На правах рукописи

ЗОТКИНА АЛЕНА АЛЕКСАНДРОВНА

МЕТОДЫ И АЛГОРИТМЫ ФОРМИРОВАНИЯ ПСИХОЛОГИЧЕСКОГО ПОРТРЕТА ПОЛЬЗОВАТЕЛЯ СОЦИАЛЬНОЙ СЕТИ ДЛЯ ЭФФЕКТИВНОГО ПОДБОРА КАДРОВ

Специальность 2.3.8. – Информатика и информационные процессы

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

ПЕНЗА – 2024

Работа выполнена на кафедре «Программирование» в Федеральном государственном бюджетном образовательном учреждении высшего образования «Пензенский государственный технологический университет».

Научный руководитель – **Мартышкин Алексей Иванович** кандидат технических наук, доцент, заведующий кафедрой «Программирование» ФГБОУ ВО «Пензенский государственный технологический университет», г. Пенза

Официальные оппоненты: **Иванов Александр Иванович,** доктор технических наук, профессор, АО «Пензенский научно-исследовательский электротехнический институт», научный консультант, г. Пенза
Ямашкин Станислав Анатольевич, кандидат технических наук, доцент, ФГБОУ ВО «Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва», г. Саранск

Ведущая организация – Федеральное государственное бюджетное образовательное учреждение высшего образования «Юго-Западный государственный университет», г. Курск

Защита диссертации состоится 24 декабря 2024 года в 12:00 часов, на заседании объединенного диссертационного совета 99.2.113.02 на базе ФГБОУ ВО «Рязанский государственный радиотехнический университет имени В.Ф. Уткина», ФГБОУ ВО «Пензенский государственный технологический университет» по адресу: 440039, г. Пенза, пр. Байдукова / ул. Гагарина, д. 1а/11, корпус 1, конференц-зал.

С диссертацией можно ознакомиться в научной библиотеке ФГБОУ ВО «Рязанский государственный радиотехнический университет имени В.Ф. Уткина, на сайте http://rsreu.ru/, в научной библиотеке ФГБОУ ВО «Пензенский государственный технологический университет» и на сайте http://www.penzgtu.ru.

Автореферат разослан «___» _____2024 г.

Ученый секретарь
диссертационного совета
доктор технических наук, доцент  А.Н. Колесенков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В современном обществе наблюдается экспоненциальный рост числа активных пользователей социальных сетей, что ведет к накоплению огромного объема данных. Этот объем данных представляет собой ценный ресурс, который может быть использован для проведения разнообразного анализа и извлечения значимой информации, например, позволяет оценить поведение пользователей и их личностные черты, а также выявить изменения в настроениях и критические психологические ситуации, включая депрессию или суицидальные наклонности путем анализа разнообразного цифрового контента, размещаемого человеком в виде публикаций, комментария событий и т.д. Однако, с увеличением объема данных возникает необходимость в разработке эффективных методов анализа, направленных на понимание поведения пользователей, их предпочтений и интересов. Это требует создания и совершенствования методологий и инструментов анализа данных, способных обрабатывать и интерпретировать большие объемы информации, выявлять скрытые закономерности и тенденции в пользовательском поведении. Исследование социальных сетей позволяет оценить поведение пользователей и их личностные черты, а также выявить изменения в настроениях и критические психологические ситуации, включая депрессию или суицидальные наклонности путем анализа разнообразного цифрового контента, размещаемого человеком в виде публикаций, комментария событий и т.д.

Исследования в области социального профилирования опираются на труды в области анализа данных, теории графов и сетей, авторами которых являются *J. Golbeck, C. Robles, K. Turner, S. Adali, W. Youyou, M. Kosinski, Stillwell D.* и другие. Современные ученые, среди которых *R. B. Tareaf, P. Berger, P. Hennig, C. Meinel, M. Vaidhya, B. Shrestha, B. Sainju, K. Khaniya, Liu F., Perez J., Nowson S.* и другие, активно изучают применение общедоступных данных Интернета, особенно в контексте социализированных данных. В России также существует научное сообщество, включая ученых, таких как Е.И. Большакова, Н.В. Лукашевич, П.И. Браславский, Е.В. Котельников, Ю.В. Рубцова, которые занимаются обработкой неструктурированных социализированных данных различного происхождения.

Анализ существующих исследований в этой области, показывает, что они в основном сосредоточены на изучении всей сети в целом, не уделяя должного внимания детальному изучению индивидуальных показателей и их особенностей, что ограничивает возможности персонализированного подхода к анализу поведения личности. Кроме того, большинство современных исследований ограничивается анализом данных только из одной социальной сети, что существенно сужает возможности предсказательных моделей. Подобный подход учитывает только фрагмент доступного для анализа цифрового следа, что в свою очередь снижает эффективность обработки и анализа данных. Традиционные методы составления психологического портрета пользователя учитывают только один из его аккаунтов в пределах одной социальной сети. Поскольку

пользователь может иметь несколько аккаунтов в различных сетях, такой подход не способен обеспечить достаточную точность и качество формируемого психологического портрета человека. Составление психологического портрета в настоящее время осуществляется в основном «ручным» способом, процедура занимает много времени ввиду обширной информации о человеке. Применение нейронных сетей для анализа психологических портретов позволит ускорить процесс, а также прогнозировать поведение отдельных лиц в будущем. Создание цифровых образов людей востребовано в различных областях деятельности, таких как психология, рекрутинг и др. При этом, психологический портрет используется исключительно как дополнительный инструмент для более глубокой оценки личных качеств и профессиональных склонностей человека.

Сложности построения социального портрета пользователя подчеркивают важность создания новых методов и алгоритмов обработки информации, размещаемой пользователями социальной сети, для решения задач выявления и идентификации факторов риска безопасности рабочей среды.

Таким образом, тема диссертационного исследования актуальна.

Объектом исследования является информация, размещаемая пользователями социальных сетей.

Предмет исследования – методы, алгоритмы и методики сбора данных для формирования психологического портрета пользователя.

Цель работы – совершенствование методов для формирования психологического портрета пользователя социальной сети, основанных на анализе информации, размещаемой ими, с учетом их индивидуально-психологических характеристик согласно типологии *Myers-Briggs Type Indicator* (*MBTI*) для прогнозирования профессионального поведения, разработки эффективных стратегий развития сотрудников и повышения уровня их удовлетворенности.

Для достижения поставленной цели в диссертации решаются следующие **задачи:**

1) проведение анализа существующих методов и моделей, применяемых для интеллектуальной обработки данных пользователей социальных сетей;

2) разработка метода для сравнения характеристик выражений и текстовых сообщений пользователей с аналогичными аккаунтами в социальных сетях;

3) разработка метода интеграции данных, размещаемых пользователем на различных платформах социальных сетей, который позволит восстанавливать данные активности, учитывая разнообразные аспекты его онлайн-поведения, с целью составления более полного и подробного психологического портрета и определения отклоняющегося поведения.

4) разработка методики анализа тональности текста, учитывающей контекст и особенности каждого текста, независимо от тематики и смыслового контекста.

5) разработка нейросетевой методики определения психологических характеристик пользователя социальной сети, с использованием типологии *MBTI*.

6) проведение экспериментального исследования для проверки предложенных методов и алгоритмов, а также разработка рекомендаций по их практическому применению.

Методы исследований. В диссертации применены методы интеллектуального анализа данных, методы теории вероятностей и математической статистики для обработки экспериментальных данных, методы обработки естественного языка, методы теории анализа социальных сетей (*Social Network Analysis, SNA*).

Научная новизна работы заключается в следующем.

1. Разработан метод оценки сходств признаков выражения, текстовых объектов, записей множества пользователей социальных сетей и реализующий ее алгоритм работы поиска аккаунтов пользователя, которые в отличии от существующих, учитывают разнообразные аспекты активности пользователей (публикации, участие в сообществах, комментарии, лайки к комментариям и публикациям). Это позволяет более точно идентифицировать одинаковые аккаунты пользователей.

2. Разработан метод интеграции информации, публикуемой пользователем на разных платформах социальных сетей, позволяющий восстанавливать данные о пользователях, проявивших активность хотя бы на одной из этих платформ, который отличается тем, что учитываются полные данные об активности пользователя на протяжении длительного периода времени, что позволяет составить более полный и подробный психологический портрет пользователя.

3. Разработана методика кросс-доменного аспектно-ориентированного анализа тональности текста и алгоритм на ее основе, которая в отличии от существующих, фокусируется на выделении аспектов и анализе тональности отношения к ним в тексте, что позволяет получить более детальное представление о содержании и оценке текста, в отличие от других рассматриваемых методик.

