Информационные процессы, Том 24, № 2, 2024, стр. 149–162.

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ТЕХНИЧЕСКИХ И СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМАХ

Исследование взаимосвязи онлайн-активности пользователя социальной сети с его поведенческими реакциями

К.Р. Мудрова[*], А.Н. Жданова[*], А.В. Куприянов[*,],**

[*] *Самарский национальный исследовательский университет им. академика С.П. Королева*
[**] *Институт систем обработки изображений, НИЦ «Курчатовский институт»*

Поступила в редколлегию 01.07.2024 г. Принята 01.08.2024 г.

Аннотация—В данной работе предложена технология анализа цифрового следа пользователя социальной сети «ВКонтакте», основанная на алгоритмах определения проявления отклоняющегося поведения в письменной речи и классификации текстовых публикаций с использованием методов машинного обучения. Разработанная технология реализована в виде веб-приложения на основе клиент-серверной архитектуры, которое позволяет авторизованным пользователям (HR-специалистам) получать сформированные и структурированные системой отчеты о характере возможных поведенческих реакций кандидатов на должность при приеме на работу.

КЛЮЧЕВЫЕ СЛОВА: публичные данные, HR-системы, анализ публикаций, обработка текстовых данных, веб-приложение, подбор персонала, машинное обучение.

DOI: 10.53921/18195822_2024_24_2_149

1. ВВЕДЕНИЕ

В стремлении удержаться на глобальном рынке и занять одну из лидирующих позиций крупные компании вынуждены непрерывно искать способы и средства повышения качества производимых ими продуктов, их адаптации согласно меняющимся трендам, а также способов достижения эффективного использования имеющихся в распоряжении компании ресурсов. На пути улучшения производства предприятия могут пытаться сокращать или вовсе отказываться от использования тех или иных материальных и нематериальных активов. Однако ни одна компания не в состоянии полностью отказаться от использования кадрового состава или всего персонала, иначе говоря, человеческих ресурсов – интеллектуального и исполнительного капитала производства, во многом определяющего его текущий потенциал и перспективы дальнейшего развития.

В связи с этим в построении общей стратегии работы компании особое место занимают планирование и грамотное управление человеческими ресурсами: подбор и найм на должности высококвалифицированных специалистов, обеспечение соответствующих условий работы и оплаты труда, индивидуальный и совокупный анализ результатов сотрудников. Поддержание высокого уровня эффективности работы отделов по управлению персоналом – задача, к решению которой сегодня можно подойти путем использования средств автоматизации функций кадровых служб и улучшения качества общего информационного обеспечения: с широким спектром задач по управлению персоналом HR-менеджерам призваны помочь специализированные HR-системы, включающие в себя актуальные средства рекрутинговой аналитики, выявляющие паттерны поиска, подбора и найма сотрудников, и передовые компьютерные технологии [1].

Сегодня автоматизация не заканчивается представлением полученной информации в электронном формате и последовательной ее обработкой с помощью отлаженных алгоритмов. Современная автоматизация процессов – попытка облегчить труд человека на совершенно другом уровне. Использование широкого математического аппарата, специализированных программных средств и алгоритмов машинного обучения позволяет человеку не только обработать огромные объемы формализованной информации, но и взглянуть на исследуемую предметную область под другим углом: представить любой объект как набор взаимосвязанных признаков, выявить не очевидные поначалу связи между собранными данными, ранжировать признаки по степени их значимости в ходе исследования и даже использовать для предсказания поведения и характера взаимодействия объекта с его окружением.

В данной работе предложена новая методика анализа цифрового следа, основанная на алгоритмах определения проявления отклоняющегося поведения в письменной речи и классификации текстовых публикаций, которая сочетает в себе сентимент-анализ и оценку лексикона по методу «мешка слов». Эта интеграция позволяет значительно повысить точность оценки поведенческих характеристик кандидатов, что является нововведением в области автоматизации подбора персонала, особенно в контексте использования русскоязычных социальных сетей. В рамках исследования были изучены и систематизированы различные словари, такие как тональные словари русского языка, словари разговорной экспрессивной речи и бранной лексики. На основе этих словарей были введены и классифицированы понятия девиантного, делинквентного и позитивного поведения, что позволило создать три системных словаря для анализа текстовых публикаций.

2. СОЦИАЛЬНАЯ СЕТЬ КАК ИСТОЧНИК ИНФОРМАЦИИ

Несмотря на широкие аналитические возможности внедряемого в работу HR-менеджеров программного обеспечения, они сталкиваются с проблемой нехватки данных о личности и характерном для соискателя поведении. Информация о навыках, компетенциях и опыте работы, полученная из резюме и во время официального собеседования, создает образ кандидата только в контексте профессиональной сферы и не отвечает на вопросы о том, какое поведение человек обычно демонстрирует в социуме: как взаимодействует с окружающими людьми, каких взглядов придерживается, с какими социальными группами себя ассоциирует. Ответы на эти вопросы могут сыграть не последнюю роль в принятии решения о найме кандидата на должность. В попытке найти специалиста, имеющего неподдельный интерес к работе, способного влиться в коллектив и поддержать репутацию компании, менеджеры вынуждены обращаться к дополнительным источникам информации.

Согласно результатам опроса, заказанного компанией Express Employment Professionals, и опроса, проведенного группой компаний HeadHunter, в качестве инструмента для получения недостающей информации о характере поведенческих реакций, HR-специалисты широко используют социальные сети, анализируя опубликованный соискателем контент на предмет проявления им неприемлемого или подозрительного поведения [2, 3]. Распространение такой практики стало предиктором создания целого класса программных решений, направленных на анализ публичной онлайн-активности потенциальных кандидатов на должности.

