подвергнуты предварительной обработке и очистке. Для исследования влияния размера обучающей выборки на итоговое качество модели исходные данные были преобразованы в четыре сбалансированных датасета различного объёма: 500, 1000, 1500 и 2000 строк. Сбалансированность датасетов означает, что каждый из трёх классов представлен примерно равным количеством примеров, что позволяет избежать смещения модели в сторону более частотных классов.

Токенизация текста выполнялась с использованием штатного токенизатора модели RuBERT-base-cased, который разбивает входной текст на субсловные единицы (Byte-Pair Encoding), что позволяет эффективно обрабатывать редкие слова и незнакомые термины. Длина последовательности была фиксирована и составила 128 токенов, что обеспечивает баланс между полнотой контекста, необходимого для анализа относительно коротких описаний опыта, и вычислительной эффективностью. Входные данные формировались по специальному шаблону «Компетенция: X. Описание: Y», где X — наименование профессионального навыка, а Y — текстовое описание опыта кандидата, извлечённое из резюме. Такой формат позволяет модели явно учитывать как название компетенции, так и контекст её проявления в опыте кандидата. Данные были разделены на обучающую, валидационную

и тестовую выборки в пропорции 80/10/10 с использованием стратифицированной выборки, обеспечивающей сохранение распределения классов в каждой из трёх выборок.

Результаты экспериментов по заморозке слоёв. Проведённое экспериментальное исследование продемонстрировало чёткую зависимость качества классификации от выбранной стратегии заморозки слоёв. Наилучший результат был достигнут для конфигурации с заморозкой 10 из 12 слоёв, при которой обучаются только два верхних слоя трансформерного кодировщика и классификационная голова. Точность на тестовой выборке для данной конфигурации составила 84,0%, а значение F1-меры, являющейся гармоническим средним между полнотой и точностью, достигло 0,839. При полном обучении всех слоёв модели без какой-либо заморозки точность классификации снизилась до 78,7%, что объясняется эффектом переобучения: модель с 180 миллионами параметров начинает запоминать шум и специфические формулировки из обучающей выборки объёмом всего 1500 примеров вместо обобщения закономерностей. Конфигурация с заморозкой всех 12 слоёв, при которой обучается только классификационная голова, показала точность 72,0%, что свидетельствует о недостаточности дообучения только финального слоя для адаптации к специфике задачи классификации уровней компетенций. При увеличении числа обучаемых слоёв от 2 до 6 (что соответствует заморозке 6 слоёв) качество классификации стабилизировалось на уровне 82-83%, однако не превзошло результат конфигурации с двумя обучаемыми слоями. Интересно отметить, что конфигурации с заморозкой 8 и 6 слоёв показали практически идентичные результаты с разницей менее одного процентного пункта.

Анализ ошибок классификации. Детальный анализ матрицы ошибок, построенной для оптимальной конфигурации модели, показал важную закономерность: все ошибки классификации происходят только между соседними уровнями — Junior и Middle или Middle и Senior. Ошибок между Junior и Senior, то есть между крайними уровнями шкалы, зафиксировано не было. Этот результат подтверждает эффективность применения функции потерь DistanceAwareLoss, которая благодаря штрафованию ошибок пропорционально квадрату расстояния успешно минимизирует вероятность грубых ошибок через несколько уровней.

Такое поведение является естественным даже для человека-эксперта, поскольку граница между уровнями владения компетенцией в реальной практике часто является размытой и зависит от контекста и конкретных формулировок. Распределение ошибок между двумя типами соседних переходов оказалось приблизительно равномерным, что указывает на отсутствие систематического смещения модели в сторону более высоких или более низких оценок.

Исследование влияния объёма данных. В рамках работы также было проведено исследование влияния размера обучающей выборки на итоговое качество классификации. Для этого были сформированы четыре датасета объёмом 500, 1000, 1500 и 2000 строк, каждый из которых использовался для обучения модели с оптимальной конфигурацией заморозки (10 замороженных слоёв). Результаты показали, что увеличение размера выборки с 500 до 1500 строк приводит к росту точности классификации с 76% до 84%. При этом наиболее существенный прирост качества наблюдался при переходе от 500 к 1000 строкам (прирост около 5 процентных пунктов), тогда как при переходе от 1000 к 1500 строкам прирост составил около 3 процентных пунктов. Дальнейшее увеличение датасета до 2000 строк не дало статистически значимого прироста качества: точность осталась на уровне 84% или изменилась в пределах погрешности измерения. Это позволяет сделать вывод о том, что оптимальный объём обучающих данных для решаемой задачи составляет примерно 1500 примеров. Дальнейшее увеличение размера выборки нецелесообразно с точки зрения соотношения затрат на сбор и разметку данных и получаемого улучшения качества модели.