4. Нейросетевая методика и алгоритм, ее реализующий, для определения психологических характеристик пользователя социальной сети, с использованием типологии *MBTI*, которая в отличие от других подходов, фокусируется на изолированных личностных чертах, что позволяет предоставить комплексное представление личности.

Соответствие паспорту научной специальности. Область исследования, обозначенная в паспорте специальности 2.3.8. «Информатика и информационные процессы», охватывает следующие направления:

– разработка компьютерных методов и моделей описания, оценки и оптимизации информационных процессов и ресурсов, а также средств анализа и выявления закономерностей на основе обмена информацией

пользователями и возможностей используемого программно-аппаратного обеспечения (п. 1);

– разработка методов обработки, группировки и аннотирования информации, в том числе, извлеченной из сети интернет, для систем поддержки принятия решений, интеллектуального поиска, анализа (п. 7).

Теоретическая значимость. Развитие методов составления психологического портрета пользователя социальной сети, на основе размещаемой им публичной информации.

Практическая ценность. Использование методов, методик, алгоритмов и программных решений, разработанных в рамках диссертации, способствует сокращению времени формирования психологического портрета пользователя, что позволяет значительно повысить эффективность управления кадровой системой.

Реализация и внедрение результатов работы. Разработанные методы и алгоритмы внедрены в учебный процесс на кафедре «Программирование» ФГБОУ ВО ПензГТУ и используются при подготовке студентов по направлениям бакалавриата 09.03.01 «Информатика и вычислительная техника» и 09.03.04 «Программная инженерия» в рамках дисциплин «Методы машинного обучения и искусственного интеллекта», «Сбор и управление большими данными», «Технологии больших данных». Часть разработок и программно-технических решений, созданных в ходе диссертационного исследования, была внедрена в АО «НПП «Рубин», г. Пенза в рамках выполнения научно-исследовательской работы по теме «Метрика-Р», в ООО «ТД «ПЗЭМ» (г. Пенза) в рамках выполнения научно-исследовательского проекта по теме «Кадры для цифровой экономики», в Ассоциацию разработчиков программного обеспечения Пензенской области «Секон» при разработке решений для систем подбора кадров ряда организаций входящих в Ассоциацию (CodeInside, Tortuga), в АО «ППО ЭВТ им. В.А. Ревунова» при принятии решений по подбору сотрудников.

Достоверность результатов работы подтверждаются опытом внедрения результатов исследования в практическую и научно-исследовательскую деятельность ряда организаций, а также апробацией и обсуждением результатов диссертации на международных и всероссийских научных конференциях.

На защиту выносятся.

1. Метод оценки сходств признаков выражения, текстовых объектов, записей множества пользователей социальных сетей и алгоритм работы поиска аккаунтов пользователя на ее основе для выявления одинаковых аккаунтов.

2. Метод интеграции информации, размещаемой пользователем на разных платформах социальных сетей, который обеспечивает возможность восстановления данных для пользователей, активность которых зафиксирована хотя бы в одной из социальных сетей, с целью составления более полного и подробного психологического портрета.

3. Методика кросс-доменного аспектно-ориентированного анализа тональности текста и алгоритм работы модели на ее основе, который фокусируется на выделении аспектов и анализе тональности отношения к ним в тексте, что позволяет получить детальное представление о содержании и назначении текста.

4. Нейросетевая методика определения психологических характеристик пользователя социальной сети, с использованием типологии *MBTI*, позволяющая классифицировать пользователя по 16 факторам и алгоритм на ее основе. Результаты экспериментального анализа предложенных методов и алгоритмов, а также рекомендации по их практическому применению.

Апробация работы. Ключевые результаты, полученные в рамках диссертационного исследования, были опубликованы в научных журналах и апробированы на международных и всероссийских научных конференциях: Всероссийская научная конференция с международным участием «Цифровая индустрия: состояние и перспективы развития» (ЦИСП'2023) (Челябинск, 2023); Международная научно-практическая конференция «Индустрия 4.0» (*SmartIndustryCon*) (Сочи, 2023, 2024); XVII Международная научно-техническая конференция «Оптико-электронные приборы и устройства в системах распознавания образов и обработки изображений» (Курск, 2023); II Международный научно-практический форум по передовым достижениям в науке и технике (*SciTech* 2022) (Барнаул, 2022); Всероссийская научно-технической конференция «Современные методы и средства обработки пространственно-временных сигналов» (Пенза, 2021, 2023); Международная научно-практическая конференция «Современные информационные технологии» (Пенза, 2021, 2022, 2023, 2024); XXIII Международная научно-практическая конференция «Современные научные исследования: актуальные вопросы, достижения и инновации» (Пенза, 2022); Международный научно-исследовательский конкурс «Достижения в науке и образовании 2022» (Пенза, 2022); III Международная научно-практическая конференция «Наука и образование в современном обществе: актуальные вопросы и инновационные исследования» (Пенза, 2021).

По результатам диссертационного исследования опубликовано 38 научных работ, в том числе 6 статей в журналах, рекомендованных ВАК Минобрнауки России, 2 статьи, индексируемые в международной базе данных *Scopus*, получено 2 свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад автора. Все представленные в работе результаты исследования являются оригинальными и были получены автором самостоятельно. Данные, заимствованные у других авторов, сопровождаются ссылками на соответствующие опубликованные источники.

Объем и структура диссертации. Работа состоит из введения, четырёх глав, заключения, списка литературы, который включает 133 наименования, и 2 приложений. Общий объем диссертации составляет 140 страниц. Диссертация содержит 9 таблиц и 29 рисунков.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы исследования, формулируется ее цель и задачи, решение которых способствует достижению поставленной цели, отражены научная новизна и практическая значимость, сформулированы выносимые на защиту основные положения.

В первой главе проведен обзор современного состояния больших данных и методов интеллектуального анализа текстовых данных в социальных сетях.

Сегодня социальные сети стали неотъемлемой частью повседневной жизни, предоставляя уникальную возможность виртуального общения и обмена идеями по всему миру. Разработка технологий, использующих цифровые следы для анализа свойств и состояний личности, становится все более востребованной. Построение психологического портрета пользователя на основе размещаемой им информации становится важным направлением в этой области. Одним из важных применений психологического портрета пользователя в социальных сетях является процесс отбора персонала. Работодатели стремятся обеспечить безопасность своей компании, учитывая разнообразные риски, связанные с наймом персонала: финансовые, профессиональные и социальные. Для снижения этих рисков работодатели активно исследуют дополнительную информацию о потенциальных кандидатах, включая данные из социальных сетей.

Отмечено, что основные проблемы в таких исследованиях возникают из-за необходимости обработки огромного объема данных из социальных сетей и сложности выявления нелинейных связей между реальным поведением человека и его активностью в сети. Кроме того, большинство исследований базируется на данных только одной социальной платформы, что сужает возможности предсказательных моделей из-за ограниченного набора доступных цифровых следов. Это обуславливает низкую эффективность традиционных методов обработки и анализа данных.

Также исследованы коммерческие платформы анализа данных в социальных сетях (*Social Studio, IQBuzz, Brand Analytics*), где основное внимание уделяется статистическим методам для обнаружения упоминаний брендов. Однако эти платформы ориентированы лишь на поиск отзывов о небольшом круге брендов, не уделяя достаточного внимания конкретным пользователям сети, ограничиваясь только необходимыми метриками для анализа продвижения брендов.

В первой главе приведены примеры построения психологического портрета человека на основе открытой информации из социальных сетей. Рассмотрены следующие подходы к изучению личности: системный подход Гордона Олпорта, «Большая пятерка», «*HEXACO*» и «*MBTI*». Другие личностные системы в отличие от *MBTI*, такие как «Большая пятерка» или «*HEXACO*», обычно говорят о личностных чертах изолированно, что часто менее полезно при попытке концептуализировать человека в целом. Именно

по этой причине в данной работе предлагается использовать именно метод *MBTI*.

Во второй главе на основе результатов проведенного анализа предыдущей главы приводится математическое описание социальной сети, оценка сходства признаков выражения, оценка текстовых объектов, оценка сходства между двумя записями и оценка сходства пользователей в социальной сети.

Оценка сходства признаков выражения. Поскольку содержимое признака представлено в виде набора текстовых выражений, их сходство определяется следующим образом: предположим, что $C_1 = \{c_1^1, c_1^2, ..., c_1^m\}$, $C_2 = \{c_2^1, c_2^2, ..., c_2^n\}$ – два набора выражений или строк, в которых m, n - размеры множества C_1 и C_2. Сходство между C_1 и C_2 определяется формулой:

$$Sim(C_1, C_2) = \frac{2 \times |C_1 \text{ I } C_2|}{|C_1| + |C_2|} = \frac{2 \times t}{m + n}, \quad (1)$$

где t - размер множества пересечений C_1 и C_2.