Среди наиболее популярных решений можно выделить такие продвинутые зарубежные программные системы, как Fama, Ferettly, Neotas и Good egg, предназначенные для скрининга социальных сетей соискателей и сотрудников компаний: анализа текстовых и графических публикаций [4–7]. Однако все эти системы объединяет одна общая особенность – их ориентированность на работу с зарубежными платформами, затрудняющая использование представленных средств отечественными компаниями для оценки публикаций в российских социальных сетях.

В данной работе предложено решение задачи автоматизации анализа цифрового следа соискателя в социальной сети «ВКонтакте».

3. ЭТАПЫ РАЗРАБОТКИ СИСТЕМЫ

Работа по автоматизации проведения анализа цифрового следа соискателя была разделена на два этапа:

1. этап исследования, включивший в себя определение возможных методов сбора и обработки открытых публичных данных, формализацию критериев оценки и разработку алгоритма анализа онлайн-активности пользователей социальной сети;
2. этап создания программной системы, реализующей полученный алгоритм анализа при проверке публикаций выбранного кандидата.

Решение большинства задач, связанных с обработкой информации, полученной из социальных сетей, для сбора необходимых данных требует использования предоставляемых их платформами программных интерфейсов. В качестве основного инструмента в работе с публичными данными на каждом из этапов решения поставленной задачи был использован набор средств программного интерфейса социальной сети «ВКонтакте» – «API ВКонтакте», предоставляемого сторонним разработчикам и позволяющего получать информацию напрямую из базы данных vk.com [8, 9].

Выбранный программный интерфейс «API ВКонтакте» включает в себя методы для работы с такими внутрисистемными объектами как «Пользователь», «Сообщество», «Приложение», «Беседа», медиаконтентом (фотографиями, аудиозаписями, видеозаписями), их вспомогательными объектами и наборами значений. Взаимодействие сторонних приложений и программного интерфейса осуществляется посредством протокола гипертекстовой разметки HTTP (HyperText Transfer Protocol): в ответ на GET или POST-запрос с явным указанием личного ключа доступа, имени метода и необходимых свойств объекта, API возвращает JSON-объект с запрошенными данными, согласно уровню доступа [10].

Анализ цифрового следа соискателя в рамках разработанного приложения сводится к получению обособленных результатов анализа количественных характеристик опубликованной информации (количества фотографий, заполненных полей в описании профиля пользователя, общего числа публикаций), выраженных в оценке «открытости» кандидата на должность, и результатов анализа его текстовых публикаций, выраженных в определении принадлежности публикаций к одному из выделенных по типу отклоняющегося поведения классов и демонстрации совокупной оценки проявления кандидатом в письменной речи делинквентного, девиантного и позитивного поведения.

Определение способа анализа онлайн-активности соискателей стало главной целью этапа исследования, в рамках которого работа велась в двух направлениях: обработка совокупности публичных данных профилей случайных пользователей «ВКонтакте» с целью оценки уровней проявления ими того или иного поведения и решение задачи классификации текстовых публикаций с использованием методов машинного обучения.

4. КОЛИЧЕСТВЕННАЯ ОЦЕНКА СОЦИАЛЬНОЙ АКТИВНОСТИ

Для сбора и обработки публичных данных тысячи случайных пользователей социальной сети ВКонтакте: списков и общего количества их текстовых публикаций, общего числа фотографий и заполненных полей в описании профиля пользователя в разделах «Карьера», «Образование» и «Жизненная позиция», - с целью получения количественных оценок проявления отклоняющегося поведения и социальной открытости соискателей было разработано

взаимодействующее с «API ВКонтакте» вспомогательное программное обеспечение на языке Python.

Далее была осуществлена предобработка данных в соответствии с методикой [11]: очистка текстов от знаков пунктуации, специальных символов, HTML-тэгов, приведение их к нижнему регистру и разбиение на отдельные слова – токены, удаление стоп-слов, стемматизация и лемматизация каждого полученного из текстов слова.

Особое внимание уделялось обработке персональных данных. Все данные пользователей перед обработкой прошли процедуру анонимизации, включающую удаление всех идентифицирующих признаков.

По завершении предобработки текстов для каждого из тысячи случайных пользователей был составлен личный «мешок слов», представляющий собой текст в виде мультимножества его слов без учета грамматики и их порядка, но содержащего информацию об их количестве [12]. С целью выявления в «мешках слов» терминов, указывающих на проявление девиантного, делинквентного или позитивного поведения на основе тональных словарей русского языка [13–15], трудов по лингвистической экспертизе текстов [16–18], а также словарей разговорной экспрессивной речи [19] и бранной лексики [20] были составлены три системных словаря. С использованием вспомогательной программы было посчитано количество совпадающих слов в «мешках слов» пользователей и системных словарях, а также были составлены для каждой категории слов соответствующие графики кумулятивной вероятности. Приведены графики для категорий «Делинквентная лексика» (рис. 1) и «Девиантная лексика» (рис. 2). Каждое значение на оси ординат построенных графиков характеризует вероятность того, что в «мешке слов» пользователя, выбранного случайным образом, количество обнаруженных совпадающих слов не превышает соответствующую величину, отложенную на оси абсцисс. На основе анализа графиков была получена экспертная оценка, согласно которой количественные показатели каждого из критериев были соотнесены с одним из трех уровней проявления: низким, средним и высоким. Аналогичным образом были проанализированы общее число текстовых публикаций, количество опубликованных фотографий и заполненных полей в описании профиля. Полученные количественные показатели, соответствующие трем уровням проявления, приведены в таблице 1.

