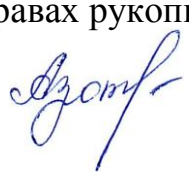
МИНОБРНАУКИ РФ
ФГБОУ ВО «ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ
ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

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

Диссертация
на соискание ученой степени
кандидата технических наук

Научный руководитель:
кандидат технических наук,
доцент МАРТЫШКИН А.И.

Пенза – 2024

ОГЛАВЛЕНИЕ

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ, СОКРАЩЕНИЯ

ИС – информационная система.

КАНОЭ (или ОКЕАН) – пятифакторная модель личности.

НЛП – нейролингвистическое программирование.

РСУБД – реляционная система управления базами данных.

API (Application programming interface) – программный интерфейс приложения.

BERT (Bidirectional Encoder Representations from Transformers) – языковая модель, основанная на архитектуре трансформер, предназначенная для предобучения языковых представлений с целью их последующего применения в широком спектре задач обработки естественного языка.

CRF (Conditional Random Fields) – статистическая модель последовательности, которая используется для моделирования зависимостей между последовательными метками, такими как метки частей речи или метки именованных сущностей.

Data Science – наука о данных.

DRF (Django REST Framework) – набор инструментов для создания веб-сервисов и *API* на основе фреймворка *Django*.

ENFJ (Extraverted, Intuition, Feeling, Judging) – психотип «даритель», согласно *MBTI*.

ENFP (Extraverted, Intuition, Feeling, Perception) – психотип «чемпион», согласно *MBTI*.

ENTJ (Extraverted, Intuition, Thinking, Judging) – психотип «предприниматель», согласно *MBTI*.

ENTP (Extraverted, Intuition, Thinking, Perception) – психотип «политик», согласно *MBTI*.

ESFJ (Extraverted, Sensing, Feeling, Judging) – психотип «энтузиаст», согласно *MBTI*.

ESFP (Extraverted, Sensing, Feeling, Perception) – психотип «исполнитель», согласно *MBTI*.

ESTJ (Extraverted, Sensing, Thinking, Judging) – психотип «директор», согласно *MBTI*.

ESTP (Extraverted, Sensing, Thinking, Perception) – психотип «командир», согласно *MBTI*.

Fasttext – эффективное векторное представление слов для русского языка.

INFJ (Introversion, Intuition, Feeling, Judging) – психотип «адвокат», согласно *MBTI*.

INFP (Introversion, Intuition, Feeling, Perception) – психотип «посредник», согласно *MBTI*.

INTJ (Introversion, Intuition, Thinking, Judging) – психотип «аналитик», согласно *MBTI*.

INTP (Introversion, Intuition, Thinking, Perception) – психотип «мыслитель», согласно *MBTI*.

ISFJ (Introversion, Sensing, Feeling, Judgment) – психотип «защитник», согласно *MBTI*.

ISFP (Introversion, Sensing, Feeling Perception) – психотип «художник», согласно *MBTI*.

ISTJ (Introversion, Sensing, Thinking, Judgment) – психотип «инспектор», согласно *MBTI*.

ISTP (Introversion, Sensing, Thinking, Perception) – психотип «изобретатель», согласно *MBTI*.

KNN (k-Nearest Neighbor) – метод ближайших соседей.

LSTM (Long short-term memory) – сеть долгосрочной и краткосрочной памяти.

MBTI (Myers-Briggs Type Indicator) – метод психологической оценки.

NLTK (Natural Language ToolKit) – пакет библиотек и программ для символьной и статистической обработки естественного языка.

NMF (Non-negative Matrix Factorization) – факторизация неотрицательных матриц.

RF (Random Forest) – метод случайного леса.

RNN (Recurrent neural network) – рекуррентная нейронная сеть.

URL (Uniform Resource Locator) – унифицированный указатель ресурса.

Word2Vec – общее название для совокупности моделей на основе искусственных нейронных сетей, предназначенных для получения векторных представлений слов на естественном языке.

ВВЕДЕНИЕ

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 создание рекомендаций по их практическому использованию.

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, четырех глав, заключения, списка литературы, который включает 133 наименования, и 2 приложений. Общий объем диссертации составляет 140 страниц. Диссертация содержит 9 таблиц и 29 рисунков.

1 ОБЗОР СОВРЕМЕННОГО СОСТОЯНИЯ БОЛЬШИХ ДАННЫХ И МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ В СОЦИАЛЬНЫХ СЕТЯХ

Какие ассоциации вызывает у вас понятие «большие объемы данных»? Для многих это может быть неясным термином, возникающим в виде визуализации массивов громадных серверных ферм, или, возможно, как ассоциация с получением персонализированной рекламы от продавцов.

В современной интерпретации, термин «Большие данные» относится к множеству разнообразных наборов данных, которые выделяются своим значительным объемом и сложностью, что затрудняет их обработку традиционными методами [1]. Наука о данных (*Data Science*) представляет собой область, занимающуюся анализом и извлечением полезной информации из огромных наборов информации [2]. Сравнение между машинным обучением, наукой о данных и областью Больших данных можно провести через аналогию с сырой нефтью и ее переработкой в различных предприятиях. Несмотря на тесные корни в области статистики и традиционных методах управления данными, Большие данные (*Big Data*) и наука о данных сегодня выросли в самостоятельные дисциплины. Понятие «*Big Data*» обычно оценивается с использованием трех ключевых критериев, известных как «правило трех *V*» [3]:

Объем (*Volume*) – количество данных в конкретном объеме.

Разнообразие (*Variety*) – включает в себя различные типы данных, содержащихся в системе;

Скорость (*Velocity*) – скорость генерации и поступления новых данных.

С течением времени предложены и другие дополнительные критерии, такие как достоверность (*Veracity*), жизнеспособность (*Viability*), ценность (*Value*), переменчивость (*Variability*) и визуализация (рисунок 1.1).

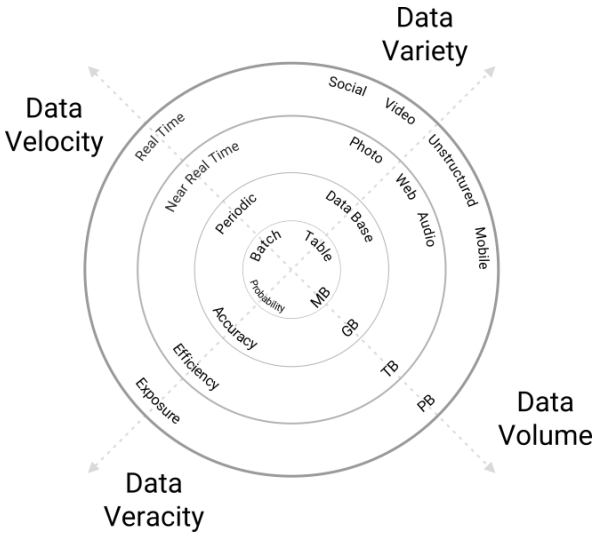


Рисунок 1.1 – Ключевые критерии *Big Data*

На сегодняшний день «*Data Science*» представляет собой научную область, использующую теоретические, математические, вычислительные и практические методы для анализа и оценки данных [4]. Основной задачей этой дисциплины является извлечение ценной информации, применимой в различных областях, таких как принятие решений, разработка продуктов, анализ тенденций и прогнозирование. «*Data Science*» тесно взаимосвязана с областью Больших данных и требует умений в области современных технологий, знаний в области машинного обучения, навыков организации вычислений и разработки алгоритмов. Это выделяет «*Data Scientist*» среди традиционных статистиков.

В наше время информационные технологии все более интенсивно влияют на повседневную жизнь людей вне зависимости от их сферы деятельности. Это приносит с собой не только выгоды, но и риски. Цифровой след, оставляемый при использовании электронных средств связи, оказывает воздействие на разнообразные научные дисциплины, включая социологию, демографию, социально-экономическую географию и историю. Операторы мобильных сетей обладают огромными объемами информации, и анализ данных, полученных от сотовых станций, может точно определить местоположение человека. Эти сведения могут использоваться для отслеживания как отдельных личностей, так

и целых популяций, несмотря на строгое законодательное регулирование такой информации [5].

Социальные сети также представляют собой важный источник данных [6]. Пользователи сами заполняют свои профили, размещая информацию о себе и своих интересах в открытом доступе. Электронные социальные сети, как понятие, возникли в начале 2000-х годов и являются онлайн-платформами для организации и управления социальными связями в сети Интернет.

Со временем, развиваясь, они стали разнообразными по целевой аудитории и целям использования. Несмотря на разнообразие, все они включают профили пользователей, которые заполняются ими добровольно и предоставляют ценную информацию для анализа. Можно утверждать, что анализ публично доступных данных в сети, включая информацию из социальных сетей, остается востребованным и находит практическое применение в разнообразных прикладных задачах. Тем не менее, для эффективного решения таких задач необходимо применение специализированных методов и алгоритмов для сбора и анализа данных.

Итак, наша работа будет сосредотачиваться на создании инструментов для сбора неструктурированной информации из открытых источников с целью построения социального профиля человека в системах принятия решений. Мы также проведем обзор информационных систем, решающих аналогичные задачи, и выявим потенциальные области применения таких систем.

1.1 Большие данные

Понятие «Большие данные» было введено специалистами по управлению данными с некоторой долей иронии, описывая его как «громадное, необузданное количество информации». Однако история анализа данных насчитывает далеко не одно столетие. В 1663 году Джон Граунт, изучая бубонную чуму, описал работу с «огромным количеством информации» и можно утверждать, что он был одним из первых, кто использовал статистический анализ данных [7]. В начале 1800-х годов область статистики стала развиваться, включая сбор и анализ данных.

Эволюция понятия «Большие данные» включает ряд этапов, и несмотря на то, что можно вернуться в 1663 год, чтобы найти первые признаки увеличения объемов данных, стоит отметить, что «Большие данные» – относительное понятие, зависящее от контекста [8]. «Большие данные» для компаний, как *Amazon* или *Google*, значительно отличаются от «Больших данных» для среднего страхового предприятия, но в обоих случаях речь идет о значительных объемах данных.

1.1.1 Основы Больших данных

Проблема обработки данных стала очевидной для Бюро переписи населения США еще в 1880 году. Оценки показывали, что на обработку данных, собранных в ходе переписи 1880 года, потребовалось бы восемь лет, и ожидалось, что перепись 1890 года займет более 10 лет для обработки. В 1881 году молодой сотрудник бюро по имени Герман Холлерит разработал табулирующую машину Холлерита, основанную на перфокартах, предназначенных для управления узорами на ткацких станках. Это изобретение сократило сроки обработки данных с десяти лет до трех месяцев [9].

В 1927 году австрийско-немецкий инженер Фриц Пфлюмер разработал метод магнитной записи информации на магнитной ленте. Этот метод использовал тонкую бумагу, покрытую порошком оксида железа и лаком [10]. С этого момента магнитные ленты стали важной технологией для записи и хранения данных.

В 1943 году британские ученые создали машину «Колосс», предназначенную для сканирования и анализа перехваченных немецких сообщений. «Колосс» был примером процессора обработки данных [11]. В 1945 году Джон фон Нейман положил начало современной компьютерной архитектуре [12]. Его идеи стали основополагающими для разработки будущих компьютеров и определили подходы к программированию, архитектуре и организации вычислительных систем.

Считается, что данные события стали катализатором для создания Агентства национальной безопасности США (*NSA*) в 1952 году. Это агентство было учреждено с целью дешифровки сообщений, перехваченных в ходе «холодной войны», что требовало значительных вычислительных мощностей и

новых технологий обработки информации. Компьютеры того времени достигли такого уровня развития, что они могли не только собирать, но и автоматически обрабатывать большие объемы данных. Это новшество открыло новые горизонты в области анализа информации и положило начало новой эре в обработке «Больших данных». В дальнейшем, способности к автоматизации и быстрому анализу данных стали ключевыми для стратегического планирования и оперативной деятельности в различных областях, включая безопасность и разведку.

1.1.2 Использование Больших данных

Время «Больших данных» свидетельствует о глубокой революции в различных сферах промышленности и влияет на культурные и поведенческие аспекты жизни человека. Эпоха информации меняет способы обучения, музыкального творчества и труда. В данном контексте приведем ряд примеров использования Больших данных [13]:

1. Здравоохранение. Большие данные применяются для создания карт вспышек болезней и тестирования новых методов лечения.

2. *NASA. NASA* использует Большие данные для исследования Вселенной, анализа космических явлений и дистанционного мониторинга планет.

3. Музыкальная индустрия. Вместо реализма Большие данные используют исследования и анализ для выявления предпочтений аудитории и формирования плейлистов и рекомендаций.

4. Утилиты. Компании энергетического сектора используют Большие данные для анализа поведения потребителей и предотвращения отключения электричества.

5. Спортивные товары и фитнес. Компании, такие как *Nike*, применяют сенсоры и устройства для мониторинга состояния здоровья клиентов и предоставления персонализированных рекомендаций.

6. Кибербезопасность. Большие данные применяются в области кибербезопасности для выявления и пресечения киберпреступности.

1.1.3 Анализ Больших данных

В 2017 году опрошено 2800 специалистов в области бизнес-аналитики, которые предсказали, что анализ данных и их визуализация станут ключевыми направлениями. Визуализация данных представляет собой эффективную форму визуальной коммуникации, включая инфографику, и позволяет наглядно отображать информацию, включая изменения и колебания [14].

Модели визуализации данных становятся все более популярными для получения информации из больших объемов данных. Однако существующие модели иногда остаются неуклюжими и требуют дополнительного усовершенствования. Среди компаний, предоставляющих инструменты для визуализации Больших данных, следует отметить такие, как *Domo* [15], *Qlik* [16], *Tableau* [17], *Sisense* [18] и т.д.

История Больших данных далека от своего завершения, и, несомненно, объем данных будет продолжать расти. С развитием этой области будут разработаны новые технологии для улучшения сбора, хранения и анализа данных, тем самым способствуя более быстрой трансформации нашего мира на основе данных.

На сегодняшний день все больше компаний активно внедряют анализ Больших данных в свою деятельность. Одной из таких компаний является «*HCL*», специализирующаяся на понимании и внедрении Больших данных в других организациях. Такие пионеры данных продолжают динамично внедрять и развивать область Больших данных.

1.2 Основные типы открытых данных

В области Больших данных и науки о данных существует множество различных типов информации, доступной в открытом доступе [19]. Рассмотрим эти типы более подробно:

1. Структурированные данные. Это данные, соответствующие определенной модели, которые могут легко храниться в табличных форматах баз данных или файлах *Excel*. *SQL* (*Structured Query Language*) часто используется для управления и запроса таких данных. Однако могут существовать структурированные данные,

которые трудно поместить в традиционные реляционные базы данных, например, иерархические данные, такие как семейные древа.

2. Неструктурированные данные. Эти данные не следуют определенной модели и могут иметь разнообразный и изменчивый контекст. Примером является электронная почта, которая содержит структурированные элементы, такие как отправитель и текст, но может быть трудной для анализа из-за разнообразия способов выражения информации.

3. Данные на естественном языке представляют собой сложный подтип неструктурированных данных, требующий глубоких знаний в области лингвистики и аналитических методов. Обработка текстов на естественном языке включает в себя такие задачи, как распознавание сущностей, анализ эмоций и множество других аспектов.

4. Машинные данные. Эти данные генерируются автоматически компьютерами, приложениями и устройствами без участия человека. Примерами являются данные сети Интернет вещей (*IoT*), журналы веб-приложений и т.п.

5. Графовые данные. Такие данные фокусируются на связях и отношениях между объектами. Эти данные используют структуры графов для представления и хранения информации. Примерами являются социальные сети и связи между объектами [20].

6. Мультимедийные данные. Эти данные включают в себя видео, аудио и графику. Обработка таких данных является сложной задачей, так как компьютеры должны анализировать мультимедийный контент.

7. Потоковые данные. Эти данные поступают в систему непрерывно при возникновении событий. Примерами являются прямые трансляции и данные о финансовых котировках.

Знание различных типов данных и умение адаптировать методы обработки к их особенностям является важной частью работы исследователей в области Больших данных и науки о данных [21].

1.3 Анализ систем для эффективной обработки объемных данных

Две ключевые задачи, которые стремятся решать современные ИС, взаимодействуя с большими объемами данных, размещаемых на электронных социальных платформах [22, 23]:

1. Мониторинг и анализ. Этап мониторинга включает в себя сбор и организацию первичных данных, построение связей между пользователями и внешними ресурсами. Этап анализа направлен на выявление структурных и статистических закономерностей в данных, а также расчет основных количественных характеристик.

2. Прогнозирование и управление. На этапе прогнозирования используется заранее определенная математическая модель информационного процесса, которая может быть представлена в виде статистической модели или модели динамического процесса. Этап управления предполагает реализацию комплекса мероприятий в социальной сети с целью достижения желаемого состояния протекающих в ней процессов.

ИС можно классифицировать по следующим критериям [24]:

1. По уровню анализа социальных сетей. ИС отслеживают активности в социальных сетях для выявления трендов и настроений, предсказывают данные, воздействуют на пользователей и контент для достижения определенных целей.

2. По моделям социальных сетей. ИС исследуют взаимосвязи между участниками сети.

3. По методам анализа данных. ИС могут анализируют данные с использованием статистических метрик и тестов, фокусируются на визуализации и анализе сетевых структур.

4. По объектам анализа социальных сетей. ИС фокусируются на изучении поведения и взаимодействия как отдельных участников, так и группы людей.

5. По режимам анализа данных. ИС позволяют отслеживать и реагировать на события по мере их возникновения.

6. По режимам сбора данных. Сбор данных может быть применен ко всему объему информации или ограничиваться определенной темой.

7. По источникам данных. ИС могут использовать один или несколько источников данных, такие как социальные сети, блоги, форумы и файловые сервисы.

8. По объемам обрабатываемых данных. Информационные системы могут работать как с большими объемами данных, так и с модельными объемами данных.

Далее будет представлен обзор таких ИС.

1.3.1 Массовые системы анализа социальных сетей

Предположим, нас заинтересовал определенный объект, персональность или событие, и мы стремимся получить информацию на эту тему. В этом случае можно воспользоваться системой массового анализа. Известно несколько таких систем:

1. Поисковые ИС. Их ключевой характеристикой является простота использования и основной функционал – поиск информации (Поиск в блогах – *blogs.yandex.ru*).

2. ИС с уведомлениями. Эти системы, подобно поисковым машинам, предоставляют возможность настроить временной интервал для поиска информации с последующей отправкой результатов пользователю (*Google* оповещения – *google.ru/alerts*).

3. Агрегирующие ИС. Эти системы отслеживают статистику и популярность запросов, поступающих в поисковые системы (например, *wordstat.yandex.ru* – показывает поисковые запросы).

4. ИС для сбора данных. Эти системы способны собирать информацию из различных источников, включая приложения для просмотра *RSS*-лент.

5. ИС для сбора и агрегации данных. Эти системы специализируются на сборе и объединении информации из разных источников, которые могут различаться по формату данных (*FeedsApi, Quadrigram*).

Рассмотренные ИС обладают рядом преимуществ, включая доступность (бесплатное использование) и низкий барьер вхождения. Однако ограниченный

функционал, связанный с анализом открытых данных, ограничивает их использование на профессиональном уровне. В первую очередь, эти ИС разработаны с целью предоставления обзорной информации.

1.3.2 Системы мониторинга и анализа социальных сетей в контексте коммерческих организаций

Современные системы мониторинга и анализа социальных сетей помогают компаниям эффективно решать как внутренние, так и внешние задачи, связанные с их развитием. [25]. Эти задачи включают в себя анализ работы сотрудников, оптимизацию бизнес-процессов, создание позитивной рабочей атмосферы, исследование рынка, привлечение новых партнеров, поддержание существующих связей, оценку результатов компании и продвижение бренда.

Сегодня мониторинг и анализ социальных сетей должны обладать рядом ключевых функций для поддержания конкурентоспособности:

1. Отслеживание упоминаний брендов, что позволяет контролировать, как часто и в каком контексте упоминаются товарные марки в социальных сетях.
2. Анализ рыночных рисков и возможностей помогает выявлять значимые темы, которые могут указывать на новые перспективы развития бизнеса, или, наоборот, на угрозы.
3. Веб-аналитика предоставляет инструменты для отслеживания и анализа поведения пользователей на сайтах.
4. Обратная связь в социальных сетях создает возможность взаимодействия с клиентами в реальном времени через социальные платформы.
5. Прогнозирование и управление в социальных сетях включает функции для формирования прогностических моделей, позволяющих предсказать поведение клиентов [26].