Все возможные значения $Sim(C_1, C_2)$ лежат в интервале [0, 1]. Эту формулу можно применить к объектам, значение которых представляет собой набор выражений. Предположим, что, $f(e_i) = \left(f_1^i, f_2^i, ..., f_p^i \right)$, $f(e_j) = \left(f_1^j, f_2^j, ..., f_p^j \right)$ – две записи, представленные их функциями. Рассмотрим признак $sign$, значением которого является набор выражений. Сходство между записями e_i и e_j по признаку $sign$ определяется по следующей формуле:

$$S_{sign}(e_i, e_j) = Sim\left(f_{sign}^i, f_{sign}^j \right), \quad (2)$$

где f_{sign}^i, f_{sign}^j – значения выражения признака $sign$ двух записей e_i и e_j.

Оценка сходства текстовых объектов. Для анализа текстов предлагается использоваться метод *TF-IDF* (*TF* (*Term Frequency*) – частота термина (*n*-граммы – последовательности из *n* слов) в записи; *IDF* (*Inverse Document Frequency*) – обратная частота записи, указывающая на редкость *n*-граммы в наборе документов). *TF-IDF* – произведение *TF* и *IDF*, которое определяет вес каждой *n*-граммы в тексте. Чем больше *TF-IDF*, тем более значимой считается *n*-грамма для данного текста.

Предложенная методология состоит из следующих этапов:

1) разбивка текста на набор *n*-грамм $k_1 = \left(a_1^1, a_2^1, ..., a_n^1 \right)$ и $k_2 = \left(a_1^2, a_2^2, ..., a_m^2 \right)$.

2) вычисление *TF-IDF* для каждой *n*-граммы в тексте, где

$$TF = \frac{Количество\ вхождений\ n\text{-}граммы\ в\ записи}{Общее\ количество\ n\text{-}грамм\ в\ записи} \quad (3)$$

$$IDF = \log\left(\frac{Общее\ количество\ записей}{Количество\ записей\ содержащих\ n\text{-}грамму} + 1\right) \quad (4)$$

$$TF\text{-}IDF = TF \times IDF \quad (5)$$

После вычисления *TF-IDF* для каждой *n*-граммы текст представляется в виде вектора, где каждая *n*-грамма соответствует первому элементу вектора, а второй элемент – его значение (*TF-IDF* вес *n*-граммы). $< n\text{-}gram, TF\text{-}IDF >$: $t^1 = (< a_1^1, t_1^1 >, < a_2^1, t_2^1 >, ..., < a_n^1, t_n^1 >)$ и $t^2 = (< a_1^2, t_1^2 >, < a_2^2, t_2^2 >, ..., < a_m^2, t_m^2 >)$.

3) определение расстояния между двумя векторами:

$$D(k^1, k^2) = \frac{1}{N} \sum_{i=1}^{N} d_k , \quad (6)$$

где N – количество различных *n*-грамм, рассматриваемых в $k^1 \cup k^2$, d_k – расстояние на каждом элементе $< a_i^1, \mathrm{t}_i^1 >$ из t^1 (или элемент $< a_j^2, \mathrm{t}_j^2 >$ из t^2, соответственно), которое вычисляется следующим образом: если существует элемент $< a_l^2, \mathrm{t}_l^2 >$ из t^2 (или элемент $< a_l^1, \mathrm{t}_l^1 >$ из t^1, соответственно) такой, что $a_l^2 = a_l^1$ тогда:

$$d_k = \frac{\left| t_i^1 - t_l^2 \right|}{\max\left\{ t_i^1, t_l^2 \right\}} \quad (7)$$

В противном случае, если *n*-грамма присутствует только в одном из векторов, то расстояние $d_k = 1$. Значение $D(k^1, k^2)$ находится в интервале [0, 1]. Тогда сходство между двумя текстовыми объектами заключается в следующем:

$$S_{Sim}(k^1, k^2) = 1 - D(k^1, k^2) \quad (8)$$

Чем ближе значение к 1, тем тексты более похожи; чем ближе к 0, тем они более различны.

Для измерения схожести между двумя записями было применено 5 критериев: контент, теги, категоризация, настроение и эмоции. Последние четыре особенности выражения записи, оцениваются как сходство по признаку выражения следующим образом:

$$S_{cate}(e_i, e_j) = Sim(f_{cate}(e_i), f_{cate}(e_j)); \quad (9)$$

$$S_{tags}(e_i, e_j) = Sim(f_{tags}(e_i), f_{tags}(e_j)); \quad (10)$$

$$S_{sent}(e_i, e_j) = Sim(f_{sent}(e_i), f_{sent}(e_j)); \quad (11)$$

$$S_{emot}(e_i, e_j) = Sim(f_{emot}(e_i), f_{emot}(e_j)). \quad (12)$$

Одним из текстовых признаков записи является ее содержание, поэтому оно оценивается как сходство текстовых признаков, рассчитываемое следующим образом:

$$S_{cont}(\mathrm{e}_i, \mathrm{e}_j) = Sim(f_{cont}(e_i), f_{cont}(e_j)) \quad (13)$$

Пусть e_i и e_j две рассматриваемые записи, значений функций, содержание, теги, категории, настроения и эмоции которых являются

характеристиками записей: $f_{cont}^{i}; f_{cont}^{j}; f_{tags}^{i}; f_{tags}^{j}; f_{cate}^{i}; f_{cate}^{j}; f_{sent}^{i}; f_{sent}^{j}; f_{emot}^{i}; f_{emot}^{j}$. На основе подхода многоатрибутного сходства двух объектов сходство между двумя записями e_i и e_j оценивается следующим образом:

$$f_{entry}(e_i,e_j) = f_{ent}(S_{cont}(e_i,e_j), S_{tags}(e_i,e_j), S_{cate}(e_i,e_j), S_{sent}(e_i,e_j), S_{emot}(e_i,e_j)), \quad (14)$$

где $f_{ent}:[0,1]^5 \to [0,1]$. Исходя из того, что сходство между двумя записями формируется на основе сходства этих записей по пяти свойствам, которые определяются введенными нами функциями, следует потребовать выполнение последующих свойств:

$$f_{ent}(v_1,w,x,y,z) \le f_{ent}(v_2,w,x,y,z) \ \ if \ \ v_1 \le v_2 \quad (15)$$

$$f_{ent}(v,w_1,x,y,z) \le f_{ent}(v,w_2,x,y,z) \ \ if \ \ w_1 \le w_2 \quad (16)$$

$$f_{ent}(v,w,x_1,y,z) \le f_{ent}(v,w,x_2,y,z) \ \ if \ \ x_1 \le x_2 \quad (17)$$

$$f_{ent}(v,w,x,y_1,z) \le f_{ent}(v,w,x,y_2,z) \ \ if \ \ y_1 \le y_2 \quad (18)$$

$$f_{ent}(v,w,x,y,z_1) \le f_{ent}(v,w,x,y,z_2) \ \ if \ \ z_1 \le z_2 \quad (19)$$

Представленные свойства характеризуют тот факт, что увеличение расхождения по частному признаку должно вести к не уменьшению общего расхождения. Данные функции реализованы в разработанной программе (Свидетельство о регистрации программы для ЭВМ №2022662518 от 05.07.2022г.).

Определение сходства между двумя пользователями происходит путем оценки сходства по каждому типу поведения с использованием средневзвешенной агрегации. Таким образом, общее сходство между ними по всем рассматриваемым видам поведения представляется следующим образом. Пусть w_1, w_2, w_3, w_4 – это веса, отражающие степень схожести на основе размещения/публикации, присоединения к группе, постановки лайков под записями и комментариев/лайков к комментариям соответственно. Они должны удовлетворять условию $w_1 + w_2 + w_3 + w_4 = 1$.

Сходство между пользователем U_1 и пользователем U_2:

$$S(U_1,U_2) = w_1 \times S_{publ}(U_1,U_2) + w_2 \times S_{conn}(U_1,U_2) + \\ + w_3 \times S_{like}(U_1,U_2) + w_4 \times S_{komm}(U_1,U_2), \quad (20)$$

где S_{publ} – сходство между двумя пользователями на основе публикации;

S_{conn} – сходство между двумя пользователями и на основе присоединения к тому или иному сообществу;

S_{like} – сходство между двумя пользователями и на основе лайков к публикациям, комментариям;

S_{komm} – сходство между двумя пользователями и на основе комментариев.