Таблица 1. Соответствие количественных показателей уровням проявления критериев.

Критерий	Уровень проявления		
	Низкий	Средний	Высокий
Количество совпадений со словарем делинквентной лексики	$x \leq 3$	$3 < x \leq 695$	$x > 695$
Количество совпадений со словарем девиантной лексики	$x \leq 1$	$1 < x \leq 255$	$x > 255$
Количество совпадений со словарем позитивной лексики	$x \leq 3$	$3 < x \leq 333$	$x > 333$
Количество заполненных полей в описании профиля	$x \leq 1$	$1 < x \leq 5$	$x = 6$
Количество опубликованных фотографий	$x \leq 29$	$29 < x \leq 2490$	$x > 2490$
Количество текстовых публикаций	$x \leq 112$	$112 < x \leq 5036$	$x > 5036$

В отличие от оценок уровней проявления девиантного, делинквентного и позитивного поведения, оценка уровня «открытости» является комплексной и складывается из оценок количества опубликованных фотографий, заполненных полей в описании профиля кандидата, общего числа текстовых публикаций. В рамках работы уровни по первым двум критериям оценивались в 1, 2 и 3 балла соответственно, уровни по критерию количества опубликованных фотографий, как более приоритетному, в 2, 4 и 6 баллов. Уровень «открытости» определялся

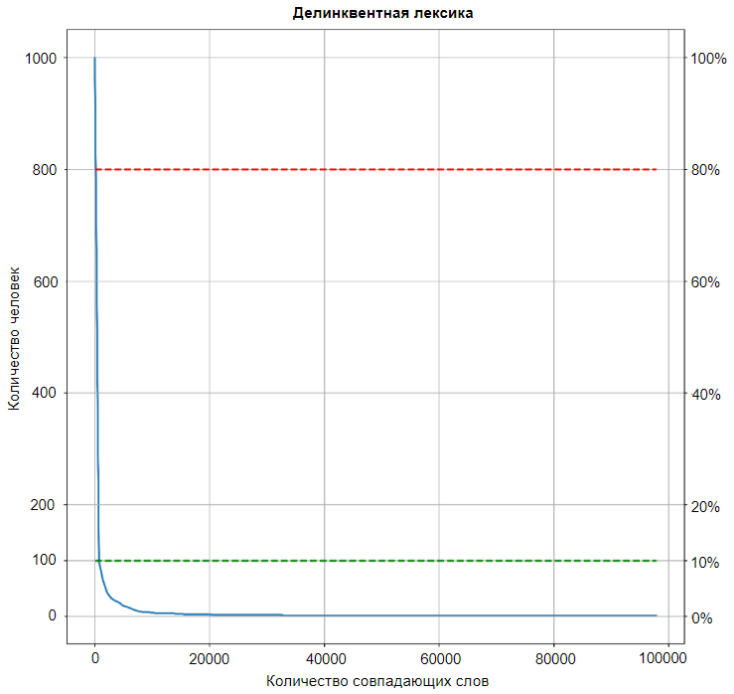

Рис. 1. График кумулятивной вероятности для категории «Делинквентная лексика».

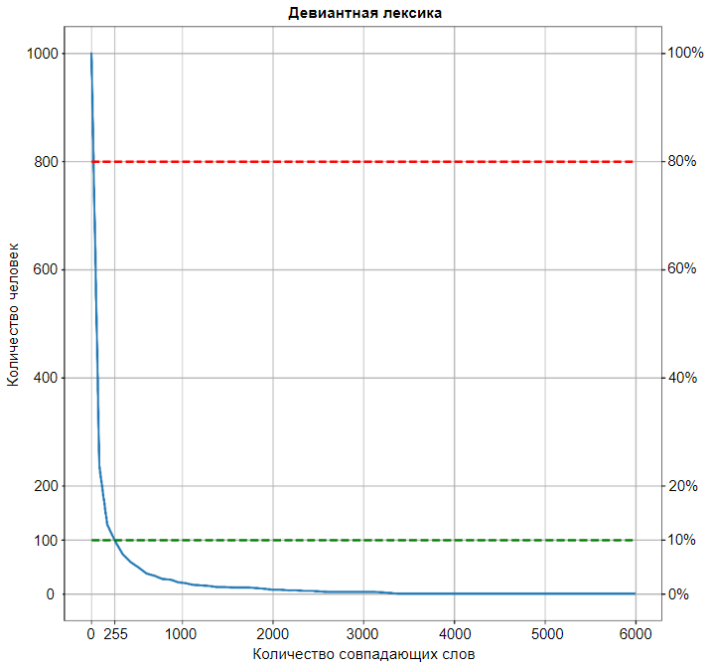

Рис. 2. График кумулятивной вероятности для категории «Девиантная лексика».

на основе суммы баллов по всем трем критериям: сумма в 4–5 баллов соответствовала низкому уровню, 6–9 – среднему, 10–12 – высокому.