В качестве примера ниже рассмотрим несколько ИС.

ИС «*Social Studio*» – платформа позволяет в режиме реального времени отслеживать упоминания торговых марок и анализировать их эмоциональную окраску с использованием морфологических анализаторов. Одним из ключевых

преимуществ является возможность быстрой реакции на упоминания. Пользователи могут настроить параметры мониторинга и получить доступ ко всем учетным записям через единый интерфейс [27].

Кроме того, система предлагает настройку рейтинговых профилей по различным критериям, таким как количество сообщений по определенной теме, уникальные комментаторы и количество входящих ссылок. Однако стоит отметить, что период ретроспективного анализа данных ограничен одним месяцем, что может оказаться недостаточным для некоторых задач. Основные характеристики системы *Social Studio* представлены в таблице 1.1.

Таблица 1.1 – Основные характеристики системы *Social Studio*.

Вендор	SalesForce
Сайт	www.marketingcloud.com
Пользователи	Коммерческие организации
Уровень анализа данных	Мониторинг и анализ
Методы анализа	Базовые методы анализа и поиска текстов на уровне ключевых слов, анализ тональности текстов (в том числе и на русском языке), визуальный анализ (инфографика)
Объекты анализа	Сеть в целом, пользователи, информационные сообщения, мнения
Режим анализа	Анализ в режиме реального времени, ретроспективный анализ с ограничением в 30 дней
Сбор данных	Сбор данных в режиме реального времени
Охват источников данных	Блоги, форумы, новостные медиа, сайты обмена изображениями и видео, социальные сети

ИС «*IQBuzz*» представляет собой мощный инструмент для мониторинга социальных сетей, который работает в круглосуточном режиме и обеспечивает пользователей актуальными данными в реальном времени [28]. Сразу несколько пользователей могут одновременно взаимодействовать с

платформой, делиться результатами анализа и обмениваться мнениями. Открытый доступ к анализируемым данным позволяет всем заинтересованным пользователям получать информацию и делать собственные выводы на основе собранных данных. Система автоматически определяет тональность сообщений, что помогает пользователям понять, как общество воспринимает определенные темы или события. На основе информации, размещенной на их страницах в социальных сетях, «*IQBuzz*» может генерировать данные о возрасте, половой принадлежности и других характеристиках пользователей. Это позволяет более точно интерпретировать полученные результаты и понимать целевую аудиторию. Основные характеристики системы «*IQBuzz*» представлены в таблице 1.2.

Таблица 1.2 – Основные характеристики системы *IQBuzz*

Вендор	Айкубаз
Сайт	iqbuzz.pro
Пользователи	Коммерческие организации и частные лица
Уровень анализа данных	Анализ, мониторинг и управление репутацией
Методы анализа	Поиск и анализ фраз (в том числе и английских), анализ тональности текста и визуальное представление полученных данных
Объекты анализа	Онлайн СМИ, социальные сети, видеохостинги и др.
Режим анализа	Анализ в режиме реального времени
Объемы обрабатываемых данных	Более 30 миллионов упоминаний в сутки
Сбор данных	Сбор данных в режиме реального времени
Охват источников данных	онлайн–СМИ, ВКонтакте, Мой Мир, Google+, YouTube, RuTube и др.

ИС «*Brand Analytics*» – многофункциональный инструмент для анализа данных, предназначенный для работы с информацией из социальных сетей [29].

Одной из основных функций «*Brand Analytics*» является сбор данных, который осуществляется на основе ключевых слов, геолокаций и авторов сообщений в социальных сетях. Это позволяет получать полную картину о том, как определенные темы или бренды обсуждаются в различных цифровых пространствах. Собранные метаданные предоставляют возможность глубокого анализа информации, включая оценку тональности сообщений, выявление дубликатов и анализ языковых особенностей. Это дает возможность определить, как аудитория реагирует на различные маркетинговые кампании или события. Система также учитывает пол и местоположение авторов сообщений, что позволяет более точно сегментировать целевую аудиторию и адаптировать маркетинговые стратегии. Одним из значительных преимуществ системы является то, что «*Brand Analytics*» разработана в России, что обеспечивает поддержку русского языка. Основные характеристики системы «*Brand Analytics*» представлены в таблице 1.3, где можно ознакомиться с ее функциональными возможностями, технологиями, используемыми для анализа, а также с другими параметрами, важными для пользователей.

Таблица 1.3 – Основные характеристики системы *Brand Analytics*

Вендор	Brand Analytics
Сайт	br–analytics.ru
Пользователи	Коммерческие организации и частные лица
Уровень анализа данных	Мониторинг и автоматический анализ
Методы анализа	Полная поддержка поиска по словам, тегам и тд.
Объекты анализа	Социальные сети, блоги, онлайн СМИ, мессенджеры, сайты госучреждений, организаций, сайты отзывов, видеохостинги
Режим анализа	Анализ в режиме реального времени
Сбор данных	Сбор данных в режиме реального времени
Охват источников данных	Соцсети (vk.com, odnoklassniki.ru и др), блоги (liveinternet.ru, diary.ru), СМИ (ria.ru, lenta.ru) и др.

ИС, представленные в настоящем обзоре, идеально подходят для высокоэффективной работы в сфере социальных сетей, что делает их уникальными по сравнению с общими системами обслуживания. Эти платформы предоставляют возможность тщательного мониторинга упоминаний торговых марок, анализа эмоциональной окраски сообщений и взаимодействия с пользователями через интеграцию учетных записей компаний в сети.

Однако, для эффективного функционирования этих платформ требуется надежное оборудование, способное справляться с большими объемами данных и обеспечивать стабильную работу системы. Это может стать значительным препятствием для небольших компаний или организаций с ограниченным бюджетом.

1.3.3 Системы и фреймворки для анализа и обработки данных

Фреймворки и движки, используемые в процессе обработки данных, представляют собой ключевые элементы в архитектуре систем обработки данных. Хотя нет строгого и авторитетного определения, разграничивающего «движки» от «фреймворков», иногда полезно рассматривать первые как компоненты, фактически выполняющие вычисления над данными, в то время как последние представляют собой набор компонентов, предназначенных для поддержки этого процесса.

Например, можно взглянуть на *Apache Hadoop* [30] как на структуру обработки данных, где *MapReduce* выступает в роли механизма обработки по умолчанию. Часто движки и фреймворки могут использоваться взаимозаменяемо или совместно. Например, *Apache Spark*, альтернативная структура, может быть интегрирована с *Hadoop* для замены *MapReduce* [31]. Эта функциональная совместимость между компонентами является одной из ключевых причин, почему большие системы обработки данных обладают высокой гибкостью (рисунок 1.2).

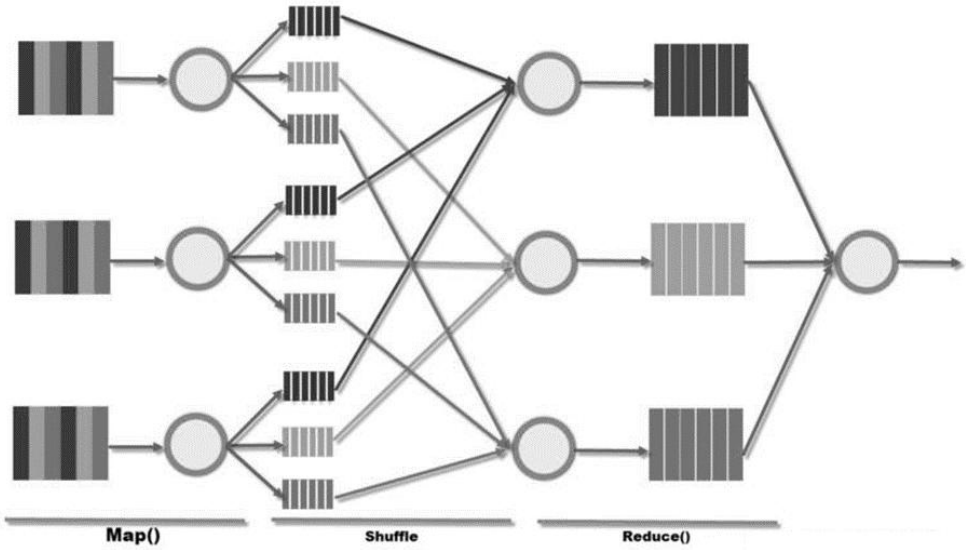


Рисунок 1.2 – Модель обработки данных *MapReduce*

Для упрощения обсуждения указанных компонентов, мы предлагаем сгруппировать структуры обработки данных в соответствии с типами данных, которые они обрабатывают. Некоторые системы выполняют обработку данных пакетами, в то время как другие работают с непрерывным потоком данных по мере их поступления в систему. Третьи могут применять как один, так и другой метод обработки.

Далее рассмотрим две ключевые концепции обработки данных, прежде чем погрузиться в детали и последствия различных реализаций.

Системы пакетной обработки данных. Пакетная обработка данных имеет долгую историю в мире анализа данных. Она включает в себя работу с большим статическим набором данных и предоставление результата после завершения вычислений.

Характеристики пакетной обработки данных обычно включают в себя:

- Ограниченность. Конечные объемы информации, что позволяет четко определить начало и конец обработки.
- Постоянство. Данные обычно хранятся в постоянном хранилище.
- Обширность. Пакетные операции часто используются для обработки очень Больших объемов данных.

Пакетная обработка хорошо подходит для случаев, когда требуется доступ ко всем записям в наборе данных. Например, при вычислении общих и средних значений, данные должны обрабатываться в целом, а не по отдельным записям. Для этого требуется поддерживать состояние данных в течение вычислений.

Примером системы, способной обрабатывать данные в пакетном режиме, является *Apache Hadoop* [32]. Эта система представляет собой платформу обработки данных, спроектированную для пакетной обработки. *Hadoop* была одной из первых крупных сред с открытым исходным кодом, которая получила значительную популярность в анализе Больших данных. Современные версии Hadoop включают в себя несколько ключевых компонентов и слоев, которые работают совместно для эффективной обработки данных пакетами [33]:

- *HDFS* (*Hadoop Distributed File System*) – распределенная файловая система, обеспечивающая управление хранением данных и их репликацию между узлами кластера. *HDFS* гарантирует доступность данных даже в случае сбоев узлов и служит как источником данных, так и для хранения промежуточных и финальных результатов вычислений.
- *YARN* (*Yet Another Resource Negotiator*) – компонент, ответственный за управление ресурсами кластера и планирование задач. Он позволяет запускать разнообразные задачи в кластере Hadoop и обеспечивает координацию ресурсов.
- *MapReduce* – пакетный процессор обработки данных *Hadoop*.

Модель пакетной обработки данных в *Hadoop* основана на механизме *MapReduce*, который включает в себя этапы чтения данных из *HDFS*, разделения данных, применения вычислений, перераспределения результатов и окончательной агрегации результатов. Поскольку *MapReduce* часто включает многократное чтение и запись данных, он может работать медленно. Однако, благодаря использованию дискового пространства в качестве хранилища данных, *MapReduce* способен обрабатывать огромные объемы информации. Это также делает его экономически более доступным, так как не требуется больших

объемов оперативной памяти. *Hadoop* имеет обширную экосистему и может использоваться как базовая платформа для множества приложений и интеграций с другими системами обработки данных.

В заключение, стоит подчеркнуть, что *Apache Hadoop* и его процессор данных *MapReduce* предлагают надежную и проверенную модель пакетной обработки данных, которая прекрасно подходит для работы с огромными объемами информации. Эта архитектура эффективна в ситуациях, где время выполнения не является критическим фактором, позволяя обрабатывать большие массивы данных с высокой эффективностью и надежностью. Благодаря своей масштабируемости и гибкости, *Hadoop* остается одним из ведущих решений для аналитики и обработки больших данных в различных отраслях. Экономическая доступность и возможность интеграции с другими платформами делают *Hadoop* привлекательным решением для множества сценариев обработки данных с использованием различных технологий.

1.4 Практическое применение Больших данных

Применение *Big Data* охватывает множество областей и сфер деятельности. Вот некоторые ключевые пример:

1. Анализ потребностей клиентов. Сегодня одним из наиболее значимых применений *Big Data* является лучшее понимание клиентов и их поведения. Данные о клиентах их предпочтениях, полученные из различных источников, включая социальные сети, браузерные журналы и аналитику текста, позволяют компаниям создавать прогностические модели. Например, *Wal-Mart* использует *Big Data* для прогнозирования продаж, а компании по страхованию автомобилей оценивают вождение клиентов для точного расчета страховых тарифов. *Netflix* применяет *Big Data* для анализа просмотров и предпочтений своих пользователей, что позволяет рекомендовать контент на основе их интересов. Также компания использует данные для определения, какие шоу и фильмы следует производить, основываясь на трендах и запросах

аудитории. Это помогает не только улучшить пользовательский опыт, но и повышает успех оригинальных проектов.

2. Усовершенствование бизнеса. Компания Target использует анализ больших данных для прогнозирования потребительского спроса на основе покупок, поведения в интернете и демографической информации. Это позволяет им оптимально управлять запасами и персонализировать предложения, что увеличивает вероятность покупки. В результате Target может заранее предлагать клиентам товары, которые они с наибольшей вероятностью захотят приобрести в определенные периоды. Даже процессы управления человеческими ресурсами улучшаются с помощью обработки Больших данных, включая оптимизацию процессов найма персонала и анализ корпоративной культуры.

3. Личное качественное оценивание и повышение производительности. Большие объемы данных имеют особое значение для личного использования. Данные с носимых устройств, таких как смарт-часы и браслеты, позволяют отслеживать здоровье и активность. Эти устройства собирают обширные объемы информации, которые анализируются для выявления новых закономерностей и улучшения общего состояния здоровья. Кроме того, Большие данные играют важную роль в онлайн-знакомствах, помогая людям находить совместимых партнеров.

4. Улучшение здравоохранения и общественного здоровья. *Big Data* способствуют развитию медицины и общественного здоровья. Их вычислительная мощность позволяет декодировать ДНК и проводить исследования заболеваний. Данные с носимых устройств могут использоваться для сбора информации о здоровье и активности миллионов людей. Такие данные позволяют оценивать эффективность лечения и внедрять инновации в здравоохранении. Примером такого применения является *ResearchKit* от *Apple*, который позволяет проводить биомедицинские исследования с использованием смартфонов пользователей. Таким образом, *Big Data* становятся все более важными для разных сфер деятельности, предоставляя новые возможности для

оптимизации и улучшения бизнес-процессов, здравоохранения и общественного здоровья, а также для личной оценки и повышения производительности.

5. Улучшение спортивных результатов. В современном спорте Большие данные стали неотъемлемой частью аналитики и стратегии. Команда NBA Golden State Warriors активно использует аналитику больших данных для улучшения игровых результатов. Они анализируют данные о каждом игроке, включая их физические показатели, эффективность бросков и тактические действия на поле. С помощью таких систем, как Second Spectrum, команда может получать детализированные статистические отчеты и видеоматериалы, что позволяет тренерам и игрокам лучше понимать стратегии соперников и адаптировать свою игру. Это помогает в подготовке к матчам и повышает общую производительность команды. Элитные спортивные команды в настоящее время расширяют применение умных технологий за пределы тренировочных полей и соревновательных арен. Они активно используют современные умные устройства для мониторинга здоровья и физической активности своих спортсменов. Такие устройства предоставляют данные о питании, качестве сна и эмоциональном состоянии, что позволяет создать полную картину о состоянии атлета и оптимизировать его тренировочные и соревновательные процессы.

6. Совершенствование науки и исследований. Например, проект Human Genome Project, который завершился в 2003 году, стал важным примером использования больших данных в науке. Этот международный проект по секвенированию человеческого генома генерировал колоссальные объемы генетической информации, что потребовало применения передовых вычислительных технологий для хранения и анализа данных. Благодаря анализу таких массивов данных исследователи смогли выявить генетические предрасположенности к заболеваниям, что в свою очередь открыло новые горизонты для медицины и генетики. Такие подходы к анализу данных сегодня применяются в различных научных областях.

7. Оптимизация производительности машин и устройств. Большие данные также играют важную роль в развитии автономных систем и устройств. Например, компании, занимающиеся производством беспилотных летательных аппаратов (БПЛА), используют большие данные для оптимизации маршрутов и повышения эффективности полетов. Беспилотники собирают информацию о погодных условиях, воздушных потоках и препятствиях в реальном времени, что позволяет им адаптироваться к изменяющимся условиям. Это позволяет обеспечить безопасное и эффективное выполнение задач, таких как доставка грузов или мониторинг сельскохозяйственных угодий. Такие технологии помогают улучшить производительность и снижать риски при выполнении различных операций. Кроме того, интеллектуальные счетчики энергии в домах клиентов позволяют отслеживать и оптимизировать потребление энергии в режиме реального времени, а умные сети и интеллектуальные счетчики поддерживают эффективное управление энергопотреблением. Розничные сети также используют Большие данные для оптимизации работы магазинов и управления персоналом. Таким образом, Большие данные стали незаменимым инструментом в различных сферах, включая спорт, науку, технологии и медицину, и предоставляют возможности для оптимизации и улучшения работы систем и устройств.

8. Совершенствование безопасности и правопорядка. Большие данные нашли широкое применение в области повышения уровня безопасности и обеспечения соблюдения правопорядка. В России примером использования больших данных для повышения безопасности является проект "Безопасный город". Эта инициатива направлена на интеграцию технологий видеонаблюдения, датчиков и аналитических систем для мониторинга общественных мест. В рамках проекта устанавливаются камеры с функцией распознавания лиц и анализом поведения, что позволяет оперативно реагировать на подозрительные действия и предотвращать правонарушения. Данные, собранные из различных источников, обрабатываются в реальном времени, что помогает правоохранительным органам более эффективно планировать патрулирование и реагировать на инциденты.

9. Улучшение и оптимизация городской и государственной инфраструктуры. Роль Больших данных в значительной степени определяется улучшением различных аспектов функционирования городов и стран. Использование данных в реальном времени открывает возможности для оптимизации движения в городах в зависимости от текущей ситуации на дорогах, а также учета социальных и метеорологических данных. Некоторые города стремятся стать «умными городами», в которых все аспекты инфраструктуры и коммунальных служб интегрированы. Это означает, что автобусы будут ожидать задержанные поезда, а светофоры будут реагировать на текущие трафиковые условия для минимизации пробок.

10. Финансовая торговля. Завершая этот обзор, стоит обратить внимание на применение Больших данных в финансовой сфере. В России примером применения больших данных в финансовой сфере является система алгоритмической торговли, используемая банком «Сбер». Банк внедрил технологии, которые анализируют большие объемы данных, включая рыночные новости, финансовые отчеты и поведение трейдеров, для автоматизации процессов покупки и продажи активов. Это позволяет «Сберу» принимать решения в реальном времени, оптимизировать портфель и улучшать результаты торговли. Также банк использует аналитические инструменты для оценки рисков и прогнозирования изменения курсов валют, что способствует более эффективному управлению активами и снижению финансовых потерь. Ведущие российские банки, в числе которых и Альфа-Банк, и Тинькофф, также долгое время успешно внедряют технологии Больших данных в своей деятельности, что подчеркивает важность данных в финансовом секторе.

1.5 Анализ современного состояния моделей и методов интеллектуального анализа данных пользователей социальных сетей

В последние годы наблюдается стремительный рост числа социальных сетей, а также числа людей, пользующихся этими сетями. Социальные сети, также называемые социальным программным обеспечением или программным

обеспечением для совместной работы, представляют собой ряд приложений, которые расширяют групповые взаимодействия и общие пространства для совместной работы, социальных связей и совокупного обмена информацией в веб-среде [34]. Аналогичным образом, социальные сети определяются как веб-сервисы, позволяющие отдельным лицам [35]:

1) создавать общедоступный или полупубличный профиль в ограниченной системе;

2) формулировать список других пользователей, с которыми у них есть общее подключение;

3) просматривать и пересматривать свой список подключений и те, которые были установлены другими пользователями в системе.

Социальные сети привлекли внимание миллионов пользователей с самого момента появления этих сайтов в публичном доступе (таких как «ВКонтакте», «Одноклассники», «Телеграмм» и т.д.). Большинство пользователей интегрировали такие сайты в свою повседневную жизнь. В России самой популярной социальной сетью является «ВКонтакте», на втором – «Одноклассники» [36]. «ВКонтакте» и «Одноклассники» определяются как «социальная утилита, которая помогает людям обмениваться информацией и более эффективно общаться со своими друзьями, семьей и коллегами». Например, по состоянию на май 2023 года «ВКонтакте» насчитывает 79,5 миллионов зарегистрированных пользователей. Год к году показатель увеличился на 9,6%. По миру количество пользователей «ВКонтакте» в месяц составило 101,7 млн.