Также во второй главе рассмотрен метод объединения информации из двух социальных сетей, который необходим для составления более полного и подробного психологического портрета пользователей. В разных социальных сетях могут содержаться данные о пользователях в разные промежутки времени, и эти данные могут быть дополняющими или самостоятельными.

Например, в одной сети могут быть данные о пользовательской активности в определенный период времени, в то время как в другой сети такие данные могут отсутствовать, но могут быть доступны другие сведения о поведении пользователя.

Помимо этого, приведено описание предлагаемой модели *IbDA-LSTM-CRF* для решения задачи совокупного выделения извлечения аспектов и тональности отношения к ним авторов текстов, относящимся к различному контексту и на разные темы. На входе в модель имеется текст, который с помощью токенизации преобразован в последовательность токенов $S = \{w_1, w_2, ..., w_n\}$, где n – длина входной последовательности. Каждому токену w_i из $S(i \in [1,...,n], i \in \ddot{y})$ требуется сопоставить метку y_i, $y_i \in \{0, B-POS, I-POS, B-NEG, I-NEG, B-NEU, I-NEU\}$.Таким образом, модель ставит в соответствие каждой входной последовательности S последовательность $Y = \{y_1, y_2, ..., y_n\}$, представленную в *BIO*-формате. Таким образом, разметка последовательности, составленная для токенов-слов, должна учитывать это разбиение. В данной работе при разбиении слова на несколько токенов все получившиеся подтокены сохраняют метку целого слова. При обучении модели для решения кросс-доменной модификации задачи *ABSA* имеется размеченный набор данных из домена-источника: $Data_S = \{(S_k, Y_k)\}^{n_s}$, где (S_k, Y_k) -– один обучающий объект, представленный в виде пары «последовательность – *BIO*-разметка», а n_S – количество таких объектов. Также пусть имеется набор данных из целевого домена $Data_T = \{S_j\}^{n_T}$, где n_T – количество текстов из целевого домена. Конечная задача модели – генерация *BIO*-разметки аспектов и тональности Y_j для каждой последовательности S_j из $Data_T$. Так, при наличии размеченных данных из домена-источника и неразмеченных данных из целевого домена, необходимо адаптировать модель для работы сразу в двух доменах. Предлагаемая в работе модель учитывает специфичность токенов из входных последовательностей для обоих доменов и путем взвешивания функции ошибки на каждом из токенов последовательностей домена-источника пытается придать больший вес токенам, близким к целевому домену.

Предлагаемая архитектура модели кросс-доменного аспектно-ориентированного анализа тональности состоит из двух частей: главная часть отвечает за выделение аспектов и анализ тональности отношения к ним в совместном виде, а вспомогательная часть отвечает за генерацию доменнозависимых весов для функции потерь, тем самым адаптируя главную часть модели под целевой домен.

В **третьей главе** приведено подробное описание созданного программного продукта, алгоритм работы которого представлен на рисунке 1.

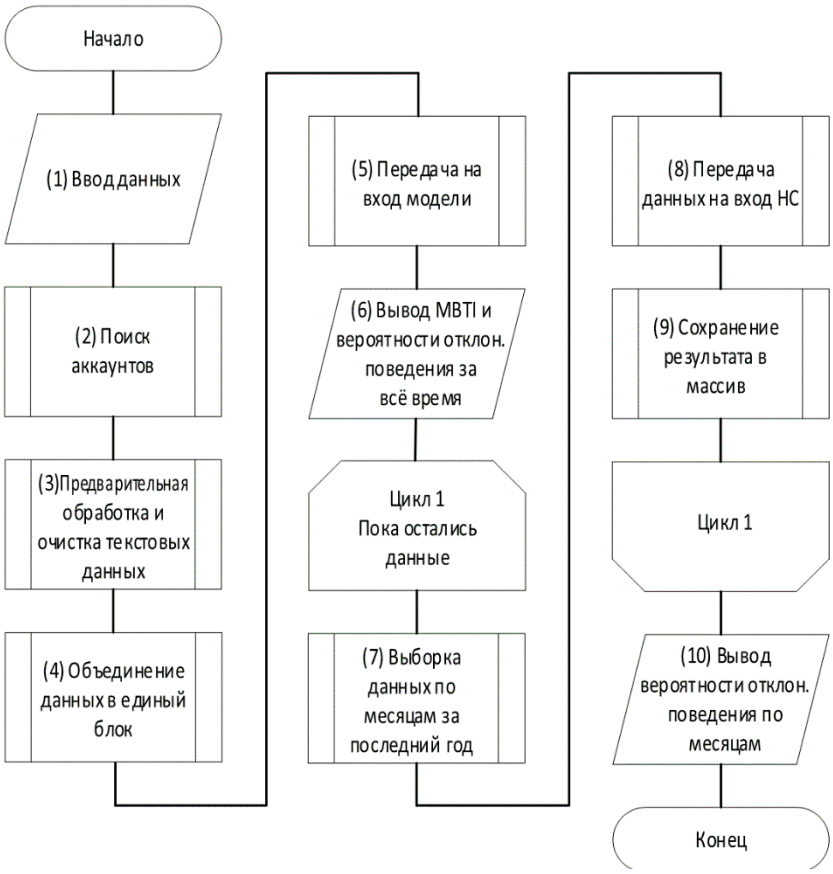

Рисунок 1 – Алгоритм работы созданного программного обеспечения

Алгоритм поиска аккаунтов пользователя представлен на рисунке 2.

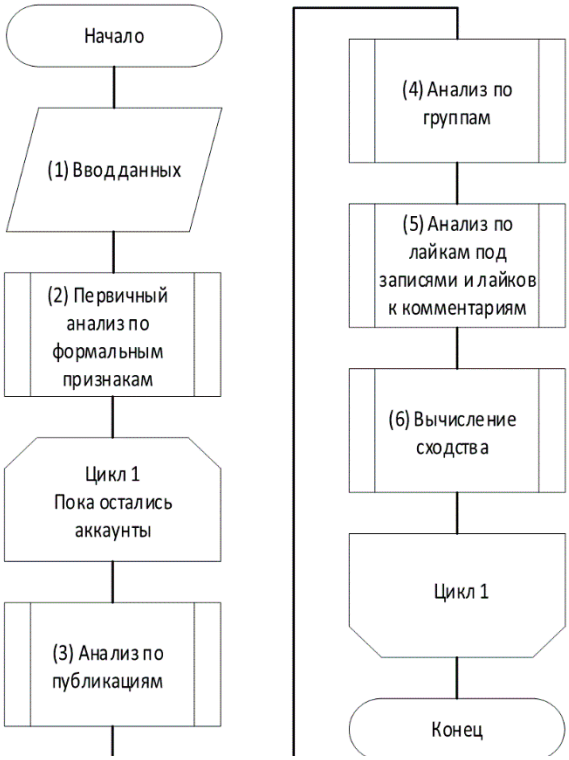

Рисунок 2 – Алгоритм работы поиска аккаунтов пользователя

Алгоритм работы модели *IbDA-LSTM-CRF* изображен на рисунке 3.

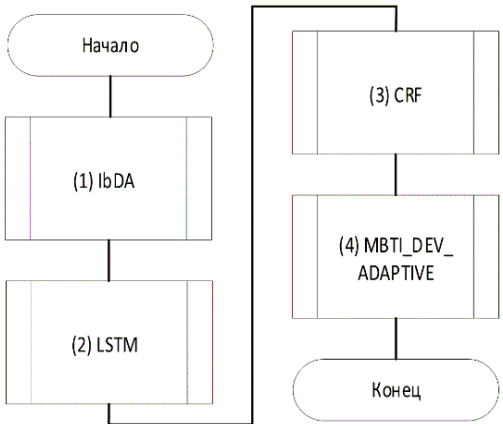

Рисунок 3 – Алгоритм работы модели *IbDA-LSTM-CRF*

Разработанные алгоритмы интегрированы в программное обеспечение, которое предоставляет возможность проводить анализ психологического портрета пользователя и оценивать вероятность отклоняющегося поведения. Это осуществляется в режиме реального времени, что позволяет учитывать данные за заданный временной интервал, обеспечивая тем самым более точное и актуальное представление о состоянии пользователя.

В четвертой главе представлены данные, подтверждающие эффективность применения разработанных алгоритмов.

1. Эксперименты по нахождению профилей пользователей в разных социальных сетях («ВКонтакте» и «Одноклассники»).