5. КЛАССИФИКАЦИЯ ТЕКСТОВЫХ ПУБЛИКАЦИЙ

Следующей за формированием критериев оценки уровней проявления поведения и социальной открытости задачей было сравнение возможностей классификации текстовых публикаций нескольких моделей машинного обучения. Классификация текстов – одна из фундаментальных задач в обработке естественного языка (NLP – natural language processing), вызывающая интерес исследователей в самых разных научных сферах к созданию уникальных классификаторов, позволяющих анализировать содержимое текстовых данных и присваивать им наборы тегов или категорий в зависимости от содержания. Разработка основанных на методах машинного обучения систем – один из возможных подходов к автоматической NLP-классификации текстов, ставший широким полем для проведения экспериментов с алгоритмами обучения, характерных как для обработки естественного языка, так и исключительно количественных данных в связи с возможностями представления целых документов или слов по-отдельности в виде числовых векторов. Целью классификации текста является изучение существующих связей между заранее известной категориальной переменной и экземплярами, состоящими из независимых переменных [21]. То есть, работа ведется с некоторым заранее подготовленным корпусом текстов, компонентам которого присвоены специальные лингвистические или внешние теги.

Необходимый для обучения классификаторов корпус текстов был сформирован с помощью вспомогательной программы: через запросы к «API ВКонтакте» были получены и сохранены тексты 4863 публикаций различных открытых сообществ «ВКонтакте». В текстах сохраненных публикаций встречались бранная и ненормативная лексика, упоминания противоречащих моральным и правовым нормам социального поведения поступков, а также описания отличных по роду деятельности массовых мероприятий (спортивных марафонов, образовательных конференций и тренингов, волонтерских акций и фестивалей), позволившие приглашенным экспертам отнести каждый текст к одному из выделенных классов: классам поведения («deviant», «delinkvent» и «positive») или нейтральному («neutral») классу публикаций, которые не удалось отнести к трем предыдущим.

В связи с преобладанием в исходной выборке публикаций, отнесенных к классу «neutral» для работы с моделями машинного обучения было принято решение ограничить их количество в сформированном наборе данных. Таким образом, в дальнейшем использовался корпус, в котором оказалось 770 предварительно обработанных программой текстов с тегом «delinkvent», 1012 – с тегом «deviant», 979 – с тегом «positive» и 1000 – с тегом «neutral». В процессе работы 80% исходной выборки было использовано в качестве обучающей выборки, 20% – тестовой. В качестве критерия оценки качества модели была использована метрика accuracy.

В качестве первой модели для решения задачи мультиклассовой классификации использовалась логистическая регрессия, в качестве результата продуцирующая вероятность принадлежности объекта к некоторому классу N на основе вычисленных на обучающей выборке коэффициентов и входных признаков, представленных числовыми значениями [22]. Создание и обучение выбранной модели было проведено с использованием Python-библиотеки Scikit-learn, предназначенной для машинного обучения. С целью извлечения признаков из текстовых данных был использован метод TF-IDF (Term Frequency-Inverse Document Frequency), отражающий важность каждого слова в документе относительно количества его употреблений в конкретном документе и во всей коллекции текстов и позволивший в ходе исследования проанализировать наиболее характерные наборы слов и биграмм для текстов каждого из классов [21, 23]. Для обучения модели логистической регрессии каждый текст в выборке был

представлен в виде матрицы TF-IDF показателей, каждый класс поведения – поставленным ему в соответствие целым числом. Значение метрики accuracy полученной модели на тестовой выборке составило 0,6813.

Наряду с моделью логистической регрессии была спроектирована модель с использованием средств открытой библиотеки глубокого машинного обучения Keras. При подготовке тексты исходного корпуса были преобразованы в матрицу числовых последовательностей, в которых каждому слову соответствовал его номер в составленном по всему корпусу словаре. Полученная в ходе работы нейронная сеть состоит из комбинации Embedding-слоя, сопоставляющего входную последовательность целых чисел последовательности плотных векторов, слоя регуляризации SpatialDropout1D, используемого для предотвращения переобучения в нейронных сетях, двух скрытых LSTM-слоев (Long short-term memory) усовершенствованной рекуррентной нейронной сети, способной обучаться долгосрочным зависимостям, и одного промежуточного плотного слоя, в качестве функции активации использующего гиперболический тангенс для генерации выходного тензора. В качестве функции потерь была выбрана категориальная кросс-энтропия, а в качестве алгоритма оптимизации – алгоритм Adam, являющийся расширением стохастического градиентного спуска и обеспечивающий индивидуальное обновление скорости обучения каждого веса нейронной сети. После 12 эпох обучения точность модели на тестовой выборке составила 0,7025.

Последняя рассмотренная модель была получена в результате настройки предварительно обученной многоязыковой модели, принадлежащей семейству языковых моделей BERT. BERT-модели отличаются использованием архитектуры Transformer, предполагающей последовательное кодирование входных предложений в векторы и декодирования в последовательность выходных лексем с использованием специального механизма внимания, позволяющего выявлять зависимости между словами в предложениях за счет фокусирования на наиболее релевантных терминах при декодировании выходных лексем. Моноязычные BERT-модели способны показывать более высокие результаты при решении задач классификации текстов и анализа их тональности [24]. В качестве основы была выбрана доступная на платформе Hugging Face Hub модель «rubert-base-cased», обученная на новостных статьях и материалах ресурса «Википедия» на русском языке [25]. В качестве входных аргументов при обучении выступили список полных текстов обучающей выборки и список соответствующих им классов. Наилучший результат полученная модель показала на тестовой выборке по прошествии 15 эпох обучения: значение метрики accuracy составило 0,7242.