В настоящее время все чаще при изучении социальных сетей, применяется интеллектуальный анализ данных [37]. Это процесс, в котором используются различные инструменты анализа данных для выявления закономерностей и взаимосвязей в данных, которые могут быть использованы в целях прогнозирования [38]. Контролируемые данные – методы интеллектуального анализа используются для моделирования выходной

переменной на основе одной или нескольких входных переменных, и эти модели могут быть использованы для прогнозирования будущих случаев.

Социальные сети становятся объектом активного исследования, поскольку поведение пользователей может служить индикатором их личностных качеств, а также для выявления настроений и критических психических состояний, таких как депрессия или суицидальные наклонности. Сложность таких исследований возникает из-за необходимости сбора и анализа больших объемов данных, извлекаемых из социальных сетей, что требует значительных вычислительных ресурсов и продвинутых методов обработки информации. Эти данные могут включать текстовые сообщения, комментарии и другую активность пользователей, что создает многомерную картину их поведения. Кроме того, существует множество нелинейных взаимосвязей между поведенческими характеристиками человека в реальной жизни и его виртуальной активностью. Например, настроение и эмоциональное состояние пользователя могут отражаться в его сообщениях и реакциях в сети, но эти проявления могут быть искажены внешними факторами, такими как влияние друзей, культурные контексты и личные обстоятельства. Это делает интерпретацию данных сложной задачей, требующей комплексного подхода и многофакторного анализа, чтобы точно уловить связи между реальным и виртуальным поведением. К тому же, большинство исследований сосредоточено на анализе данных из одной социальной сети, что ограничивает потенциал предсказательной модели, т.к. учитывается лишь небольшое подмножество всех доступных цифровых следов, а это в свою очередь приводит к низкой эффективности традиционных методов обработки и анализа данных.

Одним из решений является возможность использования методов машинного обучения.

1.5.1 Обзор существующих решений анализа социальных сетей

Современные социальные сети стали настоящим кладезем информации о пользователях. Они предоставляют не только личные данные, но и разнообразные аспекты взаимодействия, интересов и принадлежности к различным сообществам.

Это богатство информации создает уникальные возможности для анализа и структурирования данных, позволяя извлекать ценные знания.

Анализ социальных сетей способен значительно углубить наше понимание поведения пользователей и их предпочтений, а также предсказывать будущие тренды. Информация о взаимодействиях пользователей – лайки, комментарии, репосты – позволяет компаниям не только понять, какие товары вызывают наибольший интерес, но и выяснить, какие факторы влияют на принятие решений.

В итоге, социальные сети становятся не просто платформами для общения, но и мощными инструментами для бизнеса, помогающими лучше понимать свою аудиторию и адаптировать предложения в соответствии с ее потребностями. Это превращает каждый пост и комментарий в потенциальный источник стратегической информации, открывая новые горизонты для успешного взаимодействия с клиентами.

В этом разделе проводится обзор литературы, связанной с социальными сетями и определения личностных характеристик пользователя на основе размещаемой им информации.

Многочисленные исследовательские группы стремились решить задачу автоматического психологического профилирования пользователей [39, 40]. Результаты всех проведенных исследований сводятся к тому, что психотип существенно влияет на поведение людей в социальных сетях. В статье [41] отмечается, что психотип является важным фактором в человеческом поведении, взаимодействии и эмоциональном восприятии. Также необходимо отметить научные изыскания группы исследователей [42], которая проанализировала сообщения 75 тысяч волонтеров на *«Facebook»* и выявила связь между использованием определенных слов и психотипом.

Данной проблемой занимались сотрудники ФГАОУ ВО «Пермский государственный национальный исследовательский университет» (ПГНИУ). Они разработали и запатентовали компьютерную программу, предназначенную для анализа текстов комментариев в социальных сетях. Эта программа создает психологический и лингвистический портрет пользователей, используя сложные

алгоритмы обработки естественного языка. В ходе разработки платформы исследователи собрали обширную базу данных, включающую 21 тысячу текстов, написанных пользователями в различных социальных сетях. Эти тексты подверглись тщательному анализу по множеству параметров: от стилистических особенностей и жанровых характеристик до уровня информативности. В общей сложности программа учитывает более 150 различных категорий, что позволяет глубже понять особенности коммуникации пользователей [43].

Кроме того, для более точной оценки психологических характеристик пользователей в программу были интегрированы такие параметры, как возраст, пол и количество опубликованных постов. Исследователи использовали опросник «*Big Five Inventory*» для определения ключевых психологических черт пользователей, таких как открытость, добросовестность, экстраверсия, приятность и невротизм. Это позволяет программе достигать точности в определении психологических параметров собеседника до 70%. При исследовании выявлены следующие закономерности: в текстах мужчин практически отсутствуют извинения, а женщины используют их значительно чаще. Мужчины с недобросовестными тенденциями часто прибегают к сарказму и иронии в своих сообщениях. Ученые отмечают, что люди, склонные к частым шуткам, могут проявлять враждебность и признаки интровертности. В то же время пользователи, чьи тексты наполнены сочувствием, обычно являются доброжелательными и консервативными [44]. Одним из минусов данного исследования является ограниченность его применения для анализа аккаунта пользователя социальной сети в реальном времени. К тому же, в программе, разработанной сотрудниками ПГНИУ, отсутствует интерфейс и все команды выполняются через командную строку.

В проекте «Разработка алгоритма идентификации факторов риска безопасности пользователей социальных сетей на основе анализа контента и психологических характеристик его потребителей» (РНФ №19-78-10122) ученые создали комплексную модель идентификации и прогнозирования факторов риска безопасности в социальных сетях, основанной на анализе контента (содержания)

социальных сетей и индивидуально-психологических характеристик пользователей (тревожность, агрессивность, типология характера) [45]. Ученые измерили с помощью диагностических инструментов психологическое состояние человека, после чего сравнили его цифровой след с противоположными результатами, например, высокий уровень тревожности с низким. Также при помощи наблюдения за содержанием и пользователями тех сообществ в соцсети, которые могут быть потенциально опасными, подтвердили результаты измерения [46].

В статье [47] представлена модель для анализа текстов и прогнозирования личности брендов в социальных сетях. Для этого использовалась модель «Большой пятерки». Исследование включало использование наборов данных «*myPersonality*» и данных, относящихся к страницам брендов. Функции для анализа были извлечены из обоих наборов данных. При выборе функций использовались различные методы, такие как корреляция Пирсона и методы машинного обучения (*SVR, XGB*, нейронные сети с обратной связью). Модель *XGB* дала наилучший результат в прогнозировании личности брендов.

Авторы статьи [48] представили систему, способную анализировать личностные черты пользователей *«Facebook»* на основе публикуемых ими статусов. В работе была использована модель «Большой пятерки». Для анализа использовался набор данных «*MyPersonality*», содержащий информацию о 250 пользователях и около 10000 примеров статусов от пользователей. После извлечения сообщений производилась предварительная обработка путем удаления ссылок, символов и т. д. Все слова были преобразованы в нижний регистр. Для данных в реальном времени использовался алгоритм исправления орфографии. Сообщения также содержали символы, такие как хэштеги (#) и смайлы, которые были удалены, оставив только слова. Для извлечения ключевых слов из документов был рассчитан *TF-IDF*, формируя таким образом вектор признаков. Этот вектор был слишком большим, поэтому для уменьшения его размера и получения только значимых признаков был использован метод главных компонент. В работе были использованы алгоритмы машинного обучения *KNN* и *SVM*. *KNN* оказался наилучшим для классификации личностных черт.

В статье [49] авторы рассматривали сверточные нейронные сети в качестве классификатора, позволяющего выявить отклонения в психологическом состоянии пользователей интернета. Для этой цели проводилось обучение нейронных сетей с использованием данных, собранных из социальных микроблог-сервисов. Было выделено два типа признаков, описывающих анализируемый поток данных нейронными сетями: данные в сообщении ограниченной длины (твиты), содержащие текст или изображение, и статистические данные, которые включали количество сообщений от пользователя и количество комментариев к сообщению (эти данные собирались в течение определенного периода времени). В экспериментах, проведенных авторами, лучшая точность (более 78,5%) была показана четырехслойной нейронной сетью с методом субдискретизации по среднему значению со временем.

В препринте [50] предложен использующий рекуррентные нейронные сети для предсказания психотипа пользователя метод, анализирующий символы, слова и степень эмоциональной нагрузки его сообщений.

Подход к оценке личности, представленный в [51], основан на векторных семантических моделях (*VSM*), которые предполагают, что значение слова может быть распознано путем анализа слов, которые в данный момент встречаются с целевым словом в данном контексте. Текст пользователя был проанализирован путем представления его в векторной форме и измерения его сходства с заранее определенными векторами личностных качеств. Предложенный подход показал более высокую точность, чем методы машинного обучения (*ML*). Авторы предполагают, что объединение их подхода с алгоритмами *ML* может повысить эффективность оценки.

В статье [52] представлено автоматическое распознавание личностных качеств «Большой пятерки» в социальной сети (*«Facebook»*) с использованием текста статуса пользователей. Для автоматического распознавания использовались различные методы классификации, такие как *SMO* (последовательная минимальная оптимизация для машины опорных векторов), байесовская

логистическая регрессия (*BLR*) и разреженное полиномиальное наивное байесовское моделирование (*MNB*). Точность результатов составила около 63%.

Гозде и др. [53], собрали данные от 99 217 участников из 41 страны в рамках всемирного исследования *COVIDiSTRESS*, чтобы лучше понять многочисленные взаимосвязи между психологическими последствиями *COVID*-19 и личностными чертами «большой пятеркой». Для анализа данных были использованы мультигрупповой факторный анализ и многоуровневая регрессионная модель. Результаты показали, что на протяжении всей пандемии большая пятерка личностных характеристик была тесно связана с чувством стресса и одиночества, в то время как некоторые из этих связей были слабыми. Их исследования помогают выявлять восприимчивых людей и оптимизировать психологическую терапию во время и после пандемии *COVID*-19, когда невротизм играет важную роль в возникновении стресса и уязвимости от одиночества, особенно во время стихийных бедствий.

В работе [54] авторы предложил новый метод машинного обучения для прогнозирования личности людей на основе цифрового следа социальных сетей. Во время эпидемии *COVID*-19 предложенная модель была проверена для каждого соискателя работы путем онлайн-регистрации в каждой организации. Предложенный алгоритм использует динамическую многоконтекстную информацию, такую как информация об аккаунтах на *«Facebook», Twitter* и *YouTube*, а также на других сайтах. Прогнозирование личности оказалось более точным, чем другие доступные подходы. Несмотря на то, что мышление человека меняется в зависимости от сезона, предложенный алгоритм работает регулярно. превосходит другие современные традиционные методы прогнозирования когнитивных способностей человека.

Ян Ли и соавторы в работе [55] предложили новую систему многозадачного обучения, которая одновременно предсказывает личностные черты и эмоциональное поведение, основываясь на хорошо известной корреляции между личностными чертами и эмоциональным отношением. Он

также эмпирически оценил и описал различные механизмы обмена информацией между двумя задачами.

В работе Фатеме и др. [56] предложен метод распознавания личностей по тексту, основанный на применении глубокого обучения. В частности, использованы сверточные нейронные сети (*CNN*), которые зарекомендовали себя как эффективный инструмент для обработки естественного языка и распознавания личности. Для валидации предложенной методики была проведена серия тестов с использованием набора данных эссе. Эмпирические результаты демонстрируют превосходство данной методики по сравнению с существующими методами машинного и глубокого обучения в задачах определения личности.

В статье [57] предложена модель для прогнозирования личностей пользователей *Twitter* с лучшими показателями, чем у других систем прогнозирования. Это было сделано с помощью онлайн-опроса с использованием *Big-five*/ Опрос по инвентаризации (*BFI*), который был разослан 295 пользователям *Twitter* и собрал 511 617 твитов. Предложенная модель использовала метод опорных векторов (*SVM*) для тестирования двух альтернативных семантических подходов, которые объединяли *SVM* и *BERT*. Результаты показывают, что сочетание этих двух подходов дает показатель точности в 79,35 процента, а внедрение *LIWC* повышает этот показатель до 80,07 процента.

В заключение данного раздела следует отметить, что понимание личности играет критически важную роль в процессе подбора персонала. Большинство компаний стремятся получить представление о личности кандидата до принятия решения о его найме, чтобы оценить наличие необходимых качеств и избежать потенциально агрессивного поведения. Большинство исследований, рассматриваемых выше, ограничиваются данными из одной социальной сети, что снижает эффективность предсказательных моделей. Кроме того, одним из основных недостатков существующих подходов является ограниченность в оценке лишь таких характеристик, как добросовестность, доброжелательность, нейротизм, открытость опыту и экстраверсия, в соответствии с методологией *Big Five*. Это не позволяет

получить полное представление о кандидате. Именно по этой причине, принято решение использовать метод оценки личности Майерс-Бриггс, который широко используется именно в сфере рекрутинга.

Исходя из анализа помянутых исследований, в рамках диссертационной работы решено использовать метод оценки личности Майерс-Бриггс, в основе которой лежат идеи аналитического психолога Карла Густава Юнга.

1.6 Построение психологического портрета человека на основе открытой информации из социальных сетей

Типирование личности – система категоризации людей в соответствии с их склонностью думать и действовать определенным образом. Типирование личности пытается найти самые широкие, наиболее важные способы, которыми люди отличаются, и осмыслить эти различия, сортируя людей по значимым группам.

1.6.1 Системный подход Гордона Олпорта к изучению личности

Американский психолог Гордон Олпорт был одним из пионеров психологии личности, стремился определить количество существующих черт личности. Проведя исследование словаря, связанного с личностью, он пришел к выводу, что существует более 4000 терминов, описывающих различные черты. На основе этого Олпорт выделил три категории черт: кардинальные, центральные и вторичные [58].

Центральные черты личности встречаются гораздо чаще, чем кардинальные, и представляют собой основные элементы, формирующие характер большинства людей. Эти черты описывают личность в целом, такие как честность, дружелюбие, щедрость или тревожность. Олпорт полагал, что у большинства людей имеется от пяти до десяти центральных черт, при этом многие из них обладают различной степенью выраженности этих черт.

В отличие от центральных черт, вторичные черты проявляются в специфических ситуациях. Например, человек может обычно быть спокойным, но стать раздражительным под давлением, или наоборот, проявить тревогу при

необходимости публичного выступления. У большинства людей отсутствует одна кардинальная черта; чаще встречается сочетание нескольких центральных. Тем не менее, известных исторических личностей часто рассматривают через призму их наиболее ярких черт. Вот некоторые примеры:

- Мать Тереза: добрая, милосердная;
- Адольф Гитлер: злой, развратный;
- Альберт Эйнштейн: гениальный.

Кардинальные черты личности способны настолько определять человека, что их имена становятся синонимами его личности [59]. Этот подход хорошо подходит для выявления различных типов характеристик, но он громоздок и сложен в использовании. Многие из этих черт, например, очень похожи, что затрудняет отличие одних черт от других. Такая неоднозначность также затрудняет изучение этих качеств личности.

1.6.2 «Большая пятерка»

Модель «большой пятерки», также известная как пятифакторная модель, является наиболее широко распространенной теорией личности, которой придерживаются современные психологи [60]. Теория утверждает, что личность можно свести к пяти основным факторам, известным под аббревиатурой КАНОЭ или ОКЕАН.

Черты личности Большой пятерки – экстраверсия, доброжелательность (сотрудничество), открытость, добросовестность и нейротизм.

Добросовестность – импульсивность, неорганизованность против дисциплинированности, осторожности;

Доброжелательность – подозрительная, отказывающаяся от сотрудничества против доверчивой, полезной;

Нейротизм – спокойный, уверенный против тревожного, пессимистического;

Открытость к опыту – предпочитает рутину, практичность, а не воображение, спонтанность;

Экстраверсия – сдержанный, вдумчивый против общительного, веселого.

Модель Большой пятерки предполагает, что каждая личностная черта существует в виде спектра [61]. Это означает, что люди могут занимать позицию на шкале между двумя крайними значениями пяти основных измерений, как представлено это на рисунке 1.3.

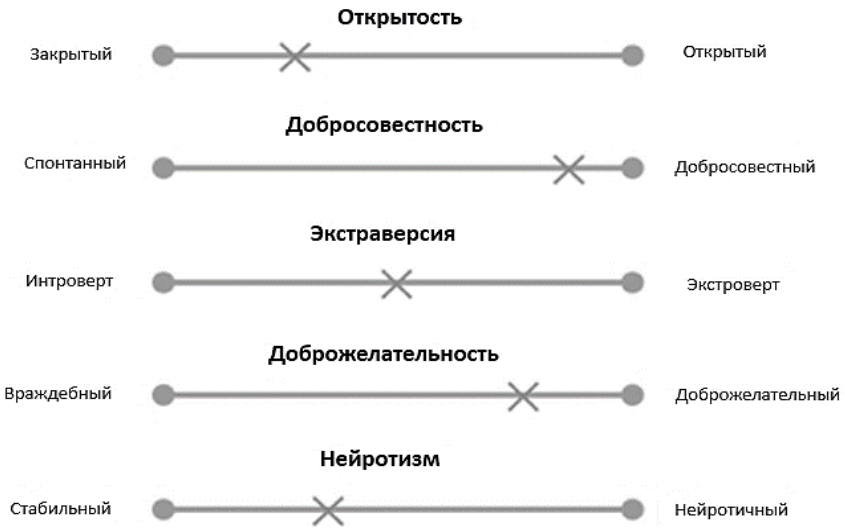


Рисунок 1.3 – Модель «Большая пятерка»

Например, при измерении экстраверсии человек не будет классифицироваться как чисто экстравертный или интровертный, а будет помещен на шкалу, определяющую его уровень экстраверсии. Ранжируя людей по каждой из этих черт, можно эффективно измерить индивидуальные различия в личности.

1.6.3 *HEXACO*

Модель *HEXACO* сохраняет многие аспекты Большой пятерки, но переосмысливает некоторые личностные факторы и вводит шестой. Сегодня как пятифакторная модель, так и *HEXACO* используются исследователями для анализа различий между людьми [62].

В состав *HEXACO* входят следующие факторы: честное смирение, эмоциональность (по ряду признаков напоминающая невротизм), экстраверсия, приятность, добросовестность и открытость опыту.

Модель *HEXACO* внесла следующие изменения в «Большую пятерку»:

1. Введен новый тип «честное смирение», который отражает степень, в которой человек ставит свои интересы выше интересов других. Этот фактор включает в себя такие аспекты, как искренность, справедливость, скромность, а также отношение к богатству и статусу.

2. *HEXACO* переопределяет некоторые факторы по сравнению с пятифакторной моделью. Например, фактор, обозначаемый как эмоциональность, примерно соответствует фактору невротизма, но включает в себя компоненты, отсутствующие в невротизме, такие как сентиментальность. Кроме того, версия приятности в модели HEXACO учитывает склонность человека к гневу, что может пересекаться с невротизмом в «Большой пятерке» [63].

1.6.4 *MBTI*

Наиболее известная и широко используемая система типирования личности была разработана в 1960-х годах Изабель Бриггс Майерс и ее матерью Кэтрин Бриггс. Они опирались на теорию личности швейцарского психиатра Карла Юнга, изложенную в его книге «Психологические типы», и создали один из самых популярных инструментов для оценки личности – индикатор типа Майерс-Бриггс, или *MBTI* (рисунок 1.4).

Система Майерс-Бриггс описывает личность человека через четыре противоположные функции личности, по-разному известные как дихотомии, предпочтения или шкалы. Первые три предпочтения основаны на трудах Юнга; Кэтрин Кук Бриггс добавила последнее предпочтение: «Суждение против восприятия», основанное на собственных наблюдениях [64].

Экстраверсия и интроверсия (*E/I*). Экстраверты ориентированы на действия и социальное взаимодействие. Например, они активно участвуют в мероприятиях, встречах с друзьями и вечеринках, черпая энергию из общения. Интроверты, напротив, предпочитают более глубокие беседы и размышления. Они могут проводить время в одиночестве, читая книгу или занимаясь творчеством, и именно в такие моменты восстанавливают свои силы.

Восприятие и интуиция (*S/N*). Люди, предпочитающие восприятие, акцентируют внимание на фактах и деталях. Например, такой человек может скрупулезно изучать инструкции перед тем, как начать новый проект. Те, кто склонен к интуиции, наоборот, больше интересуются концепциями и возможностями. Они могут мечтать о том, как их идея может изменить мир, принимая решения на основе интуиции и общей картины, а не конкретных данных.