Для проведения исследования была отобрана заранее подготовленная выборка из 153 человек, которые активно использовали социальные сети. Размер этой выборки составил 261 аккаунт в различных социальных сетях. Важно отметить, что не все участники имели несколько аккаунтов в разных социальных сетях. Диаграмма распределения пользователей по количеству аккаунтов в разных социальных сетях представлена на рисунке 4.

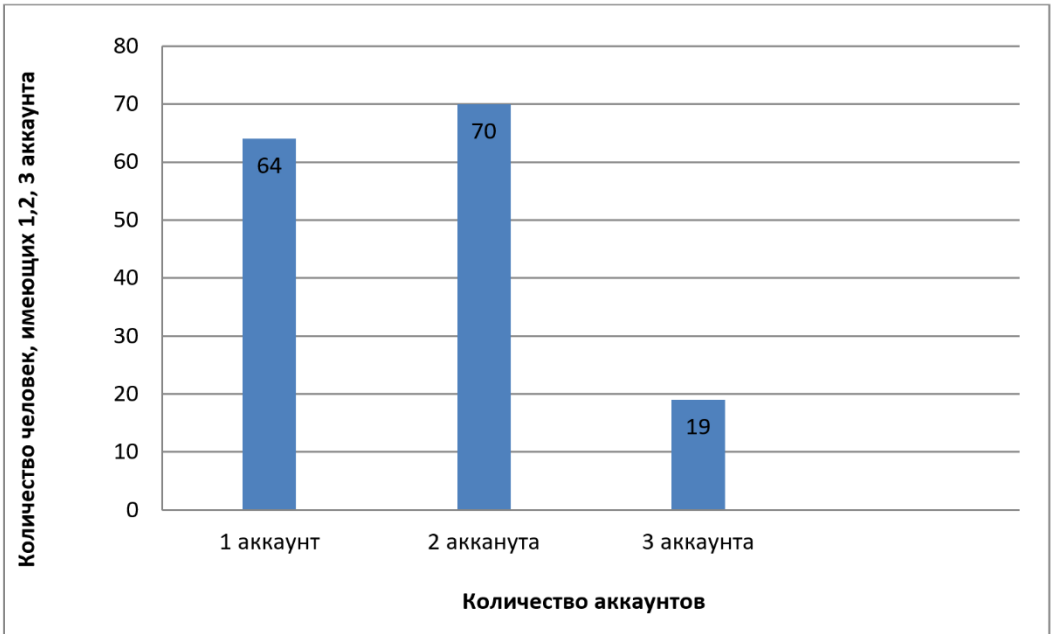

Рисунок 4 – Диаграмма распределения пользователей по количеству аккаунтов

Процентное соотношение корректно найденных аккаунтов можно увидеть на рисунке 5. Следует учесть, что в сюда входят как одиночные аккаунты, так и пользователи с несколькими аккаунтами.

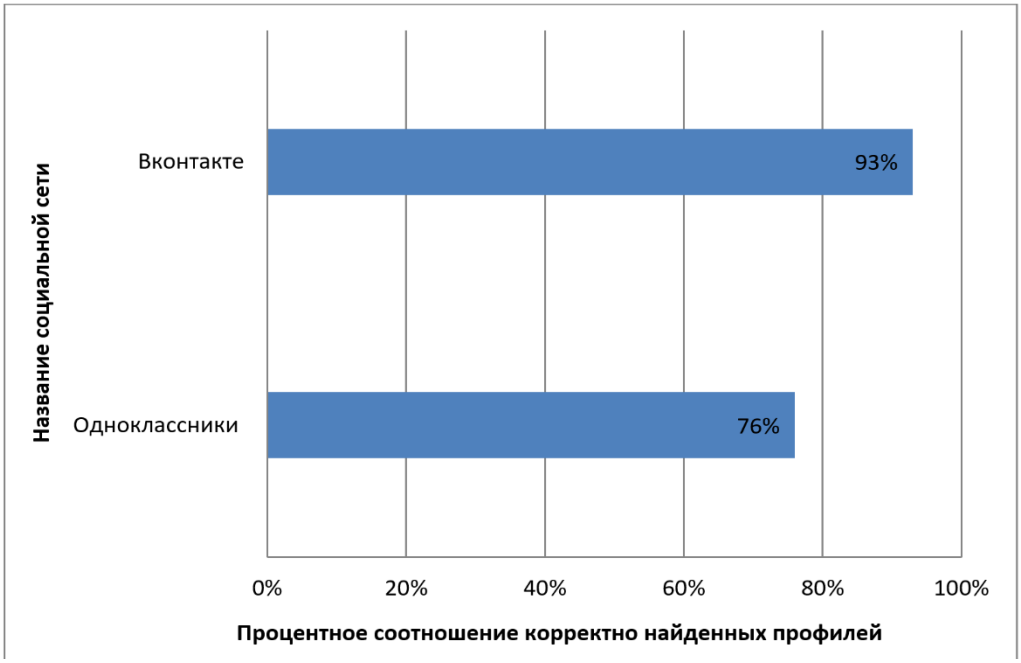

Рисунок 5 – Диаграмма процентного соотношения корректно найденных профилей

Достаточно высокие показатели правильно найденных профилей во «ВКонтакте» (93%) и «Одноклассниках» (76%) свидетельствуют о том, что разработанная система эффективно выполняет свою задачу по идентификации и визуализации пользователей. Значения совпадений указывают на то, что алгоритмы, используемые в системе, способны точно сопоставлять пользователей на разных платформах.

Кроме того, были проанализированы результаты сравнения профилей в рамках одной и той же социальной сети. Результат показан на рисунке 6.

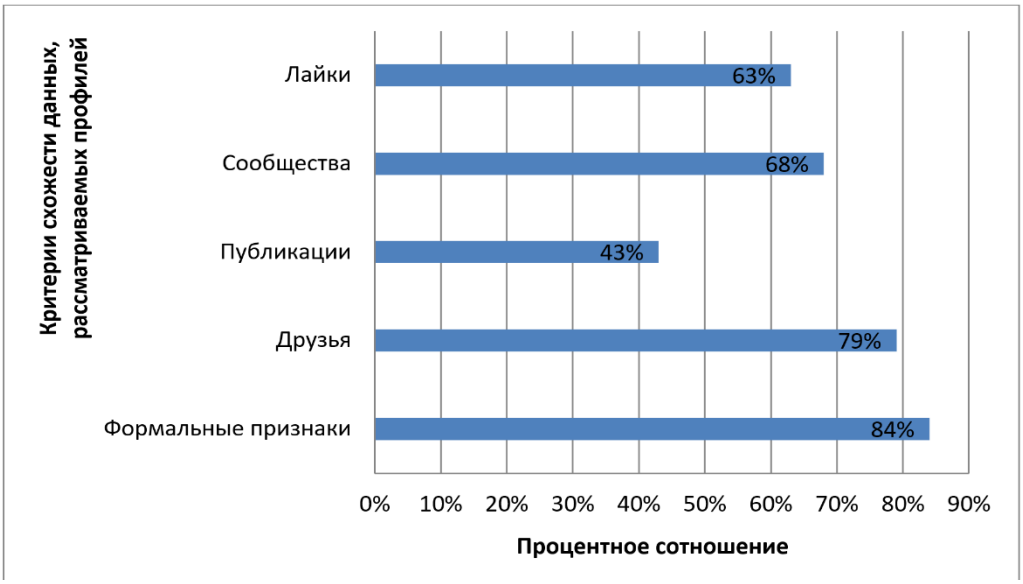

Рисунок 6 – Критерии сравнения профилей в пределах одной социальной сети

Полученные данные показали, что наблюдается высокая степень совпадения в списке друзей и подписках на сообщества между различными профилями одного пользователя. Это объясняется тем, что предпочтения пользователя в отношении друзей и сообществ остаются преимущественно постоянными в различных профилях внутри одной социальной сети. В трети проведенных экспериментов обнаружено, что профили пользователей имеют общие записи на своих страницах. Но это явление оказалось нечастым, что объясняется непостоянством в содержании профилей: пользователи не всегда заполняют свои страницы одинаковыми постами, в пределах одной социальной сети.

В контексте сравнения аккаунтов одного пользователя в различных социальных сетях (рисунок 7), наблюдается низкий уровень схожести в отношении подписок на сообщества. Это обусловлено уникальностью сообществ для каждой социальной платформы. Однако, процент совпадения публикаций оказывается выше, что объясняется частым копированием контента из одной социальной сети в другую.