В таблице 2 приведены использованные гиперпараметры моделей, в таблице 3 – полученные в ходе работы с моделями машинного обучения значения метрики качества. Для всех трех моделей с целью оценки соответствия их результатов классификации семантике публикаций была проведена дополнительная проверка на наборе предобработанных текстов публикаций, не вошедших ни в обучающую, ни в тестовую выборки: были рассмотрены восемь публикаций различных сообществ «ВКонтакте», по две для каждого из выделенных классов. Во время проверки сравнивались теги, проставленные экспертом, и результаты, сгенерированные обученными моделями.

В таблице 4 приведен пример классификации модели, основанной на архитектуре Transformer. В ходе проверки только эта модель корректно определила теги всех публикаций, и именно она по результатам исследовательского этапа была выбрана в качестве дополнительного инструмента для определения принадлежности публикации соискателя к проявлению девиантного, делинквентного или позитивного поведения в письменной речи.

Часть разработанной во время исследовательского этапа программы, позволяющей использовать методы «API ВКонтакте» для получения открытых публичных данных пользователей социальной сети и включающей в себя методы предобработки и классификации текстовых

Таблица 2. Гиперпараметры моделей.

Модель	Гиперпараметры	
Основанная на логистической регрессии	Тип регуляризации	l2
	Алгоритм оптимизации	lbfgs
	Максимальное количество итераций	100
	Обратное значение силы регуляризации	$\frac{1}{13}$
Построенная с использованием LSTM-слоев	Оптимизатор	adam
	Функция потерь	categorical_crossentropy
	Размер батча	300
	Количество эпох обучения	12
	Embedding слой	
	Размер словаря	25500
	Размер плотного embedding-a	128
	Длина входных последовательностей	1000
	SpatialDropout1D	
	Rate	0.5
	LSTM слой	
	Размерность выходного пространства	40
	Функция активации	tanh
	Функция активации, используемая на повторяющемся шаге	sigmoid
	Плотно связанный слой NN	
	Размерность выходного пространства	40
	Функция активации	tanh
	LSTM слой (аналогичен 3)	
	Плотно связанный слой NN	
	Размерность выходного пространства	4
	Функция активации	softmax
Основанная на архитектуре Transformer	Размер батча	8
	Количество эпох обучения	15
	Оптимизатор	AdamW
	Максимальная длина последовательности	128
	Базовая предобученная модель Hugging Face Transformers	rubert-base-cased

Таблица 3. Сравнение метрики качества моделей.

Модель	Значение метрики accuracy
Основанная на логистической регрессии	0,6813
Построенная с использованием LSTM-слоев	0,7025
Основанная на архитектуре Transformer	0,7242

публикаций, легла в основу центрального аналитического модуля на языке Python созданной на втором этапе работы системы.

Таблица 4. Сравнение метрики качества моделей.

Текст публикации	Оценка эксперта	Результат модели
Премия Российского общества «Знание» учреждена для признания достижений деятелей просвещения, педагогов, лекторов, авторов, блогеров, популяризаторов науки, а также чтобы отметить просветительские проекты и компании из разных областей. В 2023 году конкурс на соискание премии проходит в третий раз. Пройдет прямое включение участников Марафона с Президентом Российской Федерации в рамках церемонии награждения Международной премии.	positive	positive
Местные забулдыги спёрли 1300 метров кабеля! Сумма ущерба 345 тысяч рублей. Мало того, своими действиями они отрубили интернет и ТВ нескольким сотням жителей в Промышленном районе города. Ворюги делали вид, будто занимаются каким то монтажным работами, переодевшись в спецовку. Собранный кабель грузили в машину, далее рубили на куски, обжигали на костре и отвозили в пункт приема металлолома.	delinkvent	delinkvent
Предварительно, все произошло из-за обычного замечания. Дело в том, что мужчина в нетрезвом состоянии разлегся на диване в подъезде и покуривал сигареты. Мать с ребенком сделала замечание соседу. В итоге мужчина напал на соседку. После того, как женщина все же смогла отбиться от неадеквата, мужчина зашел в соседний подъезд, где сперва поджег лифт, а позже вырвал камеру видеонаблюдения. Местные жители обезвредили самарца и вызвали полицию.	delinkvent	delinkvent
Вчера вечером состоялась долгожданная вечиринка Halloween Немного алкоголя, кальян и танцы до утра - идеально! Наконец-то отдохнули!	deviant	deviant
Ереван красивый, исторический, модный и неповторимый город. Я был приятно удивлен от большого количества стильных мест с вкусной едой В этом городе идеальный баланс современных и честных наивкуснейших Армянских ресторанов с превосходной едой. Узнал много нового и интересного.	neutral	neutral

6. ДЕТАЛИ РЕАЛИЗАЦИИ СИСТЕМЫ

По завершении исследовательского этапа была спроектирована и разработана автоматизированная система, позволяющая авторизованным пользователям – HR-менеджерам фиксировать информацию о рассматриваемых на должности кандидатах в их личных анкетах, проводить анализ их цифрового следа в социальной сети «ВКонтакте» и в результате получать сформированные системой отчеты о характере их поведенческих реакций.

Разработанная система призвана сократить время, необходимое HR-менеджеру для получения дополнительной информации о личности соискателя на должность путем количественного анализа опубликованной им информации и выделения в общей массе его текстовых публикаций тех, которые с наибольшей вероятностью заслуживают внимания специалиста: могут указать на случаи отклонения поведения кандидата от общепринятых социальных норм или, напротив, подчеркнуть его лучшие профессиональные и личностные качества, указать на развитые навыки коммуникации и активную жизненную позицию.