Мышление и чувства (*T/F*). В процессе принятия решений мыслители стремятся к логически обоснованным решениям. Например, они могут анализировать данные и выбирать вариант, который наилучшим образом соответствует целям компании. Чувствующие, в свою очередь, принимают решения, учитывая эмоции и потребности других. К примеру, они могут выбрать менее выгодное, но более этичное решение, если это поможет команде сохранить хорошую атмосферу.

Суждение и восприятие (*J/P*). Люди, склонные к суждению, предпочитают четкую структуру и заранее спланированные действия. Например, такой человек может составить детальный план проекта с определенными сроками. В то время как те, кто ориентирован на восприятие, более гибки и открыты для изменений. Они могут менять свои планы в зависимости от ситуации, принимая новые идеи и возможности, которые возникают на пути. Эти две тенденции взаимодействуют с другими шкалами, и каждый человек иногда проявляет черты экстраверсии или интроверсии. Шкала суждения и восприятия (J-P) помогает понять, как индивидуум воспринимает и обрабатывает новую информацию [65].

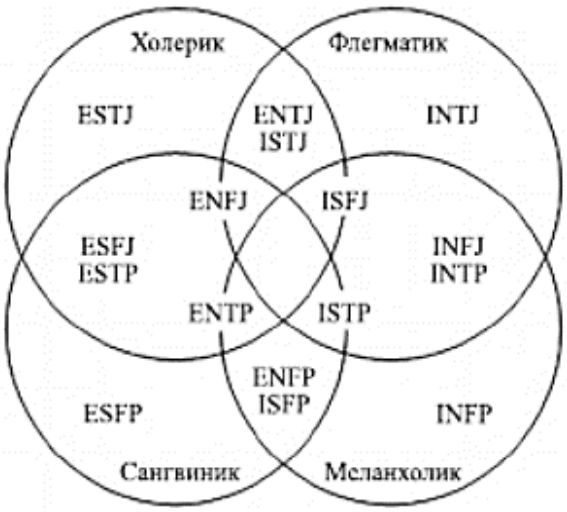


Рисунок 1.4 – Индикатор типа Майерс-Бриггс

Майерс и Бриггс выделили 16 типов личности, основанных на четырех личностных предпочтениях. Каждый тип личности обозначается четырехбуквенным кодом, каждая из которых обозначает одно из личностных предпочтений [66]:

ISTJ – Инспектор. В целом, *ISTJ* сдержанны, тихи и практичны. Они ценят организованность во всех аспектах жизни. Они лояльны и любят следовать традициям. Люди с типом личности *ISTJ* склонны к тщательному планированию, и когда дела не организованы, они могут стать беспокойными.

ISTP – Изобретатель. *ISTP* индивидуалистичны и предпочитают иметь время для размышлений в одиночестве. Они наслаждаются новым опытом, практической деятельностью и свободой работать в свое время и в своем темпе. Эти люди ориентированы на результат; их основная задача – рассмотреть проблему и определить ее первопричину, что помогает найти эффективное и разумное решение.

ISFJ – Защитник. Люди с этим типом личности наблюдательны и сосредоточены на других людях. Они проницательны и способны запоминать мелкие детали о других. Типы личности *ISFJ* чутки и чувствительны к эмоциям и чувствам других, зная, когда протянуть руку помощи. Они предпочитают факты абстрактным теориям и лучше всего учатся на практике.

ISFP – Художник. Личности *ISFP* легки и миролюбивы. Поскольку им нравится оставлять свои варианты открытыми, они откладывают принятие

решений до самой последней минуты, чтобы быть уверенными, что смогут учесть любые потенциальные изменения. Они добрые и чувствительные, но также дружелюбные и тихие – они могут проводить время с другими, но предпочитают делать это в небольших группах людей. Они заботливы, внимательны, спокойны и склонны принимать других такими, какие они есть, что делает их людьми, с которыми легко ладить.

INFJ – Адвокат. *INFJ* логичны и эмоциональны, но при этом креативны и аналитичны. У них сильная интуиция и они способны хорошо понимать эмоциональные потребности. Несмотря на то, что у них замкнутый характер, при взаимодействии с другими они способны формировать значимые связи и с удовольствием протягивают руку помощи. Однако принятие решений для них иногда может быть трудным.

INFP – Посредник. Тип личности *INFP* описывается как идеалистический, интровертный и творческий по натуре, движимый высокими ценностями. Их главный интерес заключается в поиске способов сделать мир вокруг них лучше. Они не только хотят понять себя и то, как они вписываются в мир, но также хотят понять, как они могут помочь другим.

INTJ – Аналитик. *INTJ* используют интровертную интуицию для определения значений, закономерностей и возможностей. Когда им сообщают факт, они смотрят дальше, надеясь узнать, что он на самом деле означает. Организуя свои мысли таким образом, чтобы они могли видеть причинно-следственные связи различных действий/реакций. У них высокий уровень интроверсии, поэтому они предпочитают работать самостоятельно, а не в команде.

INTP – Мыслитель. Люди с типом личности *INTP*, как правило, любят проводить время в одиночестве, размышляя о том, как все работает, а затем находя решения любых проблем, с которыми они могут столкнуться. Они предпочитают сосредотачиваться на своих внутренних мыслях и своем внутреннем мире, а не на внешнем, что делает их спокойными и аналитическими.

ESTP – Командир. Личности *ESTP* общительны, драматичны и ориентированы на действия; им нравится проводить время с самыми разными

людьми, и их интересует здесь и сейчас, а не широкий взгляд на жизнь. Они не любят рутину и чрезмерное планирование, вместо этого им нравится импровизировать и адаптироваться. Они предпочитают практичность абстрактным теориям; это позволяет им работать с простой информацией, думать о ней рационально и находить немедленное решение.

ESTJ – Директор. *ESTJ* описываются как логичные и напористые, гарантирующие соблюдение сроков и правил. Они традиционны, имеют сильные убеждения и ожидают того же от окружающих. Придавая большое значение стандартам, правилам и безопасности, для этих людей важно поддерживать статус-кво. Им легче, чем другим, брать на себя ответственность за ситуацию, поэтому они часто могут оказаться на руководящих должностях. В социальных ситуациях они чрезвычайно честны, что некоторые люди могут счесть чрезмерно критичным.

ESFP – Исполнитель. *ESFP* часто описываются как спонтанные, общительные и находчивые. Им нравится быть в центре внимания, и они обладают чрезвычайно интересным характером. Теоретическое обучение не является для них естественным, и именно поэтому они часто испытывают трудности в традиционных классах; взаимодействие с другими людьми и окружающей средой – это то, что им нравится больше всего.

ESFJ – Энтузиаст. Личности *ESFJ* добросердечны, нежны, лояльны и общительны. Черпая энергию от других людей, такие люди способны успешно поощрять других быть лучшей версией себя и с трудом выслушивают любые негативные комментарии о людях, с которыми они близки. Им нравится контролировать свое окружение, поэтому они преуспевают в среде, где могут брать на себя инициативу, организовывать и планировать.

ENFP – Чемпион. Личности *ENFP* считаются полными энтузиазма, обаятельными и творческими и делают все возможное в ситуациях, когда им предоставляется свобода исследовать различные идеи. Они обладают отличными навыками работы с людьми и склонны искренне учитывать чувства других. Они постоянно думают о новых и творческих идеях для реализации, что может стать недостатком, поскольку может заставить их отложить важные задачи на второй план.

ENFJ – Даритель. *ENFJ* – теплые личности, с общительным и чувствительным характером; настолько общительны, что способны подружиться с любым типом личности, даже если они очень интровертированы! Они обладают способностью не только чувствовать, что чувствуют другие, но и влиять на других (а иногда и манипулировать ими). При этом у них сильная система ценностей, которая уравновешивает эти характеристики и помогает им помогать другим быть лучшей версией себя.

ENTP – Политик. *ENTP* описываются как новаторские и умные люди; им нравится придумывать разные идеи и теории, но они часто делают это, прежде чем приступить к реализации предыдущих. Им нравится думать о будущем и более широкой картине, и это приводит к тому, что они не завершают проекты, поскольку меньше сосредотачиваются на текущих потребностях и насущных деталях. Им нравится общаться с широким кругом людей и нравятся дебаты; Один из распространенных мифов заключается в том, что люди с этим типом личности любят спорить просто ради спора.

ENTJ – Предприниматель. Будучи очень экстравертными, *ENTJ* любят проводить время с другими людьми и процветают за счет своей энергии. У них сильные речевые навыки, которые помогают им вести содержательные разговоры. Когда у них есть твердое мнение о ситуации, они с нетерпением ждут возможности поделиться им с другими. Однако, хотя они и хорошие собеседники, им сложно понять эмоции других, а также выразить свои собственные.

Другие личностные системы, такие как «Большая пятерка» или «*HEXACO*», обычно говорят о личностных чертах изолированно, что часто менее полезно при попытке концептуализировать человека в целом.

1.7 Сфера применения

Работа с социальными сетями может улучшить работы специалистов отдела кадров. Эффективный подбор персонала имеет важное значение для успеха компании. Наем квалифицированных и опытных специалистов является ключом к росту, инновациям и конкурентоспособности на рынке. В постоянно

меняющейся бизнес-среде компаниям необходимо применять современные подходы к привлечению лучших кандидатов, особенно через социальные сети.

Использование социальных сетей для рекрутеров имеет множество преимуществ. Фактически, им намного проще начать процесс набора персонала. Благодаря использованию социальных сетей им легче сделать первоначальный выбор получаемых заявок. С помощью фильтрации рекрутеры могут ориентироваться на профили, которые соответствуют ожиданиям, установленным для рассматриваемой должности, принимая во внимание несколько факторов. Профили кандидатов обычно содержат гораздо больше информации, чем резюме. Более того, через посты рекрутер может получить представление о личности, интересах и поведении. Изучив профили соискателя, наниматель сможет узнать о нем не только как о специалисте, но и как о человеке, имеющем определенный набор личностных качеств, необходимых для работы в команде. Рекрутеры могут учитывать эти дополнительные элементы в процессе принятия решений. Но существует огромная проблема, ручная работа с социальными сетями в процессе найма персонала является трудоемкой и время затратной, особенно при анализе большого объема информации и поиске потенциальных отклонений в поведении кандидатов [67]. Блок-схема процесса подбора персонала представлена на рисунке 1.5.

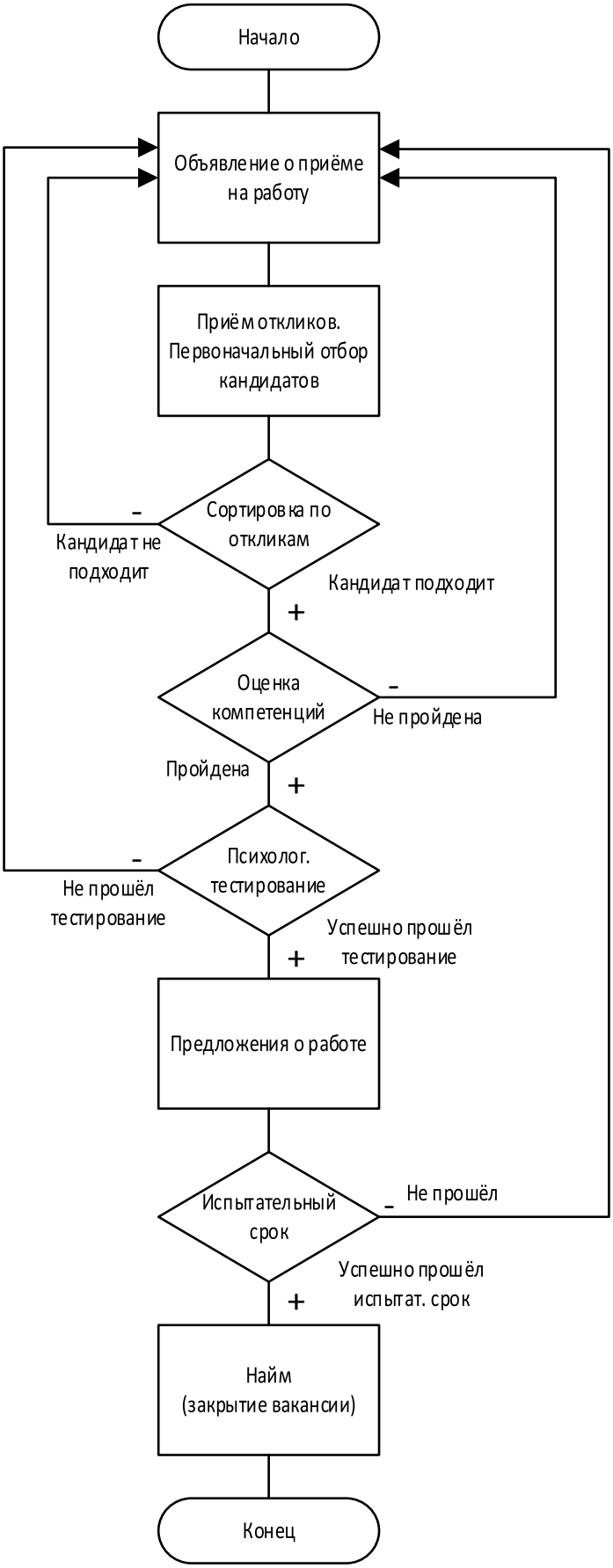


Рисунок 1.5 – Блок-схема процесса подбора персонала

Исходя из рисунка 1.5, можно сделать вывод, что ручной процесс подбора персонала приводит к ограниченности в объеме обрабатываемой информации и увеличению временных затрат на процесс найма (кроме того, если необходимо проводить дополнительное тестирование или анализировать дополнительные данные из социальных сетей, например, на предмет отклоняющегося поведения, что также может увеличить время, затраченное на процесс найма) (рисунок 1.6). Внедрение автоматизированных подходов, таких как анализ психологического портрета пользователей и оценка вероятности отклоняющегося поведения, может решить эти проблемы и значительно улучшить процесс найма.

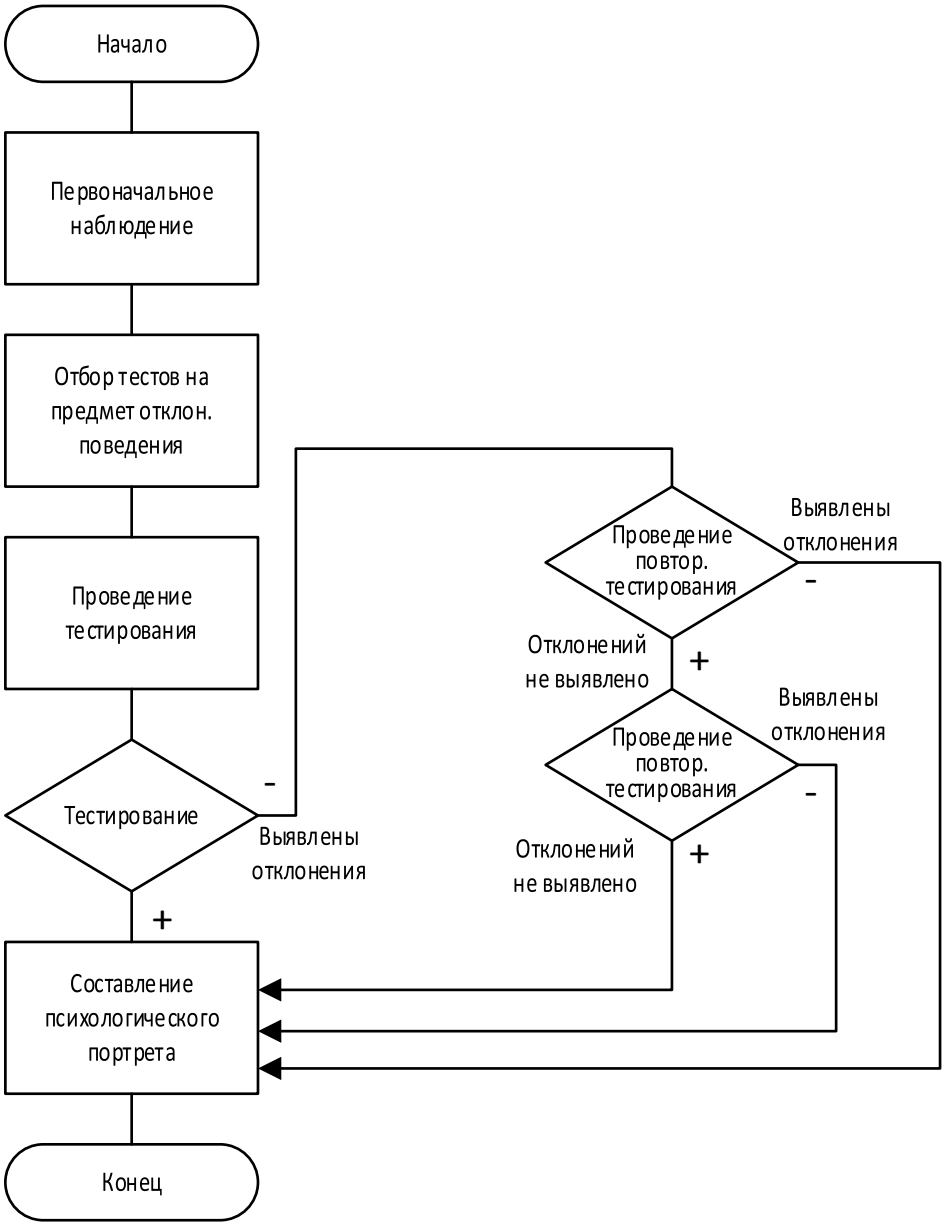


Рисунок 1.6 – Блок-схема процесса выявления отклоняющегося поведения

Использование специализированных алгоритмов и программного обеспечения для автоматизированного сбора и анализа данных из социальных сетей позволит обрабатывать большие объемы информации и выявлять потенциальные отклонения в поведении кандидатов. Применение методов машинного обучения и анализа текста для автоматической оценки психологического портрета кандидатов на основе их активности в социальных сетях позволит выявлять ключевые черты личности и предсказывать вероятность отклоняющегося поведения [68]. В контексте данной работы под отклоняющимся поведением будет рассмотрено нарушение социальных норм, которое может проявляться в форме агрессии, неуважительного общения, экстремистских высказываний.

Таким образом, внедрение автоматизированных подходов в работу с социальными сетями может существенно оптимизировать процесс найма персонала, снизить временные и ресурсные затраты и улучшить качество подбора кандидатов.

1.8 Закон о персональных данных

Общедоступные данные представляют собой персональные данные, которые субъект сделал публичными и доступными по собственному желанию, либо по своему запросу [69]. Правовые нормы единогласно подтверждают, что если субъект персональных данных сам разместил какую-либо информацию о себе в открытом доступе, то она становится публичной и может быть использована третьими лицами в соответствии с законодательством Российской Федерации.

Согласно, ч. 1 статьи 152.2. Гражданского кодекса Российской Федерации: «Если иное прямо не предусмотрено законом, не допускаются без согласия гражданина сбор, хранение, распространение и использование любой информации о его частной жизни, в частности сведений о его происхождении, о месте его пребывания или жительства, о личной и семейной жизни.

Не являются нарушением правил, установленных абзацем первым настоящего пункта, сбор, хранение, распространение и использование информации о частной жизни гражданина в государственных, общественных или иных публичных интересах, а также в случаях, если информация о частной жизни гражданина ранее стала общедоступной либо была раскрыта самим гражданином или по его воле» [70].

Наряду с приведенной статьей гражданского кодекса РФ, основными законами, регулирующими данный вопрос, являются: Федеральный закон Российской Федерации от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации»;

Федеральный закон Российской Федерации от 27 июля 2006 г. № 152-ФЗ «О персональных данных».

Статья 8 Федерального закона Российской Федерации от 27 июля 2006 г. № 152-ФЗ «О персональных данных» содержит следующее: «В целях информационного обеспечения могут создаваться общедоступные источники персональных данных (в том числе справочники, адресные книги). В общедоступные источники персональных данных с письменного согласия субъекта персональных данных могут включаться его фамилия, имя, отчество, год и место рождения, адрес, абонентский номер, сведения о профессии и иные персональные данные, сообщаемые субъектом персональных данных». Как было указано ранее в том же Федеральном законе и разъяснениях Роскомнадзора, письменное согласие равнозначно «галочке» в веб-форме: согласие на обработку персональных данных может быть получено путем установки «галочки» пользователем в соответствующей веб-форме [71].