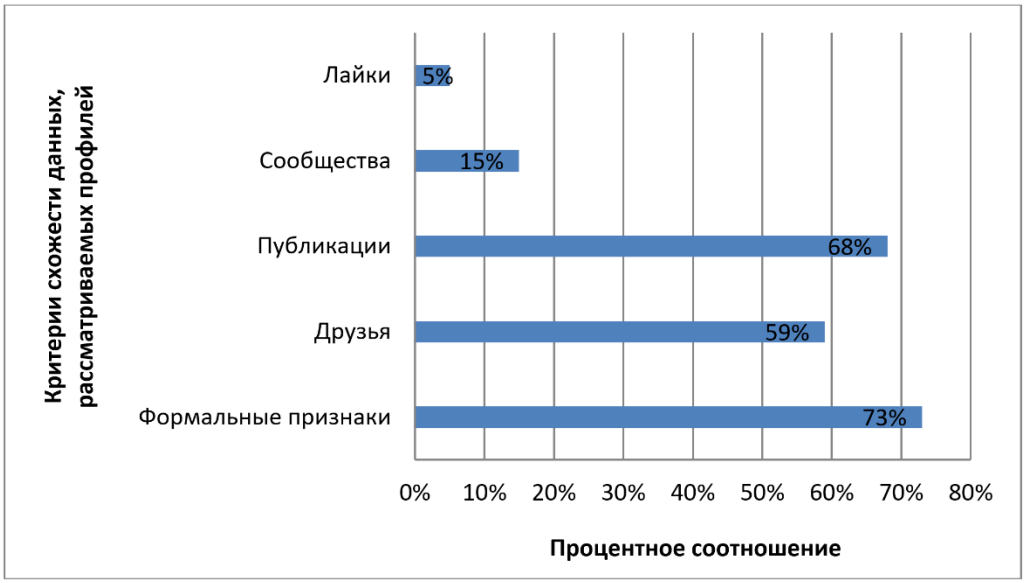

Рисунок 7 – Критерии сравнения профилей в различных социальных сетях

2. Эксперименты по определению психологического портрета пользователя и вероятности отклоняющегося поведения.

Обучение нейронной сети для определения психологического портрета и вероятности отклоняющегося поведения пользователя происходило в течение 100 эпох *LSTM*-архитектурой. Для обучения использовалась обучающая выборка, состоящая из 29000 записей из социальных сетей, включая публикации, сообщества, комментарии, лайки и другие активности пользователей и базы данных «*MBTI Dataset*», «*Personality Recognition Dataset*» и др. Общая выборка была разделена на обучающую и тестовую в соотношении 70/30.

Для проверки эффективности обучения нейронной сети использовались функции потерь и метрики, оценивающие сходство между истинными и

предсказанными данными. В данной задаче использовалась комбинация двух функций потерь: кросс-энтропия и коэффициент Дайса (таблица 1).

Таблица 1 – Статистика обучения нейронной сети LSTM.

№ эпохи	Кросс-энтропия (обучение)	Коэффициент Дайса (обучение)	Кросс-энтропия (тестовый)	Коэффициент Дайса (тестовый)
1	0,723	0,81	0,768	0,795
2	0,642	0,854	0,758	0,824
3	0,598	0,858	0,632	0,826
4	0,567	0,86	0,582	0,835
5	0,532	0,861	0,559	0,841
6	0,543	0,862	0,56	0,853
7	0,502	0,865	0,518	0,859
8	0,485	0,875	0,501	0,863
9	0,469	0,877	0,515	0,867
10	0,47	0,88	0,471	0,875
...	...	...	...	...
95	0,204	0,9982	0,226	0,981
96	0,202	0,9983	0,226	0,979
97	0,201	0,9984	0,227	0,979
98	0,201	0,998	0,225	0,983
99	0,2	0,998	0,227	0,985
100	0,199	0,998	0,223	0,986

Графики на рисунках 8 и 9 демонстрируют основные результаты обучения нейронной сети *LSTM*.

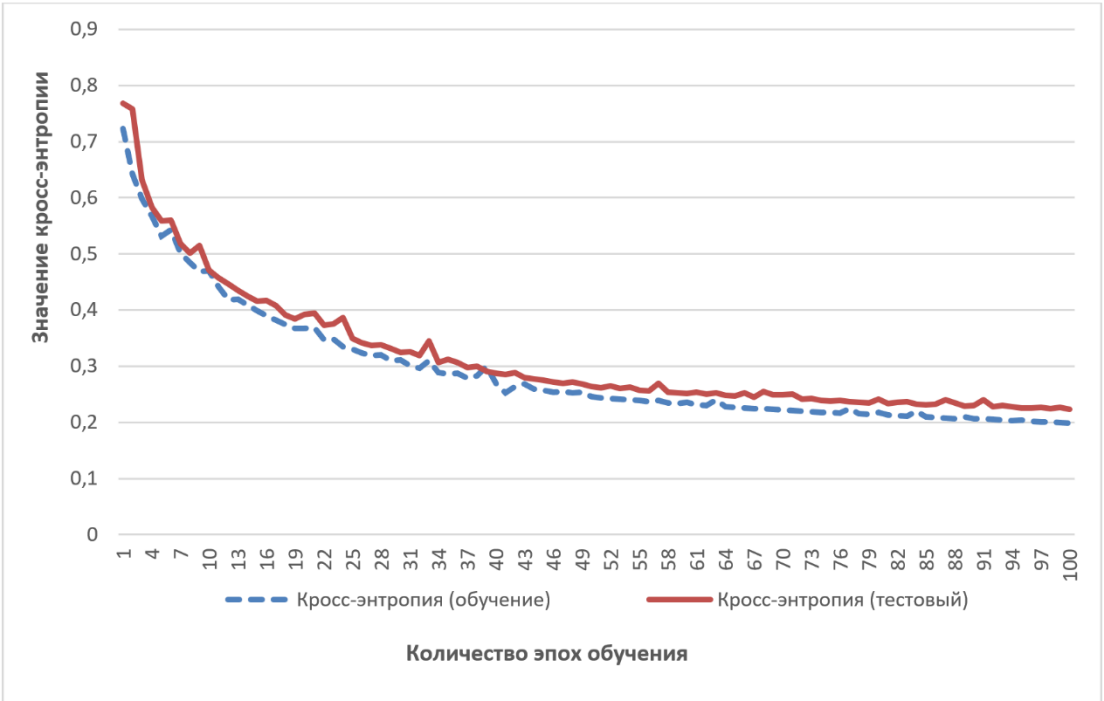

Рисунок 8 – Изменение функции потерь во время обучения

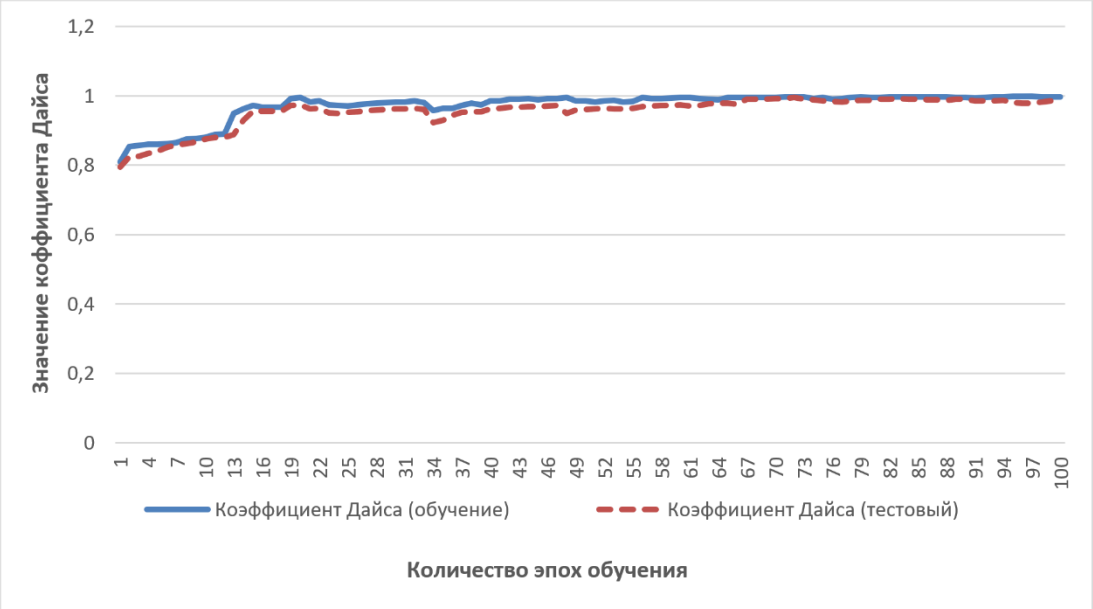

Рисунок 9 – Изменения функции схожести во время обучения

В ходе исследований проведён анализ эффективности и точности различных алгоритмов в задаче классификации текстов с целью выявления психологических характеристик авторов. Эксперименты были направлены на сопоставление нескольких алгоритмов классификации с точки зрения их эффективности при использовании модели *IbDA-LSTM-CRF*. Результаты сравнения приведены в таблице 2.