Система представляет собой веб-приложение, реализованное на основе клиент-серверной архитектуры (рис. 3), в рамках которой компьютеры-клиенты (удалённые системы) имеют возможность посылать запросы и получать ответы от сервера [26]. Серверная часть системы была разработана на языке JavaScript с помощью платформы Node.js и фреймворка Express, используемых для создания веб-приложений, средств объектно-реляционной системы управления базами данных PostgreSQL, а также отдельно реализованного аналитического модуля на языке Python. Для осуществления обработки текстов в аналитическом модуле были выбраны библиотеки, входящие в состав платформы для процессинга естественного языка NLTK, морфологический анализатор Pymorphy2, а также NLP-библиотека Simple Transformers, предназначенная для работы с моделями машинного обучения, основанными на архитектуре Transformer. Клиентская часть системы была реализована с помощью библиотеки React.js для разработки пользовательских интерфейсов и дополнительных библиотек для отображения графиков и динамических элементов.

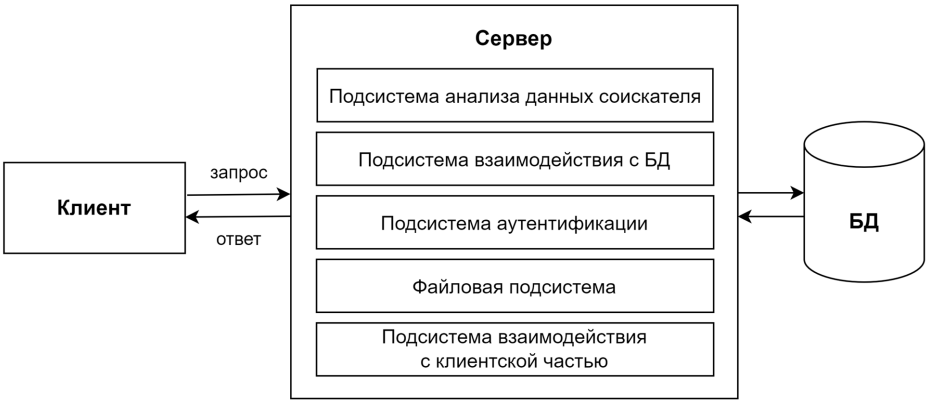

Рис. 3. Структурная схема системы.

В серверной части системы сосредоточена логика по сбору и обработке публичных данных соискателей. При поступлении соответствующего запроса на анализ данных конкретного кандидата, обработчики серверной части приложения делегируют эту задачу реализованному на языке Python аналитическому модулю, в задачи которого входит проверка настроек конфиденциальности на странице соискателя во «ВКонтакте» и возможности получения его открытых публичных данных, предобработка и двухэтапный анализ текстовых публикаций (рис. 4), вычисление частоты публикаций на основе данных за последние полгода и количественных показателей проявления отклоняющегося поведения и социальной открытости, а также сохранение необходимых данных анкеты соискателя в базе данных (БД).

Наряду с модулем обработки данных соискателей в системе присутствует модуль, отвечающий с одной стороны за формирование и сохранение в БД учетных записей HR-менеджеров – непосредственных и главных пользователей разработанной системы, с другой – связей между учетными записями менеджеров и добавленными ими анкетами соискателей. Хранение связей

позволяет не только обеспечить возможность создания анкеты одного и того же кандидата разными HR-специалистами, но и поддержать консистентность данных при редактировании ими списков анкет.

7. ФУНКЦИОНАЛЬНОСТЬ И ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ СИСТЕМЫ

Воспользоваться системой пользователь может после перехода по соответствующей ссылке в веб-браузере и прохождения обязательных этапов регистрации и авторизации. Доступ к персональным данным соискателя ограничен и предоставляется только авторизованным HR-специалистам, получившим согласие кандидата на обработку персональных данных.

Как только пользователь вводит данные своей учетной записи на главной странице и успешно авторизуется в системе, ему становятся доступны создание новых и редактирование ранее сохраненных в системе анкет кандидатов на должности, просмотр сформированных системой для каждого кандидата отчетов, настройка личного пользовательского словаря терминов и данных своей учетной записи на соответствующих страницах.

Создание новой анкеты предполагает заполнение менеджером специальной формы данными из полученного от кандидата резюме: информацией о личности кандидата, его опыте работы, должности, на которую он претендует, а также его идентификатором в социальной сети «ВКонтакте». Когда заполнение формы завершено, система запускает процесс анализа публичных данных соискателя, в течение которого отображает пользователю прогресс его выполнения. Закончив анализ и сохранив всю необходимую информацию в БД, система отображает пользователю сформированную анкету соискателя: указанные менеджером персональные данные, график активности соискателя – вычисленной частоты его публикаций в каждый из дней недели, отчеты по текстовым публикациям и определенные системой уровни проявления девиантного, делинквентного и позитивного поведения, а также уровень открытости.