В соответствии с подпунктом 1 пункта 3 статьи 6 Федерального закона Российской Федерации от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации» обладатель информации вправе определять порядок и условия доступа к информации, а также разрешать осуществление иных действий с информацией. Информация, размещаемая в форме открытых данных, является общедоступной (пункт 4 статьи 7 Федерального закона

№ 149-ФЗ), что в свою очередь означает отсутствие ограничений по ее использованию (пункт 1 статьи 7 Федерального закона № 149-ФЗ).

Политика конфиденциальности в сети «ВКонтакте»

Политика конфиденциальности социальной сети «Вконтакте» – документ о правилах обработки персональных данных посетителей сайта [72].

В пункте 6.3. «Настройка пользователем уровня конфиденциальности информации о себе» сказано: «6.3.1. Пользователь вправе, с учетом ограничений, предусмотренных п. 6.2. настоящих Правил, установить в отношении своих персональных данных (п. 4.1. Правил), а также в отношении информации, предусмотренной п.п. 4.2.3. настоящих Правил, один из следующих уровней конфиденциальности:

а) информация доступна всем Пользователям Сайта;

б) информация доступна всем Пользователям Сайта, за исключением определенных Пользователей;

в) информация доступна лицам, имеющим на Сайте статус друзей Пользователя, а также лицам, имеющим статус их друзей;

г) информация доступна только лицам, имеющим на Сайте статус друзей пользователя;

д) информация доступна только некоторым лицам, имеющим на Сайте статус друзей Пользователя;

е) информация доступна некоторым спискам друзей (в случае, если пользователь создал хотя бы один список друзей с помощью инструментария «Создать список» в разделе «Мои Друзья»);

ж) информация доступна только Пользователю».

Согласно пункту 6.3.3. «Администрация Сайта не несет ответственности за разглашение персональных данных Пользователя другими Пользователями Сайта, получившими доступ к таким данным в соответствии с выбранным Пользователем уровнем конфиденциальности».

Исходя из пункта 6.4. «Пользователь самостоятельно определяет условия и предоставляет доступ к своим персональным данным неограниченному кругу лиц, в том числе путем регистрации и использования стандартной функциональности, а также с помощью настроек приватности и видимости своей персональной страницы в рамках предоставленной Пользователю функциональности в соответствии с п. 6.3 настоящих Правил". Администрация Сайта не инициирует и не влияет на такой выбор Пользователя, а также не имеет цели получить у Пользователя разрешение на распространение его персональных данных. Обработка персональных данных, сделанных Пользователем доступными неограниченному кругу лиц, осуществляется Администрацией Сайта на основании и в соответствии с условиями Правил пользования сайтом «ВКонтакте» и настоящих Правил».

Выводы по главе

В первой главе рассмотрены ключевые аспекты работы с большими данными. Проанализированы современные системы, используемые для обработки Больших данных, такие как системы мониторинга социальных сетей. В ходе анализа установлено, что статистические методы, используемые в большинстве таких систем, позволяют эффективно выявлять частоту упоминаний брендов и общие настроения пользователей. Однако следует отметить, что эти методы зачастую не учитывают контекст взаимодействия и уникальные предпочтения отдельных пользователей, что ограничивает глубину анализа и интерпретации данных.

Проведен анализ современного состояния моделей и методов интеллектуального анализа текстовых данных в социальных сетях. Социальные сети привлекли внимание, так как поведение пользователей можно использовать как для оценки личностных качеств, так и для выявления настроений и отклоняющегося поведения.

Отмечено, что сложность таких исследований заключается в необходимости сбора и анализа большого объема данных из социальных сетей,

а также из-за наличия нелинейных взаимосвязей между реальным поведением человека и его виртуальной активностью. В реальной жизни поведение человека определяется множеством факторов, включая социальные, культурные и личные аспекты. Эти факторы могут влиять на то, как пользователь взаимодействует с контентом в социальных сетях. Например, стресс на работе может привести к активному поиску поддержки в онлайн-сообществе, тогда как позитивное эмоциональное состояние может способствовать более открытым и социальным взаимодействиям. Кроме того, многие исследования опираются на данные лишь одной социальной сети, что ограничивает предсказательную модель, так как рассматривается только небольшая часть доступных цифровых следов. Для повышения эффективности анализа важно применять мультиканальный подход, который объединяет данные из нескольких социальных сетей. Это позволит учитывать разнообразие пользовательских взаимодействий и создавать более полные и точные модели поведения,

Для повышения качества определения психологического портрета предлагается рассматривать данные, размещенные в нескольких социальных сетях, а также искать аккаунты одного пользователя в пределах одной социальной сети. Для составления портрета предлагается использовать методологию *MBTI*, т.к. она включает профессиональную ориентацию и трудоустройство, чтобы помочь работодателям лучше понять личностные характеристики кандидатов и их потенциальное поведение в рабочей среде.

2 МЕТОДЫ И АЛГОРИТМЫ ФОРМИРОВАНИЯ ПСИХОЛОГИЧЕСКОГО ПОРТРЕТА ПОЛЬЗОВАТЕЛЯ СОЦИАЛЬНОЙ СЕТИ

Основываясь на результатах предыдущей главы, можно утверждать, что в современных исследованиях обработки информации, размещаемой пользователями в социальных сетях, методы анализа текстовых данных и машинного обучения играют ключевую роль в построении психологического портрета пользователя. Эти подходы позволяют глубже понять личностные характеристики и эмоциональное состояние пользователей, анализируя их сообщения, комментарии и другие формы взаимодействия. Использование алгоритмов машинного обучения помогает выявлять скрытые паттерны и корреляции, которые не всегда очевидны при традиционном анализе, что значительно повышает точность и глубину психологического анализа [73].

2.1 Описание социальной сети

Социальная сеть представляет собой кортеж $N = < U, G, E, F, B >$, в котором:

$U = \{u_1, u_2, ..., u_m\}$ – набор пользователей, зарегистрированных в социальной сети;

$G = \{g_1, g_2, ..., g_n\}$ – совокупность сообществ, $g_i \subset U$;

$E = \{e_1, e_2, ..., e_k\}$ – записи пользователей, размещаемые в социальной сети (статус, изображение, пост). Каждый пользователь внутри записи может поставить лайк этой записи и прокомментировать ее или поделиться этой записью на своей домашней странице. Каждому пользователю u_i соответствует некоторое подмножество E_i множества E;

$F = \{f_1, f_2, ..., f_p\}$ – набор функций, каждая из которых определена на множестве записей E;

$B = \{b_1, b_2, ..., b_l\}$ – набор поведенческих характеристик каждого пользователя $u \in U$ в группе $g \in G$. Каждая поведенческая характеристика

определяется некоторым подмножеством множества всех возможных комбинаций значений набора функций из F.

Каждый пользователь u_i характеризуется набором записей, размещенных в E, и набором поведений из B в социальной сети. Каждый пользователь u_i имеет набор записей $E_i = \{e_1^i, e_2^i, ..., e_{k_i}^i\}$ (верхний индекс соответствует номеру пользователя, нижний индекс номеру записи) и каждая запись $e_j \in E$ определяет значения функций: $f^j = \left(f_1^j, f_2^j, ..., f_p^j \right), f_i \in F, f_i^j = f_i(e_j)$.

Таким образом, набор значений функций, соответствующий каждому пользователю, позволяет конкретному пользователю сопоставить соответствующие поведенческие характеристики. В итоге каждому пользователю соответствуют поведенческие характеристики $B_i = \{b_1^i, b_2^i, ..., b_{l_i}^i\}$.

Для каждой записи может быть определено несколько функций, включая явные функции, такие как публикации, и неявные функции, такие как тег, категория, тональность и эмоция. Поскольку неявные функции не могут быть напрямую извлечены из записи, модели требуется шаг для извлечения этих функций, прежде чем оценивать сходство записей.

Модель, рассматриваемая в работе, учитывает пять функций записи:

- f_{cont} – функция записи (текстовое содержание) – это явная функция, которая представляет собой текстовую часть самой записи.
- f_{tags} – функция тегов. Запись может быть помечена набором тегов. Каждый тег представляет собой самостоятельное слово или выражение. В некоторых случаях теги могут быть помечены пользователем напрямую (явно). В некоторых других случаях он не помечен пользователем явно (неявно). Функция может быть описана следующим образом: допустим, запись, представлена как набор слов $W = \{w_1, w_2, ..., w_n\}$, где w_i – это отдельное слово. Если текущее слово w_i содержит первым символом тег («#»), мы удаляем символ «#» из этого слова и добавляем результат к строке-результату, разделяя его пробелом. Если слово не содержит тег, то пропускаем его и переходим к

следующему слову. После обработки всех слов, результатом работы функции будет строка, состоящая только из слов с тегами (без символов «#»), разделенная пробелами: $f_{tags}(W) = h(W, n)$, где n – количество слов в записи.

- f_{cate} – функция категории записи, отнесенной к той или иной категории, которая представлена отдельным словом или выражением.
- f_{sent} – функция настроения записи. Настроение может быть положительным, отрицательным, нейтральным.
- f_{emot} – функция эмоций в записи, которая отражает эмоции пользователя.

Реализация функций $f_{cate}, f_{sent}, f_{emot}$ происходит на уровне нейронных сетей. В частности, $f_{cate} = HC_1(W)$, где W – запись, HC_1 – нейронная сеть, обученная на классификации текстов по категориям; $f_{sent} = HC_2(W) \in \{positive, negative, neutral\}$, HC_2 – нейронная сеть, обученная на классификации текстов по настроению; $f_{emot} = HC_3(W) \in \{эмоция_1, эмоция_2, ..., эмоция_z\}$, HC_3 – нейронная сеть, обученная на классификации текстов по эмоциям пользователя.

Каждое поведение $b \in B$ определяется значением набора функций $\left(f_{cate}, f_{tags}, f_{sent}, f_{emot}, f_{cont}\right)$. Множество всевозможных комбинаций выдает класс B, каждый класс которого определяет поведенческую характеристику.

2.2 Оценка сходства признаков выражения

Поскольку содержимое признака представлено в виде набора текстовых выражений, их сходство определяется следующим образом: предположим, что $C_1 = \{c_1^1, c_1^2, ..., c_1^m\}$ и $C_2 = \{c_2^1, c_2^2, ..., c_2^n\}$ – 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:

$$Sim(C_1, C_2) = \frac{2 \times |C_1 \cap C_2|}{|C_1| + |C_2|} = \frac{2 \times t}{m + n}, \qquad (2.1)$$

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:

$$S_{sign}(e_i, e_j) = Sim\left(f_{sign}^i, f_{sign}^j \right), \tag{2.2}$$

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

2.2.1 Оценка сходства текстовых объектов

Для анализа текстов предлагается использоваться метод *TF-IDF*. Метод *TF-IDF* используется для представления текстов в виде векторов, которые могут быть сравниваемы по их содержанию.

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

$$TF = \frac{Количество\ вхождений\ n\text{-}граммы\ в\ записи}{Общее\ количество\ n\text{-}грамм\ в\ записи} \tag{2.3}$$

$$IDF = \log\left(\frac{Общее\ количество\ записей}{Количество\ записей\ содержащих\ n\text{-}грамму} + 1 \right) \tag{2.4}$$

$$TF\text{-}IDF = TF \times IDF \quad (2.5)$$

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:

$$D(k^1, k^2) = \frac{1}{N} \sum_{i=1}^{N} d_k , \quad (2.6)$$

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

$$d_k = \frac{\left| t_i^1 - t_l^2 \right|}{\max\left\{ t_i^1, t_l^2 \right\}} \quad (2.7)$$

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:

$$S_{Sim}(k^1, k^2) = 1 - D(k^1, k^2) \quad (2.8)$$

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

2.2.2 Оценка сходства между двумя записями

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:

$$S_{cate}(e_i, e_j) = Sim(f_{cate}(e_i), f_{cate}(e_j)); \quad (2.9)$$

$$S_{tags}(e_i, e_j) = Sim(f_{tags}(e_i), f_{tags}(e_j)); \quad (2.10)$$

$$S_{sent}(e_i, e_j) = Sim(f_{sent}(e_i), f_{sent}(e_j)); \quad (2.11)$$

$$S_{emot}(e_i, e_j) = Sim(f_{emot}(e_i), f_{emot}(e_j)). \quad (2.12)$$

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

$$S_{cont}(\mathrm{e}_i, \mathrm{e}_j) = Sim(f_{cont}(e_i), f_{cont}(e_j)) \quad (2.13)$$

Пусть e_i и e_j две рассматриваемые записи, значений функций, содержание, теги, категории, настроения и эмоции которых являются характеристиками записей: $f^i_{cont}; f^j_{cont}; f^i_{tags}; f^j_{tags}; f^i_{cate}; f^j_{cate}; f^i_{sent}; f^j_{sent}; f^i_{emot}; f^j_{emot}$. На основе подхода многоатрибутного сходства двух объектов сходство между двумя записями e_i и e_j оценивается следующим образом:

$$f_{entry}(e_i, e_j) = f_{ent}(S_{cont}(e_i, e_j), S_{tags}(e_i, e_j), S_{cate}(e_i, e_j), S_{sent}(e_i, e_j), S_{emot}(e_i, e_j)), \quad (2.14)$$

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:

$$f_{ent}(v_1, w, x, y, z) \le f_{ent}(v_2, w, x, y, z) \ \ if \ \ v_1 \le v_2 \quad (2.15)$$

$$f_{ent}(v, w_1, x, y, z) \le f_{ent}(v, w_2, x, y, z) \ \ if \ \ w_1 \le w_2 \quad (2.16)$$

$$f_{ent}(v, w, x_1, y, z) \le f_{ent}(v, w, x_2, y, z) \ \ if \ \ x_1 \le x_2 \quad (2.17)$$

$$f_{ent}(v, w, x, y_1, z) \le f_{ent}(v, w, x, y_2, z) \ \ if \ \ y_1 \le y_2 \quad (2.18)$$

$$f_{ent}(v, w, x, y, z_1) \le f_{ent}(v, w, x, y, z_2) \ \ if \ \ z_1 \le z_2 \quad (2.19)$$

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.).

2.2.3 Оценка сходства между двумя пользователями

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 – веса, отражающие степень важности 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:

$$S(U_1,U_2) = w_1 \times S_{publ}(U_1,U_2) + w_2 \times S_{conn}(U_1,U_2) + \\ + w_3 \times S_{like}(U_1,U_2) + w_4 \times S_{komm}(U_1,U_2), \quad (2.20)$$

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.

Указанные выше сходства вычисляются на основе формулы (2.1).

2.3 Объединение информации из двух социальных сетей

Объединение информации из разных социальных сетей является необходимым 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 разнообразными. 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.

Для решения этой проблемы был использован метод факторизации неотрицательных матриц [74] (*non-negative matrix factorization, NMF*) для каждого временного периода отдельно. Данный метод позволяет восстанавливать данные для тех пользователей, которые проявляли активность хотя бы в одной из социальных сетей. Процесс заполнения пропусков продемонстрирован на рисунке 2.1. После применения данного метода получены полностью заполненные данные об активности в обеих рассматриваемых социальных сетях на протяжении всех временных промежутков.

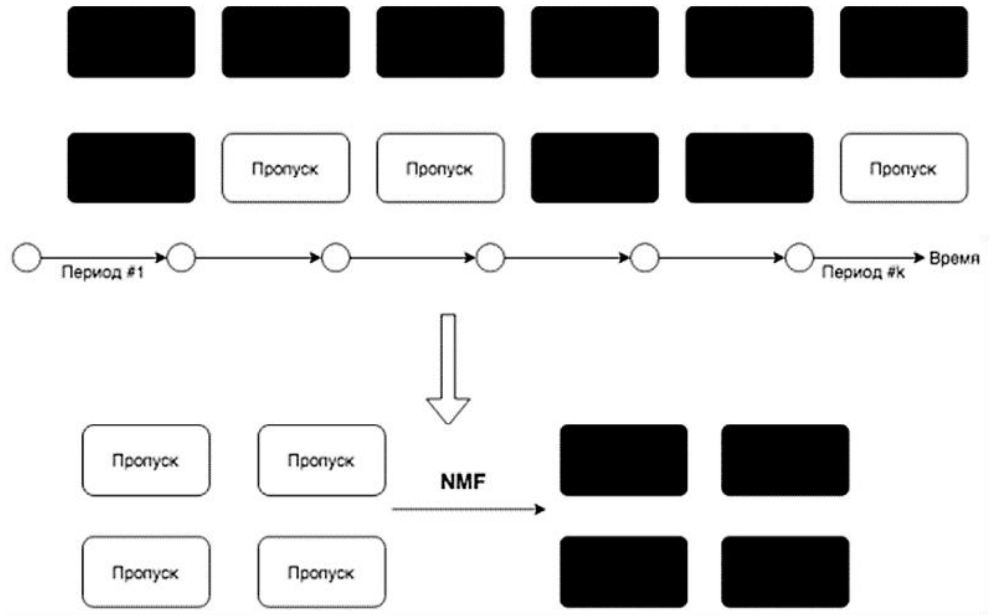


Рисунок 2.1 – Процесс объединения информации из двух социальных сетей

Для объединения информации из разных социальных сетей математическая модель может быть представлена следующим образом: имеются две социальные сети A и B, где каждая представлена в виде графа. Граф социальной сети – математическое представление социальной структуры, где узлы (вершины) графа соответствуют пользователям, а ребра графа

представляют связи между ними. G_A и G_B – графы социальных сетей A и B, соответственно.

$N(x)$ – множество соседних вершин для узла x в графе G_A. Обозначим $PR(n(x)i)$ как множество проекций вершины $n(x)i$ из графа G_A в граф G_B, при этом текстовое сходство должно превышать пороговое значение. Объединение всех множеств проекций $PR(n(x)i)$ образует множество $PR(x)$, которое содержит информацию из социальной сети B, соответствующую узлу X из сети A. Математически это можно записать следующим образом:

$$PR(x) = \cup_{n(x)i \in N(x)} PR(n(x)i), \qquad (2.21)$$

где $PR(x)$ – множество проекций для узла X из сети A, содержащее информацию из сети B.

Эта модель позволяет эффективно объединять информацию из различных социальных сетей, учитывая социальные связи пользователей и их атрибуты профиля.

2.4 Кросс-доменный аспектно-ориентированный анализ тональности текста

Подавляющее большинство моделей аспектно-ориентированного анализа тональности разработано для работы с текстами, относящимися к определенному смысловому домену, однако такие модели становятся непригодными при анализе документов, принадлежащих другому домену, относящихся к другой тематике [75]. Эта проблема появляется, когда первоначальная задача подразумевает анализ разнородных текстов. Публикации в социальных сетях отличаются многообразием тем, поэтому тексты каждого пользователя будут индивидуальны и специфичны. Так, модель аспектно-ориентированного анализа тональности, обученная на постах одной тематики, не сможет эффективно обрабатывать посты другой тематики, так как не обладает свойством извлекать информацию из терминов и выражений, специфичных для профиля (домена) последнего [76].

Также следует отметить, что на сегодняшний день большинство решений не подразумевают применение современных глубоких нейросетевых моделей для обработки естественного языка.

Для решения задачи совокупного выделения извлечения аспектов и тональности отношения к ним авторов текстов, принадлежащих различным доменам, предлагается использовать модель *IbDA-LSTM-CRF*.

Модель *IbDA-LSTM-CRF* разработана для решения задачи совокупного выделения извлечения аспектов и тональности отношения к ним авторов текстов, относящимся к различному контексту и на разные темы.

Модель *IbDA-LSTM-CRF* представляет собой метод для анализа тональности и выделения аспектов в текстах, основанный на комбинации нескольких технологий:

1. *LSTM* – используется для моделирования последовательностей текстовых данных и способна учитывать контекст и зависимости между словами в тексте [77].

2. *CRF* (*Conditional Random Fields*) – статистическая модель последовательности, которая используется для моделирования зависимостей между последовательными метками, такими как метки частей речи или метки именованных сущностей [78]. *CRF* учитывает взаимосвязи между соседними элементами, что позволяет моделировать контекстуальные зависимости.

3. *IbDA* (*Instance-Based Domain Adaptation*) – метод адаптации к домену, который позволяет модели адаптироваться к текстам из различных доменов. Это важно для задачи анализа тональности и выделения аспектов, так как тексты из разных доменов могут содержать уникальные особенности, требующие индивидуального подхода.

Итак, модель *IbDA-LSTM-CRF* использует *LSTM* для понимания контекста текста, *CRF* для моделирования зависимостей между аспектами и тональностями, а также метод *IbDA* для адаптации к различным доменам текстовых данных.