Таблица 2 – Сравнение различных методов классификации

	Precision	*Recall*	*F*-мера
LSTM	0,94	0,95	0,945
KNN	0,89	0,91	0,89
Random Forest	0,88	0,89	0,88

Лучший результат показал *LSTM*, достигающий точности 0,95, благодаря самому высокому показателю *F*-меры, который является средним между точностью и полнотой.

Как видно из графика, представленного на рисунке 10, *LSTM* показывает наилучшие результаты точности типов личности *INTJ, ENTJ, INFP, ENFP, ISTJ, ISFJ, ISTP, ISFP, ESTP, ESFP* (от 0,93 до 0,96). Это указывает на то, что *LSTM* справляется лучше всего с классификацией этих типов личности.

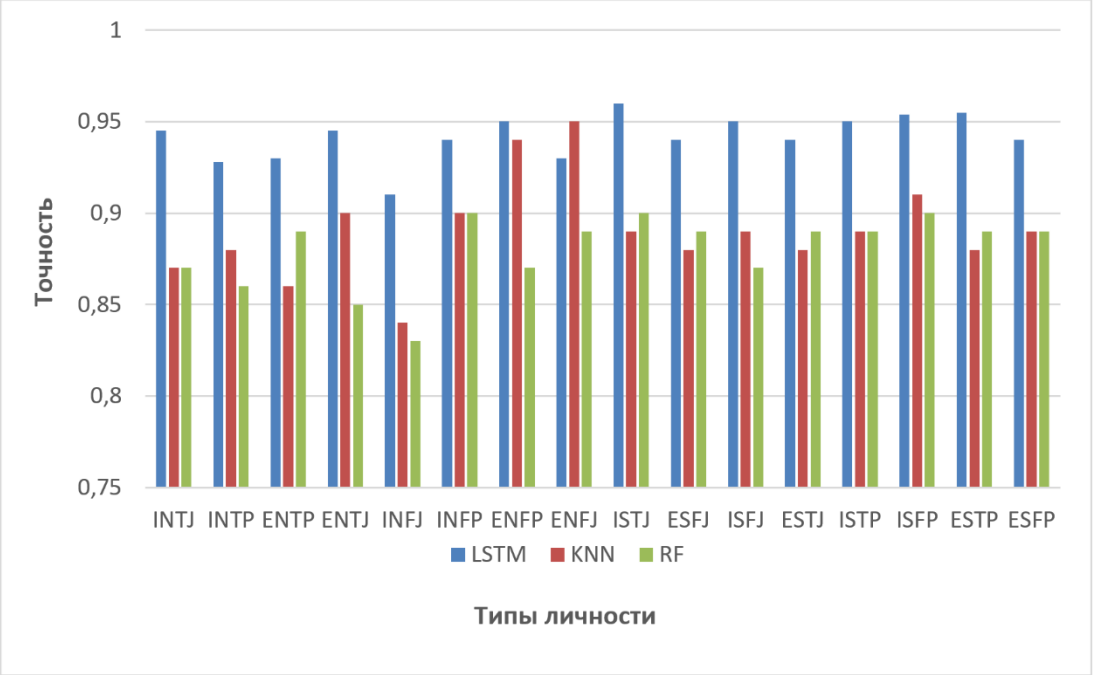

Рисунок 10 – Результаты проведенных экспериментов

Для следующих типов личности, таких как *INTP, INFJ, ESFJ, ESTJ*, результаты классификации *LSTM* также выше, хотя не столь значительно, как для других типов личности. Для таких типов личности, как *ENTP, ISFP, KNN* и *RF* показывают более высокую точность по сравнению с *LSTM*. Это может указывать на то, что для определенных типов личности другие методы классификации могут быть более эффективными.

Таким образом, предлагаемая модель позволяет более глубоко погрузиться в контекст данных, собранных из различных социальных сетей. Это обеспечивает более точное определение смысла текстов, соответствующих различным типам личности в соответствии с *MBTI*, и более точную оценку вероятности нестандартного поведения. Комбинирование этого метода позволяет учитывать не только сами слова, но и их контекст, а также взаимосвязи между различными элементами текста. Это обеспечивает высокую точность классификации и более глубокое понимание текстовых данных, что особенно важно при анализе психологических характеристик и поведения пользователей в социальных сетях.

Следует отметить, что полученные данные прошли валидацию. В эксперименте приняли участие 153 пользователя, прошедших психологическое тестирование по шкале *MBTI*. Тестирование было проведено профессиональным экспертом, который классифицировал каждого пользователя по одной из 16 категорий *MBTI*. Этот результат стал эталонным для сравнения с результатами нейросетевой модели. Разработанная модель достигла точности 0,95 в предсказании типа личности. Таким образом, в 95% случаев тип личности, предсказанный нейросетевой моделью, совпадал с результатом, который был определён экспертом.

В заключении сформулированы основные выводы и перечислены результаты, полученные в ходе работы по формированию психологического портрета пользователя для эффективного подбора кадров.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

В ходе подготовки диссертации получены следующие результаты, имеющие существенное значение для развития страны:

1. Проведено исследование методов и моделей интеллектуального анализа данных пользователей социальных сетей.

2. Разработан метод сравнительного анализа признаков выражений и текстовых объектов пользователей с целью выявления однотипных аккаунтов в социальных сетях, предвосхищая потенциальные угрозы безопасности.

3. Разработан метод интеграции данных, размещаемой пользователем в разных социальных сетях, который позволяет восстанавливать данные активности, учитывая разнообразные аспекты его онлайн-поведения, для составления более полного и подробного психологического портрета и определения отклоняющегося поведения.

4. Предложена методика кросс-доменного аспектно-ориентированного анализа тональности текста *IbDA-LSTM-CRF*, которая решает проблему аспектно-ориентированного анализа тональности, т.к. в свою очередь, она, обученная на постах одной тематики, не может эффективно обрабатывать посты другой тематики, так как не обладает свойством извлекать информацию из терминов и выражений, специфичных для профиля (домена) последнего. Данная методика учитывает контекст и особенности каждого текста, независимо от тематики и смыслового контекста.

5. Разработана нейросетевая методика определения психологических характеристик пользователя социальной сети, с использованием типологии *MBTI*. Точность классификации достигает 0,93-0,96.

6. Проведено экспериментальное исследование предлагаемых методов и алгоритмов, на основе которого были сформулированы рекомендации по их использованию.

7. Разработан программный комплекс определения психологического портрета пользователя и вероятности нестандартного поведения, применение которого в ООО «ТД «ПЗЭМ» позволило повысить эффективность управления кадровой системы на 13%.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях из перечня ВАК Минобрнауки России

1. Мартышкин А.И., Перекусихина А.Н., Зоткина А.А. Исследование групп пользователей в социальных сетях по их интересам и поведению на основе множества источников данных // XXI век: итоги прошлого и проблемы настоящего плюс. – 2020. – Т. 9. – № 4 (52). – С. 30-35.

2. Зоткина А.А., Мартышкин А.И., Новоселова О.В. Методика оптимизации обучающего алгоритма нейронных сетей // XXI век: итоги прошлого и проблемы настоящего плюс. – 2021. – Т. 10. – № 4 (56). – С. 21-24.

3. Зоткина А.А., Мартышкин А.И. Анализ полярности настроений пользователей социальных сетей в период COVID-19 // XXI век: итоги прошлого и проблемы настоящего плюс. – 2022. – Т. 11. – № 1 (57). – С. 15-18.

4. Зоткина А.А. Анализ депрессивного состояния пользователей социальной сети «ВКонтакте» // XXI век: итоги прошлого и проблемы настоящего плюс. – 2022. – Т. 11. – № 4 (60). – С. 52-55.

5. Зоткина А.А., Мартышкин А.И. Применение методов машинного обучения в задаче прогнозирования киберзапугивания пользователей социальной сети // Современные наукоемкие технологии. – 2022. – № 10-2. – С. 249-253.

6. Зоткина А.А., Мартышкин А.И. Обнаружение депрессии среди пользователей социальной сети с использованием методов машинного обучения // Computational Nanotechnology. – 2023. – Т. 10. – № 4. – С. 16-22.