Сравнение с базовыми моделями. Для обоснованного вывода о качестве разработанной модели было проведено сравнение с двумя базовыми подходами. Первым базовым подходом является логистическая регрессия с векторизацией текста методом TF-IDF. Данный подход представляет собой классический метод машинного обучения для задач классификации текста, не использующий нейросетевые архитектуры и предобученные языковые модели. Векторизация TF-IDF преобразует текст в разреженный вектор, компоненты которого отражают важность каждого слова в данном документе относительно корпуса документов в целом.

Логистическая регрессия, обученная поверх таких признаков, достигла точности 73,3%, что является достойным результатом для классического подхода, однако уступает разработанной модели на более чем 10 процентных пунктов. Вторым базовым подходом является использование базовой модели RuBERT без какого-либо дообучения, то есть в том виде, в котором она была выпущена после предобучения на общих корпусах русских текстов. Для получения предсказаний использовалась классификационная голова, инициализированная случайными весами. Данный подход показал точность 30,7%, что лишь незначительно превышает вероятность случайного угадывания для трёх классов, составляющую 33,3%. Этот результат демонстрирует необходимость дообучения модели на целевой задаче.

Практическая значимость. Разработанная в рамках выпускной квалификационной работы система автоматической классификации уровней профессиональных компетенций обладает высокой практической значимостью и может быть использована для автоматизации первичного скрининга резюме в HR-процессах компаний различного масштаба. Внедрение системы позволяет кратно сократить временные затраты рекрутеров на ручной анализ резюме, стандартизировать процесс оценки уровня владения профессиональными навыками и снизить влияние субъективного фактора на принимаемые кадровые решения. Система интегрирована в программный комплекс «HR-Интеллект» и может функционировать как в составе распределённой микросервисной архитектуры, так и в качестве самостоятельного модуля для обработки текстовых описаний компетенций. Полученные в работе экспериментальные результаты, включая оптимальную стратегию заморозки слоёв (10 из 12) и оптимальный объём обучающих данных (1500 примеров), могут быть использованы в качестве практических рекомендаций при разработке аналогичных систем классификации текстов.

Основные результаты работы. В рамках выпускной квалификационной работы разработана система автоматической классификации уровней профессиональных компетенций на основе анализа текстов резюме с использованием предобученной языковой модели RuBERT-base-cased. Проведён систематический анализ существующих методов анализа текстовых данных, включая классические подходы на основе векторного представ-

ления текста и методы глубокого обучения на основе рекуррентных нейронных сетей и архитектуры Transformer. Детально рассмотрены механизм внимания, лёгший в основу архитектуры Transformer, его модификация Self-Attention, а также принципы построения двунаправленных энкодерных моделей семейства BERT. Исследованы стратегии дообучения предобученных языковых моделей, включая полное дообучение, заморозку слоёв и добавление специализированных классификационных голов. Экспериментально обоснован выбор оптимальной конфигурации заморозки, предполагающей заморозку 10 из 12 слоёв модели, что обеспечивает максимальную точность классификации при ограниченном объёме обучающих данных. Реализована и апробирована специализированная функция потерь DistanceAwareLoss, учитывающая упорядоченность классов и снижающая вероятность грубых ошибок классификации. Определён оптимальный объём обучающих данных для решаемой задачи, составляющий 1500 примеров, превышение которого не даёт статистически значимого прироста качества.

Разработанная модель достигла точности 84,0% и значения F1-меры 0,839 на задаче различения трёх уровней владения профессиональными компетенциями, что подтверждает практическую применимость предложенного подхода для автоматизации первичного скрининга резюме в HR-процессах. Анализ матрицы ошибок показал, что все ошибки классификации происходят только между соседними уровнями, при этом грубые ошибки между крайними уровнями полностью отсутствуют. Сравнение с базовыми моделями подтвердило преимущество разработанного подхода. Разработанная модель интегрирована в программный комплекс «HR-Интеллект» и может быть использована в качестве отдельного сервиса для оценки уровня владения профессиональными навыками на основе текстового описания опыта кандидата. Разработанный программный продукт соответствует формату «Стартап как диплом» и представляет собой готовое к практическому внедрению решение, ориентированное на использование в HR-отделах компаний различного масштаба для автоматизации процессов первичной оценки и отбора кандидатов.