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 $S(i \in [1,...,n], i \in \mathbb{Z})$ 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 домена-источника и неразмеченных данных целевого домена, необходимо адаптировать модель для работы сразу в двух доменах. Предлагаемая в работе модель учитывает специфичность токенов 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.

Вначале входная последовательность S представляется в виде последовательности векторных представлений $T = \{t_1, t_2, ..., t_n\}$, $t_i \in R^{d_{emb}}$,

являющихся комбинациями эмбеддингов токенов и позиционных эмбеддингов, где d_{hid} – размерность эмбеддинга токена. Затем последовательность T с помощью модели *LSTM* кодируется в векторное представление $H = \{h_1, h_2, ..., h_n\}$, где $h_i \in R^{d_{hid}}$ – векторное представление элемента последовательности, получаемые на выходе *LSTM*, d_{hid} – размерность элемента последнего скрытого слоя *LSTM*. Данное преобразование можно записать в следующем виде $H = LSTM(\text{T})$.

Полученное векторное представление последовательности H затем подается на вход модели генерации доменно-зависимых весов и на вход слоям классификации токенов *Softmax* и *CRF* для получения выходной разметки.

Распределение классов токенов выходной последовательности, получаемое с помощью условных случайных полей выражается как:

$$p(y|h,\lambda) = \frac{1}{Z(h)} t^{\sum_{i=1}^{n} \lambda_i f_i(h, y_{i-1}, y_i)}, \quad (2.22)$$

где: h – последовательность признаковых представлений членов исходной последовательности; f – признаковая функция, выражающая связь между двумя последовательными токенами, которая может принимать значение 0 или 1; λ – обучаемые веса, определяющие вклад признаковых функций в распределение; $Z(h)$ – нормирующий коэффициент, который выражается следующим образом:

$$Z(h) = \sum_{y' \in y} \sum_{i=1}^{n} \lambda_i f_i(h, y'_{i-1}, y'_i). \quad (2.23)$$

Обучение весов λ условного случайного поля происходит путем минимизации отрицательного логарифма функции правдоподобия.

$$L_{CRF} = L(y, h, \lambda) = -\log \prod_{k=1}^{m} p(y^k | h^k, \lambda). \quad (2.24)$$

Второй параллельный *CRF* слой классификации токенов представляет собой обычный полносвязный слой с функцией активации *Softmax*, который переводит векторное представление каждого токена h_i в распределение

вероятностей p_i^{SM} принадлежности его к одному из семи классов, принадлежащих выходной *BIO*-разметке: $p_i^{SM} = soft\max(W^{SM} h_i + b^{SM})$, где $W^{SM} \in R^{d_{hid} \times |y|}$ – обучаемая матрица весов *Softmax* слоя, а $b^{SM} \in R^{|y|}$ – обучаемый вектор смещения *Softmax* слоя. В отличие от классического процесса обучения весов слоя с использованием кросс-энтропии в качестве функции потерь, модель *IbDA-LSTM-CRF* использует взвешенную кроссэнтропию в качестве функции потерь, применяя для генерации весов отдельную модель.

2.5 Источники данных и выборка для обучения нейронной сети

В качестве признаков для обучения ИНС используются данные из открытых баз данных (*MBTI Dataset, Kaggle Personality Datasets, Personality Recognition Datase* и д.р.), а также две следующие группы исходных данных:

– численные статистические параметры со страницы пользователя – скалярные характеристики, такие как количество подписчиков, количество друзей и число сообществ. Эти параметры предоставляют количественные данные о взаимодействии пользователя в социальной сети, что может быть полезным для анализа его социального поведения и влияния на его личностные характеристики [79];

– текстовые параметры – последовательность слов в публикации (посте) и в статусе, комментарии, название и информацию сообществ, теги и т.д. [80]. При этом эмодзи-символы заменяются соответствующим текстовым описанием, согласно таблице 2.1.

Таблица 2.1 – Таблица символов выражений эмоций

Эмоции	Кодировка	Эмодзи-символ
1	**2**	**3**
Радость/Счастье/Влюбленность	❤&#65039	❤️
	☺&#65039	☺️

1	2	3
Радость/Счастье/Влюбленность	&#128515	
	&#128538;	
	☺️	
	&#128514;	
	&#129303;	
	&#129315;	
	&#128526;	
Грусть	☹️	
	&#128543;	
	&#128542;	
	&#128577;	
	&#128546;	
Страх	&#128534;	
	&#128561;	
Гнев	&#128545;	
	&#128556;	
	&#128548;	
Отвращение	&#128551;	
	&#128556;	

1	2	3
Отвращение	&#128548;	
Удивление	&#128561;	
	&#128563;	
	&#128565;	
Разочарование	&#128551;	
	&#128560;	
	&#128553;	
	&#128565;	

Каждому набору эмодзи соответствует своя характеристика, согласно типологии личности Майерс-Бриггс.

В качестве исходных групп для обучения ИНС также был подобран следующий список сообществ, обладающих значимым количеством текстовых сообщений, характерных для психологического состояния пользователей с отклоняющимся поведением. В формируемую выборку для последующего обучения ИНС добавлялась как нормализованная текстовая информация, так и доступная статистическая информация из профилей подписчиков и комментаторов группы [81].

Примеры сообществ:

– *psyhelp_online* («*Help* | Психологическая помощь»);

– *psychoconsult* («Психологическая помощь»);

– *caps_rage.*

Выходными значениями для нейронной сети являются:

Exit={INTJ, INTP, ENTP,ENTJ, INFJ,INFP, ENFP, ENFJ, ISTJ, ESFJ, ISFJ, ESTJ, ISTP,ESTP,ISFP, ESFP, DB}

где *INTJ* – Аналитик;

INTP – Мыслитель;

ENTP – Политик;

ENTJ – Предприниматель;

INFJ – Адвокат;

INFP – Посредник;

ENFP – Чемпион;

ENFJ – Даритель;

ISTJ – Инспектор;

ESFJ – Энтузиаст;

ISFJ – Защитник;

ESTJ – Директор;

ISTP – Изобретатель;

ESTP – Командир;

ISFP – Художник;

ESFP – Исполнитель;

DB – отклоняющееся поведение.

Алгоритм психологического анализа создает описание психологического портрета человека, основываясь на модели анализа личности.

2.6 Алгоритм предварительной обработки и очистки текстовых данных

Следует учитывать, что тексты в социальных сетях обладают рядом особенностей, которые усложняют семантический анализ. Главной из них является наличие грамматических ошибок, которые могут искажать смысл и затруднять интерпретацию сообщений. Кроме того, пользователи часто используют неформальный лексикон, включая сленг, аббревиатуры и эмодзи, что также добавляет сложности в анализ. Эти факторы могут влиять на точность извлечения значений и понимание эмоциональной тональности текста, что требует применения специальных методов обработки и

предобработки данных [82]. По этой причине, публикации, комментарии, теги и т.д. были пропущены через этап предобработки данных, который включает в себя описанные ниже этапы (рисунок 2.2).

1. Токенизация – первый шаг в обработке текстовых данных, который включает разделение строк на более мелкие части, называемые токенами. Поскольку язык в его исходной форме не может быть точно воспринят машиной, токенизация помогает упростить понимание текста. [83]. Основной метод токенизации заключается в разделении текста на токены на основе пробелов и знаков препинания. Это позволяет выделить отдельные слова и фразы, что облегчает дальнейший анализ. [84].

2. Нормализация [85]. На этапе нормализации удаляются элементы, которые не влияют на смысл текста, такие как стоп-слова (например, «если», «но», «а», «значит») и знаки препинания. Также приводятся все слова к единому регистру (нижнему), чтобы уменьшить количество уникальных токенов и упростить обработку. Нормализация позволяет сосредоточиться на значимых частях текста, улучшая качество анализа.

3. Орфография [86]. Процесс исправления орфографических ошибок является важным шагом, так как ошибки могут искажать смысл сообщений. Для этой цели может использоваться библиотека *SpaCy*, которая автоматически исправляет орфографические ошибки, обеспечивая более высокую точность дальнейшего анализа.

4. Лемматизация [87]. Этот этап позволяет уменьшить вариативность словоформ и сосредоточиться на основных значениях слов, что значительно улучшает эффективность семантического анализа и извлечения информации.

5. Векторизация – это процесс преобразования текстовых данных в числовые векторы, которые могут быть обработаны алгоритмами машинного обучения [88]. Векторизация позволяет представить слова или тексты в виде массивов чисел, где каждое число отражает определенные характеристики или свойства токена. Векторизация происходит с помощью моделей «*BERT*» [89].

Рисунок 2.2 – Этапы предварительной обработки данных

Модель «*BERT*» производит перевод таких текстовых признаков, в числовые представления. В работах [90] слова представлялись либо в виде уникально индексированных значений (горячее кодирование), либо, что более удобно, в виде нейронных внедрений слов, где словарные слова сопоставляются с внедрениями признаков фиксированной длины, которые являются результатом таких моделей, как *Word2Vec* [91] или *Fasttext* [92]. *BERT* имеет преимущество перед такими моделями, как *Word2Vec*, потому что, хотя каждое слово имеет фиксированное представление в *Word2Vec* независимо от контекста, в котором оно появляется, *BERT* создает представления слов, которые динамически информируются словами вокруг них. Помимо учета очевидных различий, таких как многозначность, контекстно-зависимые вложения слов охватывают и другие формы информации, которые приводят к более точному представлению признаков, что, в свою очередь, приводит к повышению производительности модели.

В векторной модели текстовый документ представляется в следующем виде: $t = \{w_1, w_2, ..., w_k\}$, где w_i – слово, содержащееся в тексте. Если каждому слову w_i назначить вес m_i согласно важности этого термина в записи, то запись может быть описана формулой:

$$d = \{m_1, m_2, ...m_k\} \tag{2.25}$$

Рассматривая элементы $w_1, w_2, ..., w_k$ в системе координат k-мерного пространства, величины $m_1, m_2, ...m_k$ будут являться значениями этих элементов

в данной системе координат. Тогда множество $\{m_1, m_2, ...m_k\}$ будет представлять собой вектор в k-мерном векторном пространстве.

Все эти шаги служат для уменьшения шума, присущего обычному тексту, и повышения точности результатов классификатора. Для решения указанных задач была использована библиотека *NLTK* [93] и *spaCy* [94].

Библиотека *NLTK* – ведущая платформа для создания программ на языке *Python* для работы с данными человеческого языка.

SpaCy предоставляет широкий спектр предварительно обученных моделей, которые могут быстро анализировать текст и извлекать различные лингвистические особенности. Эти функции включают в себя теги части речи, именованные объекты, синтаксические зависимости, границы предложений и многое другое. *SpaCy* известен своей исключительной производительностью и масштабируемостью. Библиотека реализована на *Cython*, языке программирования, который компилирует код Python в высокоэффективные модули C/C++. Это позволяет *SpaCy* невероятно быстро обрабатывать текстовые данные, что делает его пригодным для крупномасштабных приложений НЛП и систем реального времени.

Эти шаги помогут подготовить текст для дальнейшего анализа [95, 96], а именно для построения моделей машинного обучения [97] в задаче построения психологического портрета пользователя и вероятности отклоняющегося его поведения [98].

2.7 Парсинг данных из социальной сети

Парсинг данных из социальной сети может быть осуществлен различными подходами в зависимости от целей и объема информации. В данной диссертационной работе использовался метод *API* (*Application Programming Interface*) [99, 100]. Использование *API* обеспечивает структурированный и авторизованный доступ к данным.

Для автоматизированного сбора информации был получен токен «ВКонтакте» и «Одноклассники» (ключ доступа). Ключ доступа сообщает

серверу, от имени какого пользователя осуществляются запросы, и какие права доступа он выдал приложению. Официальный *API* предоставляет несколько типов ключей. Для данной работы был использован тип «ключ доступа пользователя». Данный ключ используется для работы со всеми открытыми методами *API* (за исключением методов секции *secure*). Данный набор методов полностью удовлетворяет требованиям для автоматизированного сбора информации. Процесс парсинга данных из социальной сети представлен на рисунке 2.3.

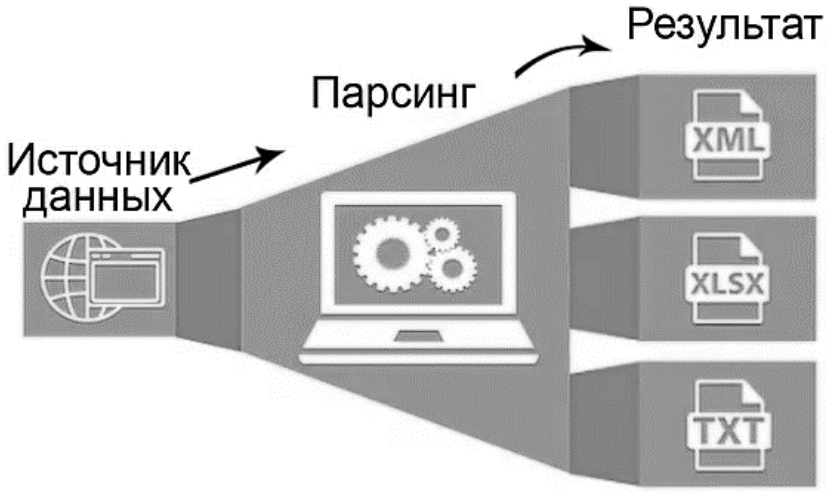


Рисунок 2.3 – Процесс парсинга данных

Например, в таблице 2.2 приведена общая информация об анализируемых параметрах и соответствующие методы *API* «ВКонтакте», осуществляющие доступ к базе данных [101].

Таблица 2.2 – Анализируемые параметры и соответствующие методы *API* «ВКонтакте»

№	Параметр	Метод	Тип исходных данных
1	Список записей со стены пользователя или сообщества	*wall.get*	Последовательность слов
2	Текст статуса пользователя или сообщества	*status.get*	Последовательность слов
3	Расширенная информация о пользователях (в т.ч. число подписчиков)	*users.get* (параметр *followers_count*)	Скаляр
4	Список идентификаторов друзей пользователя или расширенная информация о друзьях пользователя (в т.ч. общее число друзей)	*friends.get* (параметр *count*)	Скаляр
5	Список сообществ, в которых состоит пользователь		Скаляр
6	Текстовое содержание сообщества, в котором состоит пользователь (включая посты, комментарии)	– *getSubscriptions*	Последовательность слов

Эти параметры включают текстовые данные, такие как записи со стены и статус пользователя, а также скалярные данные, например, количество друзей и подписчиков, что позволяет формировать более точное представление о поведении и интересах пользователя.

2.8 Архитектура нейронной сети

В работах [102, 103, 104] были описаны основные характеристики нейрона и проведен алгоритм работы нейронных сетей. При проведении исследования были рассмотрены различные архитектуры нейронных сетей [105, 106, 107]. В конечном итоге были выбраны три классификации: *LSTM, RNN* и *RF*. *LSTM, RNN* и *RF* обеспечивает всесторонний подход к составлению психологического портрета пользователя, позволяя эффективно обрабатывать как последовательные, так и структурированные данные, предоставляя одновременно высокую точность и возможность интерпретации результатов.

1. Сеть долгосрочной и краткосрочной памяти (*LSTM* (*Long Short-Term Memory*) является модифицированной формой рекуррентной нейронной сети или *RNN* [108]. У *RNN* есть обратные связи, которые помогают им запоминать последовательную информацию со временем [109, 110]. Однако для гораздо более длинных последовательностей это создает проблему. С течением времени градиент функции потерь экспоненциально затухает, что вызывает проблему исчезающего градиента. В нейросетевых моделях *LSTM* используются три типа ворот (*Input Gate, Forget Gate, Output Gate*), которые играют ключевую роль в управлении информацией в ячейках памяти [111, 112] (рисунок 2.4). Эти ворота помогают *LSTM* запоминать или забывать информацию, что делает их более эффективными по сравнению с традиционными рекуррентными нейронными сетями (*RNN*).

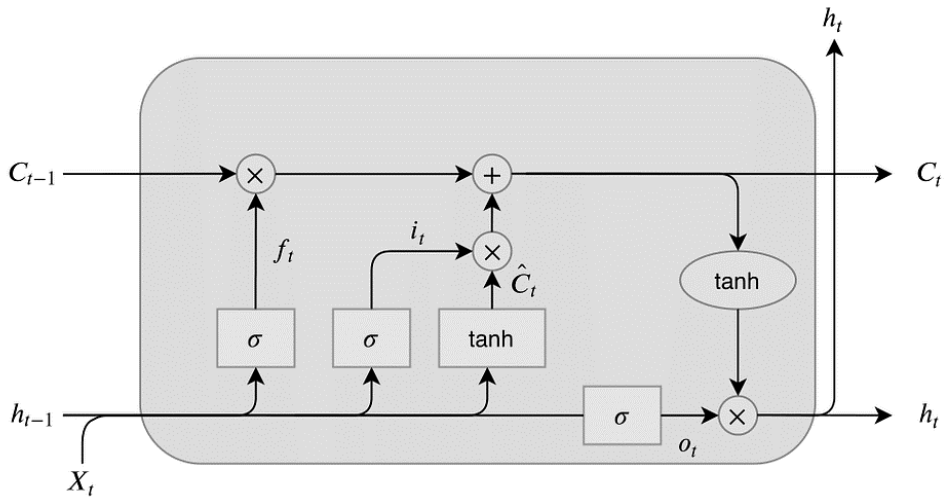


Рисунок 2.4 – Структура *LSTM*-ячейки

Для определения психологического портрета пользователя и выявления отклоняющегося поведения, используются следующие слои *LSTM*:

1. Входной слой (*Input Layer*). В этом слое происходит подача входных данных, представляющих текстовую информацию.

2. Три полносвязных слоя *LSTM* (*LSTM Layer*). Эти слои состоят из нескольких блоков *LSTM*, каждый из которых включает в себя ячейку памяти, входной вентиль, выходной вентиль и вентиль забывания. Ячейки памяти позволяют модели сохранять и запоминать информацию в течение длительных периодов, что особенно важно для анализа текстов с социальных сетей. Первый *LSTM*-слой включает 100 ячеек, которые обеспечивают начальную обработку информации. Второй *LSTM*-слой имеет 150 ячеек, что позволяет модели извлекать более сложные шаблоны и взаимосвязи в данных. Увеличение количества ячеек увеличивает выразительность модели, позволяя ей улавливать больше информации из входного текста. Третий *LSTM*-слой состоит из 200 ячеек, что позволяет дополнительно углубить анализ, учитывая более тонкие детали и контексты в текстовых данных.

3. Как говорилось ранее, при первичном обучении следом за слоями *LSTM* идет слой *Softmax*/*CRF* (слой преобразует выходные данные из *LSTM* в вероятностное распределение по различным классам).

4. Выходной слой. Этот слой генерирует выходные данные, который включает в себя классификации, соответствующие психологическим характеристикам пользователя и вероятности отклоняющегося поведения.

При дообучении НС на дополнительном смысловом домене используется слой доменно-зависимых весов *IbDA*, который является полносвязным слоем, который находится между *LSTM*-слоями и *Softmax*/*CRF*, позволяя учитывать специфику нового домена данных, улучшая адаптацию модели (рисунок 2.5).