По умолчанию при добавлении пользователем новой анкеты кандидата система формирует только «системный» отчет из публикаций, разбитых на две группы: «negative» – группу публикаций, в текстах которых были выделены слова, содержащиеся в системных словарях девиантной и делинквентной лексики, и «positive» – в текстах которых были найдены термины, совпадающие со словами из словаря позитивной лексики. Активация функции формирования системой «пользовательского» отчета требует от менеджера предварительного заполнения его личного словаря терминов. Сформированные «пользовательские» отчеты вместе с общим количеством выявленных совпадений также доступны для просмотра на страницах анкет кандидатов. Доступная пользователю функциональность не ограничивается лишь предварительным просмотром созданных отчетов. Ему обеспечена возможность просмотра подробной информации о каждой выбранной и сохраненной в отчете публикации: полный текст, дата и время размещения на странице пользователя, признак авторства, список ключевых слов, ставших причиной для добавления публикации в открытый отчет, а также метка класса, к которому была отнесена публикация языковой моделью. Более того, менеджер может отредактировать список анкет соискателей, удалив выбранную анкету, и любой из отчетов, удалив публикацию, которая, по его мнению, обладает не подходящей для конкретного отчета семантикой или не представляет интереса для менеджера в процессе рассмотрения открытой анкеты кандидата. Система отслеживает все изменения, внесенные пользователем в «системный» и «пользовательский» отчеты, и тут же обновляет результаты анализа, отображаемые в анкете, соответствующим образом. На рисунке 5 приведена экранная форма анкеты кандидата.

8. ЗАКЛЮЧЕНИЕ

Использование в работе HR-специалистов программных систем, предназначенных для скрининга социальных сетей кандидатов на должность, является актуальным и современным реше-

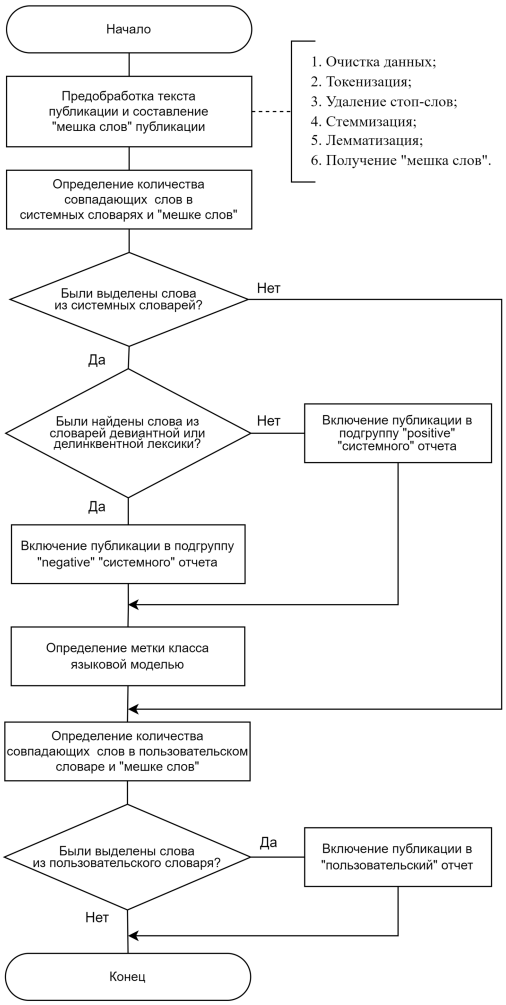

Рис. 4. Алгоритм обработки текстовой публикации кандидата.

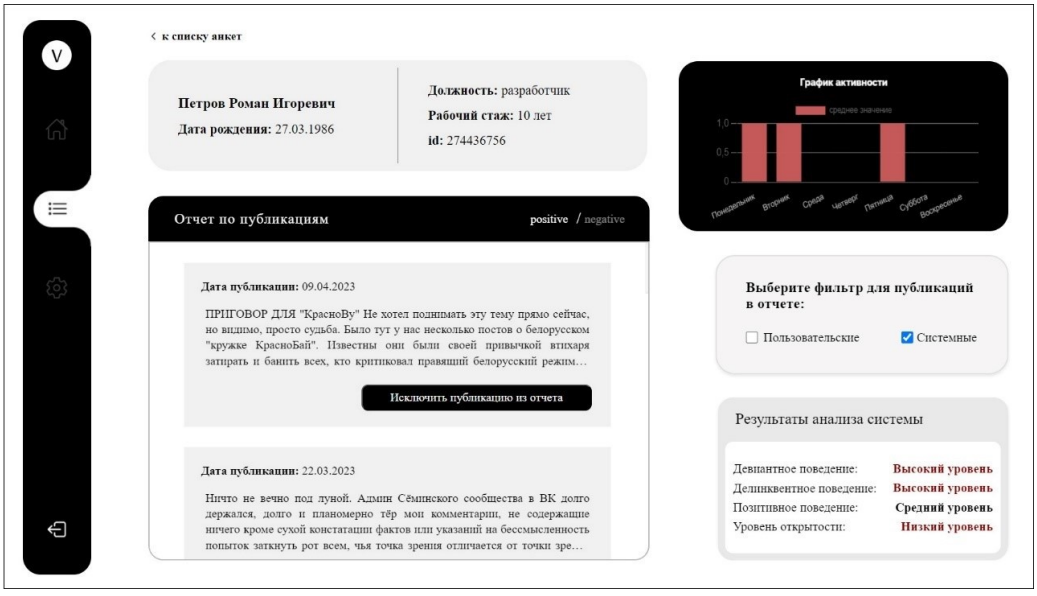

Рис. 5. Экранная форма анкеты соискателя.

нием при подборе персонала. В данной работе предложена технология анализа цифрового следа соискателей и рассмотрены этапы ее реализации. Описан ход проведенного исследования, в результате которого были определены количественные показатели отклоняющего поведения и социальной открытости пользователей социальной сети, с помощью методов машинного обучения построена и обучена модель классификации текстовых публикаций. Разработанная технология реализована в виде веб-приложения, позволяющего HR-менеджерам упростить процесс сбора публичных данных и анализа социальной активности кандидатов на должность. В будущих работах авторы планируют расширить функционал разработанного решения.