Публикации в изданиях, индексируемых базой Scopus

7. Zotkina, A.A., Martyshkin, A.I. Identification of a Depressive State among Users of the Vkontakte Social Network // Proceedings – 2023 International Russian Smart Industry Conference, SmartIndustryCon 2023. – 2023. – pp. 335-339.

8. Zotkina AA., Martyshkin A.I., Detection of Cyberbullying in Texts Posted by Users of Social Networks Using Machine Learning, 2024 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2024. – pp. 639-643.

Публикации в прочих изданиях

9. Ильичов Д.Э., Лысцов Н.А., Зоткина А.А. Основные характеристики и алгоритм обучения нейронных сетей // Наука и образование в современном обществе: актуальные вопросы и инновационные исследования: сборник статей III Международной научно-практической конференции. Пенза, 2021. – С. 25-27.

10. Ильичов Д.Э., Лысцов Н.А., Зоткина А.А. Характеристики и математическое описание нейрона // Наука и образование в современном обществе: актуальные вопросы и инновационные исследования: Сборник статей III Международной научно-практической конференции. Пенза, 2021. – С. 28-30.

11. Зоткина А.А., Мартышкин А.И. Персептрон как простейший вид искусственной нейронной сети на примере построения однослойной модели сети // Современные методы и средства обработки пространственно-временных сигналов: Сборник статей XIX Всероссийской научно-технической конференции, посвященной 60-летию первого полета в космос Юрия Алексеевича Гагарина. Под редакцией И.И. Сальникова. Пенза, 2021. – С. 33-38.

12. Мартышкин А.И., Зоткина А.А. К вопросу профилирования пользователей социальных сетей // Современные информационные технологии. – 2021. – № 34 (34). – С. 77-81.

13. Зоткина А.А., Мартышкин А.И. Анализ методов определения тональности текстовых данных пользователя социальных сетей // Современные информационные технологии. – 2021. – № 34 (34). – С. 81-84.

14. Зоткина А.А., Шиндина Н.С. Обзор существующих параметров обработки естественного языка // Современные научные исследования: актуальные вопросы, достижения и инновации: Сборник статей XXIII Международной научно-практической конференции. Пенза, 2022. – С. 56-58.

15. Зоткина А.А. Обзор интерфейса прикладного программирования-*API* как метода для взаимодействия и извлечения информации // Достижения в науке и образовании 2022: сборник статей Международного научно-исследовательского конкурса. Пенза, 2022. – С. 34-36.

16. Зоткина А.А. Рекуррентные нейронные сети как алгоритм последовательности данных // Современные информационные технологии. – 2022. – № 35 (35). – С. 24-26.

17. Мартышкин А.И., Зоткина А.А. Обзор существующих методов анализа настроений пользователей социальных сетей // Современные информационные технологии. – 2022.– № 35 (35). – С. 70-72.

18. Мартышкин А.И., Зоткина А.А. Особенности работы сверточных нейронных сетей: архитектура и применение // Современные информационные технологии. – 2022. – № 36 (36). – С. 11-13.

19. Мартышкин А.И., Киндаев А.Ю., Зоткина А.А., Поленова Т.А. Базовые составляющие центров обработки данных // Современные информационные технологии. – 2022. – № 36 (36). –С. 13-16.

20. Зоткина А.А. Анализ настроений пользователей социальных сетей как инструмент прогнозирования трендов // Современные информационные технологии. – 2022. – № 36 (36). – С. 77-79.

21. Зоткина А.А., Шиндина Н.С. Интерфейс прикладного программирования // Современные информационные технологии. – 2022. – №36 (36). – С. 79-82.

22. Зоткина А.А., Ткаченко А.В. Обработка данных при помощи рекуррентной нейронной сети // Современные методы и средства обработки пространственно-временных сигналов: сборник статей XIX Всероссийской научно-технической конференции. Под редакцией И.И. Сальникова. Пенза, 2023. – С. 98-101

23. Зоткина А.А., Павлов А.А. Описание общих признаков портрета пользователя социальной сети «ВКонтакте» // Современные методы и средства обработки пространственно-временных сигналов: сборник статей XIX Всероссийской научно-технической конференции. Под редакцией И.И. Сальникова. Пенза, 2023. – С. 95-98

24. Зоткина А.А., Мартышкин А.И. Известные методы анализа настроений пользователей социальных сетей // Современные методы и

средства обработки пространственно-временных сигналов: сборник статей XIX Всероссийской научно-технической конференции. Под редакцией И.И. Сальникова. Пенза, 2023. – С. 28-32

25. Мартышкин А.И., Зоткина А.А. Сбор данных из социальных сетей для анализа профиля человека // Современные информационные технологии. – 2023. – № 38 (38). – С. 96-100.

26. Мартышкин А.И., Зоткина А.А. Проблемы девиантного поведения пользователей социальных сетей // Современные информационные технологии. – 2023. – № 38 (38). – С. 93-96.

27. Зоткина А.А., Мартышкин А.И. Системы мониторинга социальных сетей // Современные информационные технологии. – 2023. – № 38 (38). – С. 69-73.

28. Зоткина А.А., Холкина В.М. Обзор методов анализа настроений // Современные информационные технологии. – 2023. – № 38 (38). – С. 55-59.

29. Зоткина А.А., Мартышкин А.И. Определение данных для обучения нейронных сетей, предназначенных для анализа отклоняющегося поведения пользователей // Современные информационные технологии. – 2023. – № 38 (38). – С. 35-37.

30. Зоткина А.А. Психологическое профилирование пользователей социальных сетей при помощи машинного обучения // Современные информационные технологии. – 2023. – № 37 (37). – С. 145-147.

31. Зоткина А.А., Мартышкин А.И. LIWC как метод компьютерной лингвистики и обработки естественного языка // Современные информационные технологии. – 2023. – № 37 (37). – С. 134-137.

32. Зоткина А.А., Шиндина Н.С. Решение проблем рекуррентной нейронной сети при помощи модели "долговременной кратковременной памяти" // Современные информационные технологии. – 2023. – № 37 (37). – С. 18-20.

33. Зоткина А.А., Шиндина Н.С. Основные задачи NLP и как их решают нейронные сети // Современные информационные технологии. – 2023. – № 37 (37). – С. 14-17.

34. Зоткина А.А., Холкина В.М., Балаба У.Н. Векторизация текста при помощи модели *BERT* // Современные информационные технологии. – 2024. – № 39 (39). – С. 16-19.

35. Мартышкин А.И., Зоткина А.А. Основные проблемы в области определения тональности текста // Современные информационные технологии. – 2024. – № 39 (39). – С. 85-88.

36. Мартышкин А.И., Зоткина А.А. Некоторые подходы к определению тональности текста // Современные информационные технологии. – 2024. – № 39 (39). – С. 88-92.

37. Зоткина А.А. Графовое представление структуры социальной сети // Современные информационные технологии. – 2024. – № 39 (39). – С. 96-98.

38. Зоткина А.А., Мартышкин А.И. Программа для автоматизированной очистки базы гетерогенных данных // Современные информационные технологии. – 2024. – № 39 (39). – С. 102-105.

Свидетельства о государственной регистрации программ для ЭВМ

1. Свидетельство о государственной регистрации программ для ЭВМ № 2022662518. Программа для анализа архетипов пользователей социальных сетей с использованием открытых данных профиля // Зоткина А.А., Мартышкин А.И., Данилов Е.А. Пензенский государственный технологический университет. 05.07.2022.

2. Свидетельство о государственной регистрации программ для ЭВМ № 2023682329. Программа для автоматизированной очистки базы гетерогенных данных // Зоткина А.А., Мартышкин А.И. Пензенский государственный технологический университет. 24.10.2023.

ЗОТКИНА АЛЕНА АЛЕКСАНДРОВНА

МЕТОДЫ И АЛГОРИТМЫ ФОРМИРОВАНИЯ ПСИХОЛОГИЧЕСКОГО ПОРТРЕТА ПОЛЬЗОВАТЕЛЯ СОЦИАЛЬНОЙ СЕТИ ДЛЯ ЭФФЕКТИВНОГО ПОДБОРА КАДРОВ

Специальность 2.3.8. – Информатика и информационные процессы

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Подписано в печать __.10.2024. Формат 60x84 $^{1}/_{16}$
Бумага офсетная. Печать цифровая. Усл. печ. л. 1,2.
Тираж 100 экз.
Пензенский государственный технологический университет.
440039, Россия, г. Пенза, пр. Байдукова/ул. Гагарина, $1^{а}$/11