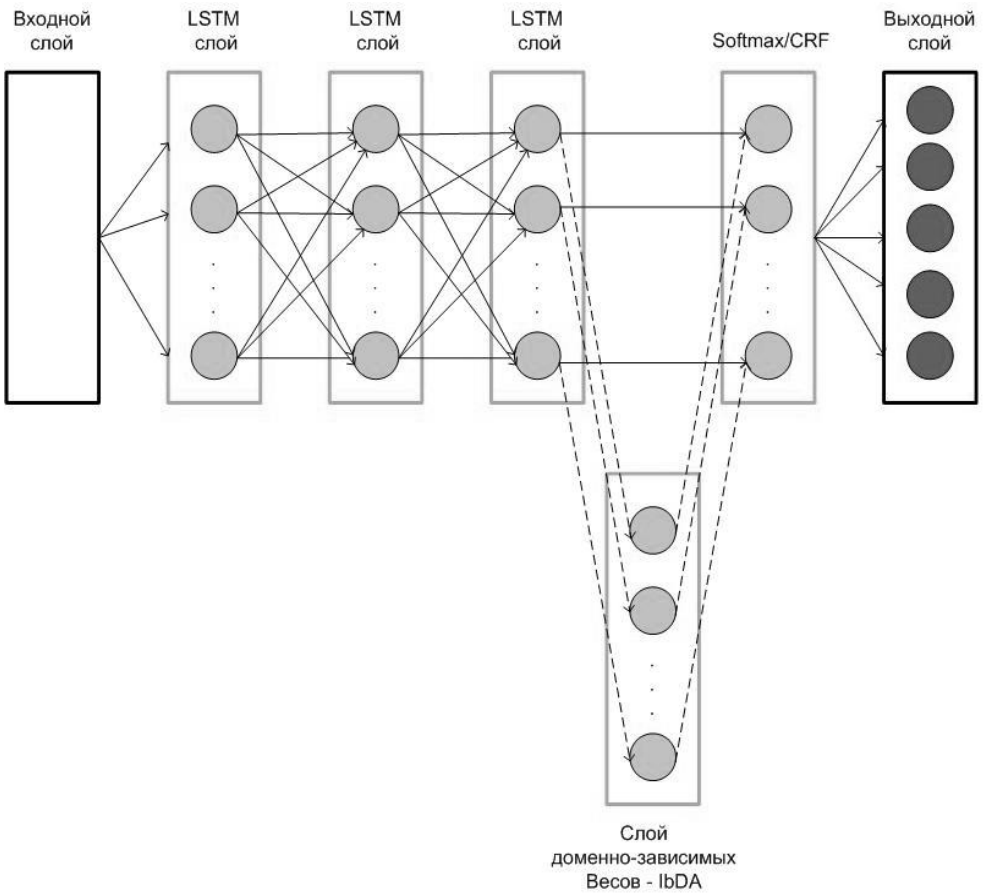


Рисунок 2.5 – Архитектура предлагаемой модели

2. Среди нескольких алгоритмов машинного обучения *Random Forest* является разнообразным, гибким, стабильным и довольно точным механизмом, который эффективно работает с большими наборами данных для получения оптимальных результатов. Этот метод может быть применен к регрессии или классификации – задачам обоих типов, разрешимым путем построения нескольких деревьев решений [113]. Затем все эти деревья решений объединяются для точного прогнозирования стабильного решения. Это очень хорошо известный алгоритм контролируемого обучения, который использует метод пакетирования, где исходные подмножества входных данных одинакового размера извлекаются с заменой, а затем на них выполняется модель; позже они собираются вместе. В деревьях принятия решений важность признака является одним из основных процессов идентификации признаков точек данных [114]. Он вычисляется путем вычисления вероятности того, что дерево достигнет узла. Чем больше вероятность, тем важнее признак.

Случайный лес выполняет две формы рандомизации для создания деревьев. Первый – начальная выборка, где для оценки параметра совокупности выполняется пакетирование. Для второй рандомизации, для построения деревьев, выбирается конечное число независимых переменных.

3. Алгоритм *KNN*, также известный как метод k-ближайших соседей, представляет собой алгоритм обучения с учителем, который широко используется для решения задач классификации и регрессии [115]. Этот алгоритм является простым и удобным для работы с обоими типами задач. В отличие от некоторых других алгоритмов, *KNN* не делает предположений о фундаментальных данных, поэтому его называют непараметризованным алгоритмом. *KNN* алгоритм сначала сохраняет обучающий набор данных, а затем выполняет свои действия. Именно по этой причине его также называют «ленивым» алгоритмом. Во время фазы обучения *KNN* алгоритм только сохраняет набор данных для будущего использования. Позднее, когда появляется новый набор данных, обучающий набор данных классифицируется в тот же самый категории, что и новый набор данных. Алгоритм *KNN* строится на концепции того, что похожие точки данных могут быть классифицированы вместе [116]. Схожесть этих точек данных чаще всего вычисляется с использованием общего евклидова расстояния, представленного в уравнении:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ...(q_n - p_n)^2}, \qquad (2.26)$$

где p, q представляют собой точки данных, $d(q,p)$ обозначает расстояние между этими двумя точками [117]. Алгоритм оценивает сходство между всеми доступными данными и классифицирует новую точку данных. Таким образом, с использованием *KNN*-алгоритма новую точку данных легко можно классифицировать в категорию, которая ей соответствует.

Выводы по главе

Во второй главе рассмотрены основные проблемы анализа тональности текстов и психологического профилирования пользователей социальных сетей

[118, 119, 120, 121]. Основные результаты исследований данной главы были описаны в соответствующих научных статьях [122, 123].

Изложены разработанные алгоритмы, методы и подходы к анализу данных из социальных сетей, предназначенные для создания психологического портрета пользователя и оценки вероятности отклоняющегося поведения.

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.

3 РАЗРАБОТКА И РЕАЛИЗАЦИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ АНАЛИЗА ПСИХОЛОГИЧЕСКОГО ПОРТРЕТА ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ

3.1 Алгоритм работы программы, предназначенной для формирования психологического портрета пользователя социальной сети

Разработанные модели и методики были реализованы в виде программы. На рисунке 3.1 представлен алгоритм работы предлагаемой системы.

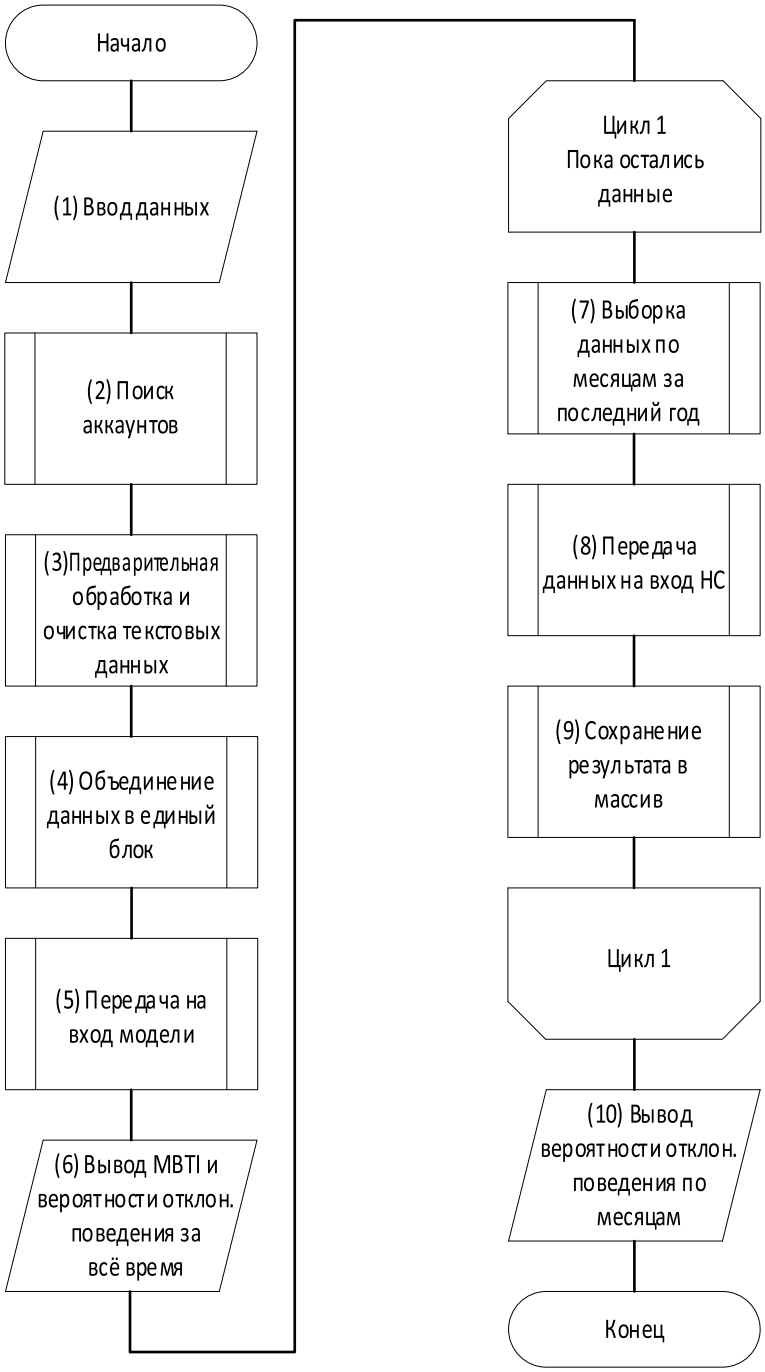
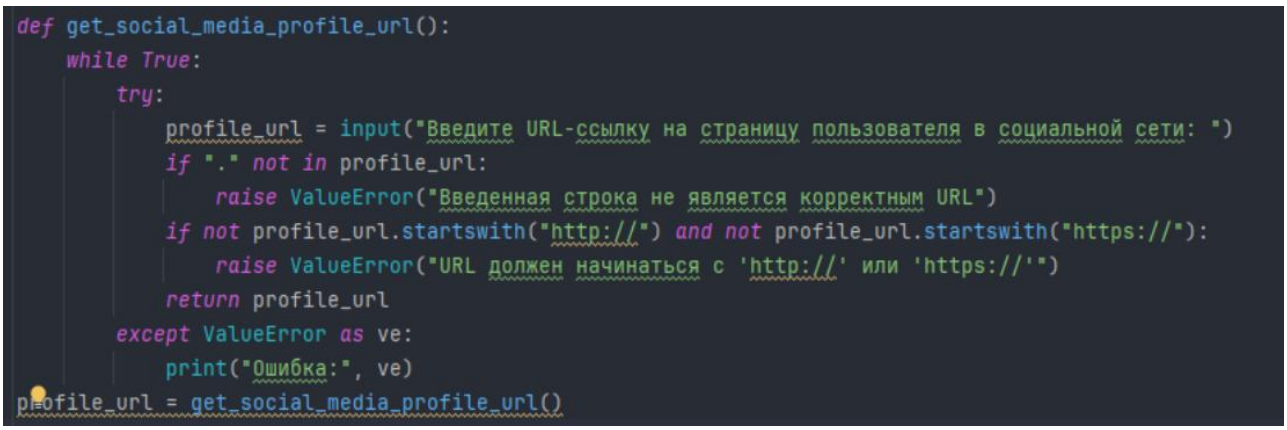


Рисунок 3.1 – Алгоритм работы предлагаемой системы

Блок 1. Ввод *URL*-ссылки на страницу искомого пользователя в социальной сети. На рисунке 3.2 показан фрагмент кода обработки ошибки для случаев некорректного ввода ссылки пользователем.



```
def get_social_media_profile_url():
    while True:
        try:
            profile_url = input("Введите URL-ссылку на страницу пользователя в социальной сети: ")
            if "." not in profile_url:
                raise ValueError("Введенная строка не является корректным URL")
            if not profile_url.startswith("http://") and not profile_url.startswith("https://"):
                raise ValueError("URL должен начинаться с 'http://' или 'https://'")
            return profile_url
        except ValueError as ve:
            print("Ошибка:", ve)
profile_url = get_social_media_profile_url()
```

Рисунок 3.2 – Фрагмент кода обработки ошибки для случаев некорректного ввода ссылки пользователем

Этот код будет повторять запрос на ввод *URL* до тех пор, пока пользователь не введет корректный *URL*.

Блок 2. Система выполняет поиск всех профилей конкретного пользователя, используя предоставленную ссылку. После этого система выполняет анализ других социальных сетей для поиска профилей того же пользователя, используя доступные данные или дополнительные параметры. Процесс поиска профилей представлен на рисунке 3.5 Фрагмент кода, реализующий поиск аккаунтов конкретного пользователя представлен на рисунке 3.3.

```python
def get_comments(post_id):  2 usages
    version = 5.131
    response = requests.get('https://api.vk.com/method/wall.getComments',
                            params={
                                'access_token': token,
                                'v': version,
                                'post_id': post_id,
                                'count': 100
                            }
                            )
    comments_data = response.json().get('response', {}).get('items', [])
    return comments_data
def take_emojis(text):  2 usages
    emojis = []
    for char in text:
        if char in emoji.UNICODE_EMOJI['en']:
            emojis.append(char)
    return emojis


def file_writer(data):  1 usage
    with open('data/victim.csv', 'w', encoding='utf-16') as file:
        a_pen = csv.writer(file)
        a_pen.writerow(('likes', 'body', 'url', 'groups', 'comments', 'emojis'))
        for post in data:
            try:
                if post['attachments'][0]['type'] == 'photo':
                    img_url = post['attachments'][0]['photo']['sizes'][-1]['url']
                else:
                    img_url = 'pass'
            except IndexError:
                img_url = 'pass'

            if post['text'] != '':
                find = GoogleTranslator(source='auto', target='en').translate(post['text'])
            else:
                find = post['text']

            groups = get_user_groups(post['from_id'])
            comments = get_comments(post['id'])
```

Рисунок 3.3 – Фрагмент кода, реализующий поиск аккаунтов конкретного пользователя

Блок 3. Предварительная обработка и очистка текстовых данных. На этом этапе выполняются такие операции, как токенизация, нормализация, исправление ошибок, лемматизация и векторизация данных. Фрагмент программного кода, иллюстрирующий процесс предварительной обработки, представлен на рисунке 3.4.

Рисунок 3.4 – Фрагмент программного кода, реализующий предварительную обработку данных

Блок 4. Данные, собранные со страницы социальных сетей, объединяются в единый блок для дальнейшего анализа.

Блок 5. Данные подаются на вход модели *IbDA-LSTM-CRF*. Работа данной модели представлена на рисунке 3.6.

Блок 6. Генерация результатов: на основе обработки данных моделью система генерирует результаты, включая психологический портрет пользователя, согласно типологии *MBTI*, и вероятность отклоняющегося поведения пользователя от 0 до 1. Следует учитывать, что для описания типа личности полезнее основываться на данных полученных за длительный период (их больше). Однако, так не видны перемены настроения за это время (в

частности, за последние месяцы) – поэтому дополнительно анализируем данные активности за последний год.

Блок 7-8. Выборка данных по месяцам за последний год. Система использует модель *LSTM* для анализа содержания и определения признаков отклоняющегося поведения. В результате анализа система получает числовую оценку вероятности отклоняющегося поведения в диапазоне от 0 до 1.

Блок 9. Сохранение результата в массив для дальнейшей обработки и агрегации.

Блок 10. Вывод результата. На основе результатов анализа система подсчитывает среднюю вероятность отклоняющегося поведения за каждый месяц в разрезе 1-го года. Полученные данные выводятся в формате от 0 до 1, где значения до 0,5 указывают на отсутствие отклонений, от 0,5 до 0,7 – на необходимость внимательного наблюдения, а значения от 0,7 до 1 – на наличие опасных признаков поведения [124].

3.2 Алгоритм поиска аккаунтов пользователя

Алгоритм поиска аккаунтов пользователя представлен на рис. 3.5.

Блок 1. Ввод *URL*-ссылки на страницу искомого пользователя в социальной сети.

Блок 2. Проведение первичного анализа сопоставления профилей по формальным признакам (ФИО, дата рождения, никнейм, статус, образование, город проживания, список друзей, подписки на сообщества, контакты и другие формальные атрибуты профиля).

Блок 3, 4, 5. Отсеивание неподходящих аккаунтов и сравнение оставшихся пользователей по различным параметрам, указанных в блоках.

Блок 4. Вычисление степени сходства между профилями, путем использования математических моделей, описанных в главе 2.

Рисунок 3.5 – Алгоритм работы поиска аккаунтов пользователя

3.3 Алгоритм работы модели *IbDA-LSTM-CRF*

Алгоритм работы модели *IbDA-LSTM-CRF* представлен на рисунке 3.6

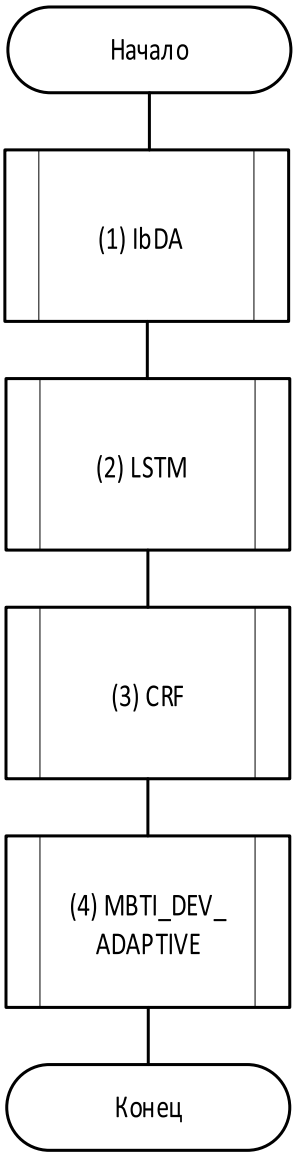


Рисунок 3.6 – Алгоритм работы модели *IbDA-LSTM-CRF*

Блок 1. В блоке *IbDA* происходит адаптация модели к текстам, относящимся к различному контексту и на разные темы, из различных доменов пользователей социальной сети. В данном случае, адаптация модели будет включать в себя преобразование представления данных из социальных сетей «Вконтакте» и «Одноклассники» (социальные сети ориентированы на разные группы населения и контекст там может быть разным, что важно учитывать). Таким образом, модель начинает лучше понимать контекст, что увеличивает точность.

Блок 2. *LSTM* извлекает важные признаки из текстов и передает их для дальнейшего анализа.

Блок 3. На данном этапе происходит моделирование зависимости между различными аспектами и тональностями в тексте. Учитываются контекстуальные зависимости между различными элементами текста, такими как выделенные аспекты и их тональность.

Блок 4. Предобученный слой *MBTI_DEV adaptive* используется для преобразования выхода *CRF* модели к информации, соответствующей системе *MBTI* и вероятности отклоняющегося поведения. Позволяет оценить тип личности пользователя в соответствии с системой *MBTI* и определить вероятность отклоняющегося поведения на основе анализа текстовых данных.

3.4 Особенности программной реализации

Модули программы реализованы на языке программирования *Python*, выбор которого обусловлен рядом причин:

1. *Python* известен своей простотой и легкостью в изучении, что делает его отличным выбором для быстрой разработки программного обеспечения.

2. *Python* имеет огромное сообщество разработчиков, что означает наличие множества библиотек и инструментов для обработки текстовых данных, машинного обучения, веб-скрапинга и других задач, связанных с анализом данных из социальных сетей.

3. *Python* обладает множеством библиотек для обработки текстовых данных, таких как *NLTK* (*Natural Language Toolkit*), *SpaCy*, *Gensim* и другие, которые облегчают анализ текста и извлечение ключевой информации.

4. *Python* является популярным выбором для разработки моделей машинного обучения благодаря таким библиотекам, как *scikit-learn*, *TensorFlow*, *PyTorch* и другие, что позволяет создавать и обучать модели для выявления психологических характеристик пользователей на основе их текстовых данных.

5. *Python* также широко используется для веб-скрапинга данных из социальных сетей или других онлайн-ресурсов, что может быть полезным для сбора данных для анализа.

В разработанной и зарегистрированной в процессе написания диссертации программе применяются следующие библиотеки: *NLTK*, содержащий широкий спектр инструментов для обработки и анализа естественного языка, включающий в себя функции токенизации, лемматизации, стемминга, разметки частей речи, анализа сентимента; *SpaCy*, предоставляющий быструю и эффективную обработку текста с помощью предварительно обученных моделей; *NumPy*, реализующая операции при работе с многомерными массивами; *Pandas*, предназначенная для манипуляции с табличными данными; *TensorFlow* и *Keras*, предоставляющие широкие возможности для построения и обучения нейронных сетей; *matplotlib* для визуализации данных, *scikitlearn* – включает методы классификации, регрессии и кластеризации, предоставляет функции для нормализации, позволяет проводить кросс-валидацию.

3.4.1 Реализация программного обеспечения

Разработанное программное обеспечение позволяет в режиме реального времени анализировать информацию, размещаемую пользователями в социальной сети, с целью формирования их психологического портрета и выявления вероятности отклоняющегося поведения. Это программное обеспечение предоставляет возможность эффективного и автоматизированного анализа Больших объемов данных, учитывая разнообразные факторы, влияющие на пользовательское поведение.

Программное обеспечение состоит из двух ключевых компонентов: фронтенд и бэкенд. Фронтенд создан с использованием JavaScript и фреймворка *React.js* [125], а бэкенд реализован с помощью *Api*–фреймворка *DRF* (*Django Rest Framework*) [126].

Фронтенд на *React.js*:

- Основан на *JavaScript*-фреймворке *React.js*.;

- Использует библиотеку *Axios* для отправки данных на сервер;
- Получает данные от пользователя через интерфейс;
- Обновляет интерфейс в соответствии с данными, полученными от бэкенда.

Бэкенд на *Django Rest Framework* (*DRF*):

- Основан на *DRF* для создания *API*;
- Обрабатывает *HTTP*-запросы от фронтенда;
- Выполняет необходимые операции: анализ данных и доступ к данным;
- Возвращает ответы в формате *JSON*.