СПИСОК ЛИТЕРАТУРЫ

1. Рекрутинговая аналитика: три уровня оптимизации рекрутинга. URL: https://www.aihr.com/blog/implement-hris-human-resources-information-system (06.01.2024).
2. 71% of Hiring Decision-Makers Agree Social Media is Effective for Screening Applicants. URL: https://www.expresspros.com/Newsroom/America-Employed/71-of-Hiring-Decision-Makers-Agree-Social-Media-is-Effective-for-Screening-Applicants.aspx (06.01.2024).
3. Не пригласили на работу? Возможно, проблема в вашем профиле в соцсетях. URL: https://samara.hh.ru/article/24311 (06.01.2024).
4. Software for Screening Social Media Fama.io. URL: https://fama.io/product (06.01.2024).
5. The Ferretly Platform. URL: https://www.ferretly.com/product (06.01.2024).
6. Neotas. URL: https://www.neotas.com (06.01.2024).
7. Good Egg Pricing, Features, Reviews and Alternatives. URL: https://www.getapp.com/hr-employee-management-software/a/good-egg (06.01.2024).
8. Использование API. Быстрый старт. VK для разработчиков. URL: https://dev.vk.com/api/getting-started (06.01.2024).
9. Михайлов Д.В., Козлов А.П., Емельянов Г.М. Выделение знаний и языковых форм их выражения на множестве тематических текстов: подход на основе меры TF-IDF. Компьютерная оптика, 2015, т. 39, № 3, стр. 429–438.
10. Справочник API. Объекты. Список объектов. VK для разработчиков. URL: https://dev.vk.com/reference/objects (06.01.2024).
11. Плавное введение в Natural Language Processing (NLP). URL: https://datastart.ru/blog/read/plavnoe-vvedenie-v-natural-language-processing-nlp (06.01.2024).
12. Harris Z. Distributional Structure. WORD, 1954, v. 10(2-3), pp. 146–162.
13. Тональный словарь русского языка КартаСловСент. URL: https://github.com/dkulagin/kartaslov/tree/master/dataset/kartaslovsent (06.01.2024).
14. NRC Word-Emotion Association Lexicon. URL: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (06.01.2024).
15. Словарь оценочных слов и выражений русского языка РуСентиЛекс. URL: https://www.labinform.ru/pub/rusentilex/index.htm (дата обращения: 06.01.2024).
16. Стернин И.А., Антонова Л.Г., Карпов Д.Л., Шаманова М.В. Выявление признаков унижения чести, достоинства, умаления деловой репутации и оскорбления в лингвистической экспертизе текста. Ярославль: Канцлер, 2013.
17. Еленская К.В. Лингвистический паспорт лексемы "насилие". Преподаватель XXI век, 2020, № 4–2, стр. 386–392.

18. Сердюкова В.М. Практикум по специальному переводу (внешнеполитические отношения). Сетевое электронное учебное издание. УО «Минский государственный лингвистический университет», 2016. URL: http://elearning.mslu.by/assignments/88/about/ (дата обращения: 06.01.2024).

19. Химик В.В. Большой словарь русской разговорной экспрессивной речи. СПб.: Норинт, 2004.

20. Мокиенко В.М., Никитина Т.Г. Словарь русской брани (матизмы, обсценизмы, эвфемизмы). СПб.: Норинт, 2003.

21. Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. СПб.: Питер, 2019. (Bengfort B., Bilbro R., Ojeda T. Applied Text Analysis with Python. Enabling Language-Aware Data Products with Machine Learning. O'Reilly Media, Incorporated, 2018.)

22. Воронина В.В., Михеев А.В., Ярушкина Н.Г., Святов К.В. Теория и практика машинного обучения: учебное пособие. Ульяновск: УлГТУ, 2017.

23. Рыцарев И.А., Кирш Д.В., Куприянов А.В. Кластеризация медиа-контента из социальных сетей с использованием технологии BigData. Компьютерная оптика, 2018, т. 42, № 5, стр. 921–927.

24. Куратов Ю., Архипов М. Адаптация глубоких двунаправленных многоязычных трансформеров для русского языка. Вычислительная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог». Москва, 2019, вып. 18, стр. 333–339.

25. Hugging Face rubert-base-cased. URL: https://huggingface.co/DeepPavlov/rubert-base-cased (06.01.2024).

26. Таненбаум Э., Уэзеролл Д. Компьютерные сети. СПб.: Питер, 2012, 5-е изд. (Tanenbaum A., Wetherall. D. Computer networks. Upper Sadle River, New Jersey: Prentice Hall, Incorporated, 2011, 5th ed.).

Development of automated system for analyzing the digital footprint of a social media user

K.R. Mudrova, A.N. Zhdanova, A.V. Kupriyanov

The paper considers the role of software systems designed for social media background screening of job seekers and the relevance of their implementation in the process of automating the work of recruitment departments. One of the possible approaches of solving the automation problem for analyzing a social media digital footprint by creating a special web application is considered. The algorithm and results of the study of random VKontakte social media users' data are presented, which allowed to determine the indicators of the social openness level and the level of deviant behavior manifestation used in the implementation of the corresponding application.

KEYWORDS: publicly available data, HR systems, recruitment, posts analysis, text processing, web application, machine learning.