Обмен данными происходит следующим образом:

- Используются стандартные форматы данных, такие как *JSON*, для обмена данными между фронтендом и бэкендом;
- Фронтенд отправляет данные на бэкенд через *HTTP*-протокол;
- Бэкенд обрабатывает запросы и возвращает ответы в формате *JSON*.

Основная логика работы: пользователь взаимодействует с фронтендом, вводя данные в интерфейс. Фронтенд отправляет данные на бэкенд. Бэкенд обрабатывает запросы, выполняет операции и возвращает результаты. Фронтенд обновляет интерфейс на основе данных, полученных от бэкенда.

Пользовательский интерфейс представлен на рисунке 3.7.

Рисунок 3.7 – Пользовательский интерфейс программы

Как видно из рисунка пользователю предоставляется поле ввода, где он может ввести *URL*-ссылку на страницу искомого пользователя в любой из

поддерживаемых социальных сетей, таких как «ВКонтакте» или «Одноклассники». Пользователь может ввести ссылку на страницу в одной из социальных сетей или в обеих сразу.

После ввода *URL*-ссылки и отправки запроса, система запускает процесс поиска профилей пользователя в обеих указанных социальных сетях. Результаты поиска профилей отображены на экранной форме, представленной на рисунке 3.8.

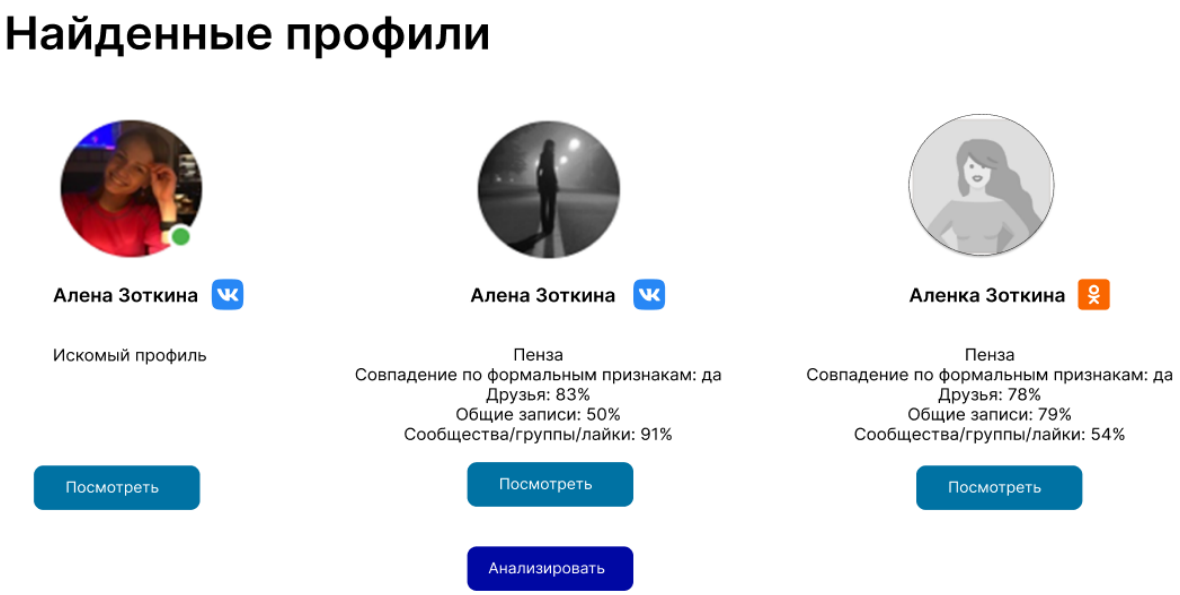


Рисунок 3.8 – Отображение всех найденных профилей в социальных сетях «Вконтакте» и «Одноклассники»

Каждый аккаунт представлен в виде карточки с основной информацией о пользователе, включая:

- Фотография профиля;
- Имя пользователя;
- Город;
- Совпадение по формальным признакам;
- Процент совпадения друзей: например, «80%» указывает на то, что пользователь имеет 80% общих друзей с искомым пользователем.
- Процент общих записей: например, «50%» указывает на то, что пользователь имеет 50% общих записей с искомым пользователем;
- Процент совпадения сообществ/групп/лайк.

Кроме того, форма содержит кнопку "Просмотреть все аккаунты", которая позволяет пользователю просматривать все найденные аккаунты пользователя в различных социальных сетях.

Далее, нажав кнопку «Анализировать», выводятся результаты анализа выводятся в новом диалоговом окне программы, где пользователь видит определенный тип *MBTI* и интерпретацию этого типа, а также вероятность отклоняющегося поведения за все время и за частичный период – год. Результат представлен на рисунке 3.9:

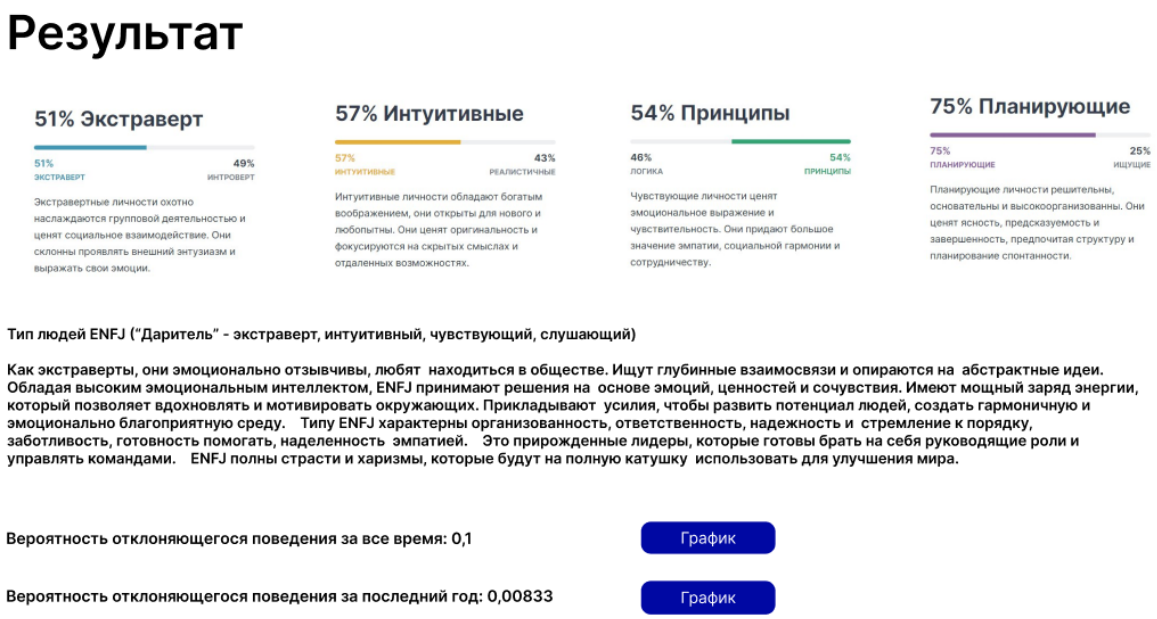


Рисунок 3.9 – Результат анализа

При анализе типа личности и отклоняющегося поведения пользователя более длинные временные периоды предоставляют больше данных для более точной оценки его поведения. Информация, собранная за длительный период времени, может помочь выявить общие тенденции и характеристики личности пользователя. На рисунке 3.10 представлен график вероятности отклоняющегося поведения за все время, с начала даты регистрации пользователя по текущий год.

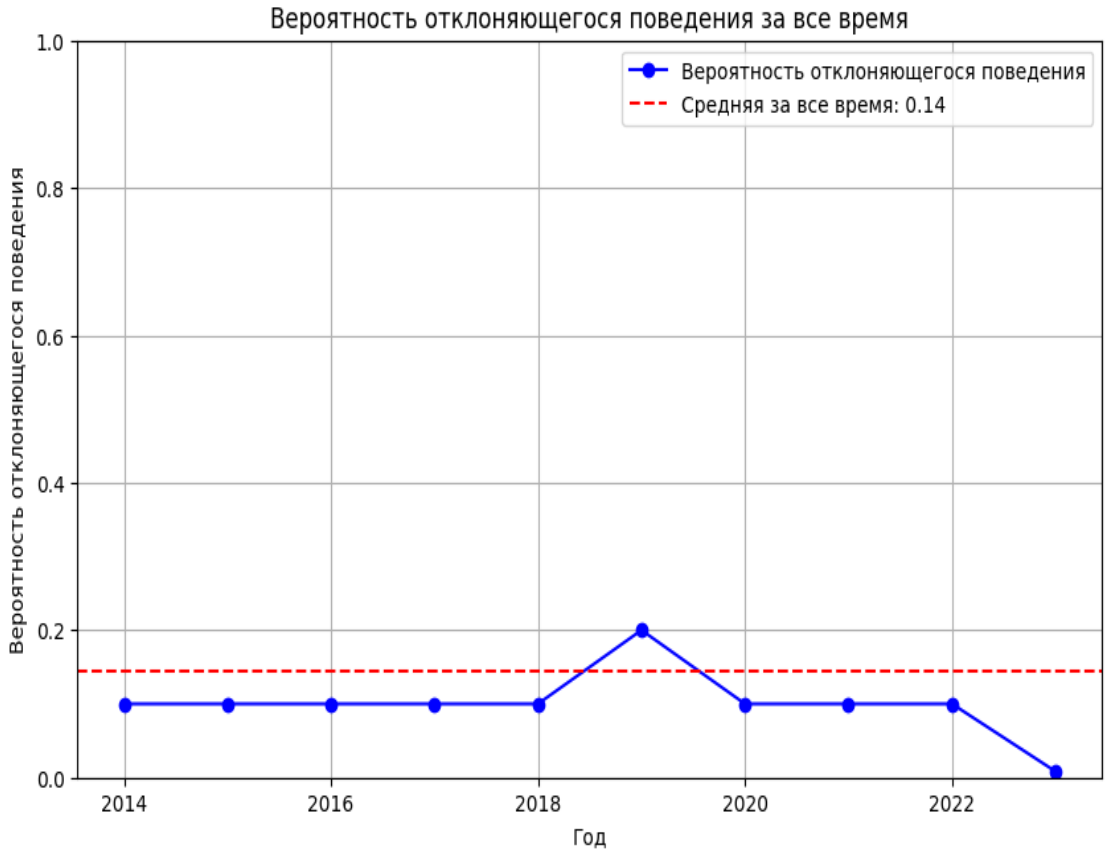


Рисунок 3.10 – График, показывающий вероятность отклоняющегося поведения за все время

Однако, при анализе длительных временных периодов, таких как несколько лет, не всегда ясно видны изменения в настроении или поведении пользователя в последние месяцы. Это может быть связано с тем, что влияние недавних событий или изменений в жизни пользователя не отражается на общей картине за длительный период. Для оценки текущего состояния пользователя и выявления возможных изменений в его поведении за последние месяцы можно построить график вероятности отклоняющегося поведения за последний год. Этот график позволяет наглядно отслеживать изменения в поведении пользователя в течение последних 12 месяцев (рисунок 3.11).

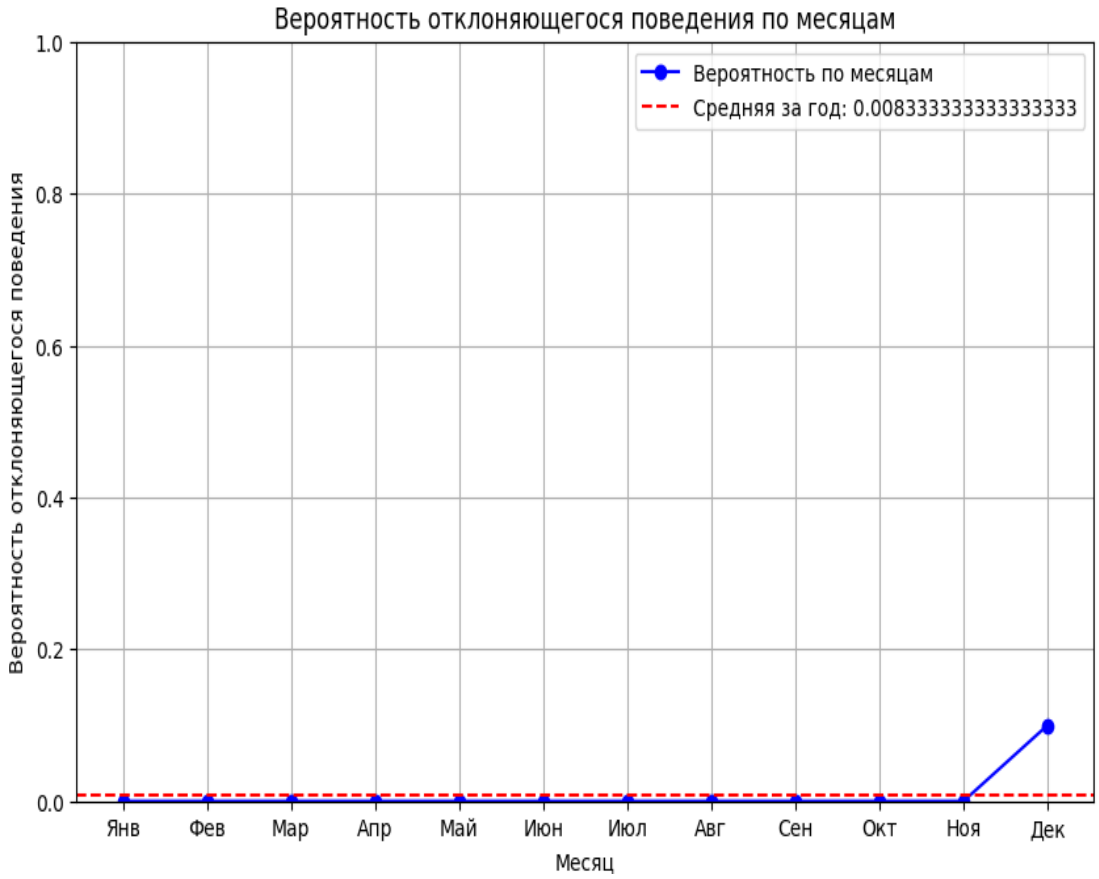


Рисунок 3.11 – График, показывающий вероятность отклоняющегося поведения за последний год

Выводы по главе

В третьей главе представлены алгоритмы программной реализации обработки информации, размещаемой пользователями в социальной сети, с целью формирования их психологического портрета. Эти алгоритмы разработаны с учетом основных принципов анализа данных, а также учитывают специфику социальных сетей и их влияние на поведение и ментальное состояние пользователей.

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.

4 ПРОВЕДЕНИЕ ЭКСПЕРИМЕНТАЛЬНОГО ТЕСТИРОВАНИЯ АЛГОРИТМОВ

4.1 Подготовка данных для эксперимента

Предложенные методы обработки информации, размещаемой пользователями в социальной сети, в задачах идентификации факторов риска безопасности с использованием социального портрета пользователя необходимо проверить на точность идентификации и сравнить с существующими аналогичными алгоритмами.

Для экспериментальной верификации предложенного метода необходимо осуществить всестороннее тестирование его отдельных компонентов: проверка одного пользователя на наличие нескольких аккаунтов как в пределах одной социальной сети, так и в нескольких, определение психологического портрета пользователя и оценка вероятности отклоняющегося поведения.

Для обеспечения прозрачности экспериментов были использованы общедоступные базы данных, содержащие тексты, которые широко применяются для тестирования методов определения психологического портрета пользователя.

Для проведения экспериментов применяются следующие базы данных:

- *MBTI Dataset* – набор данных, содержащий более 8000 текстовых постов с ресурса *Reddit*, размеченные согласно типам личности *MBTI* [127];
- *Personality Prediction* – включает в себя 10000 записей. Эти записи состоят из текстовых данных пользователей и их соответствующих типов личности, что позволяет проводить анализ текста для предсказания личностных черт [128];
- *MBTI Personality Types* 500 *Dataset* – набор данных, содержащий около 10000 записей, предварительно обработанных постов и типов личности их авторов. Сообщения имеют одинаковый размер: 500 слов в выборке [129];
- Данные, содержащие эмоционально-окрашенные сообщения, собранные самостоятельно из сообществ социальных сетей, описанных в главе 2.5. Содержат более 6700 записей.

4.2 Эксперименты по идентификации профилей пользователей на различных платформах социальных сетей

В рамках эксперимента создана предварительно подготовленная выборка из 153 пользователей, каждый из которых имел аккаунты в различных социальных сетях. В итоге общее количество аккаунтов, принадлежащих этим пользователям, составило 261. При этом следует учитывать, что не все участники эксперимента обладали несколькими аккаунтами на разных платформах; некоторые пользователи имели только один профиль в определенной социальной сети. Это разнообразие аккаунтов обеспечивало более широкий спектр данных для анализа. Диаграмма распределения пользователей по количеству аккаунтов в разных социальных сетях представлена на рисунке 4.1.

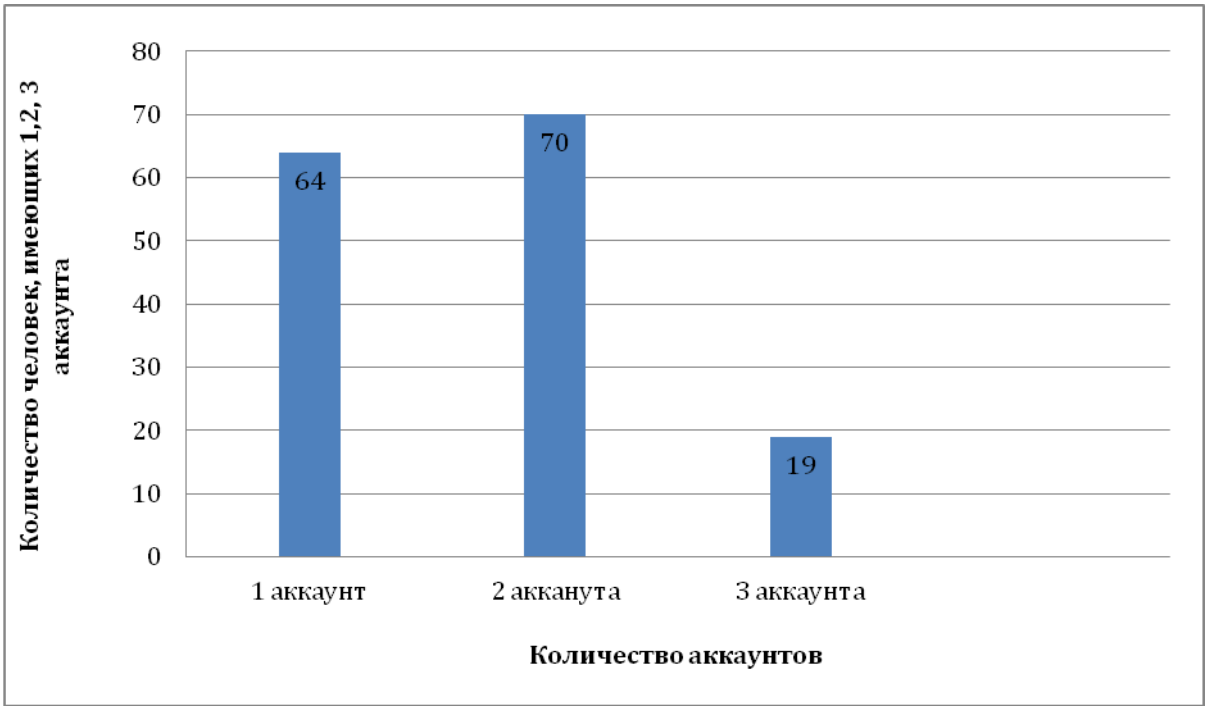


Рисунок 4.1 – Диаграмма распределения пользователей по количеству аккаунтов

Детальное распределение количества одинаковых аккаунтов пользователей представлено в таблице 4.1. Например, число 28 в столбце 4 означает, что 14 человек имеют по два аккаунта в одноклассниках.

Таблица 4.1 – Детальное распределение количества одинаковых аккаунтов пользователей

Число аккаунтов	1 аккаунт	2 аккаунта			3 аккаунта	
		Оба в ВК.	Оба в Од.	1 ВК.+1 Од.	2 ВК.+1 Од.	2 Од.+1 ВК.
1	2	3	4	5	6	7
Одноклассники	26	–	28	40	9	20
ВКонтакте	38	32	–	40	18	10
Всего	64	32	28	80	27	30
Общее количество аккаунтов	**261**					

Процентное соотношение корректно найденных аккаунтов можно увидеть на рисунке 4.2. Следует учесть, что в сюда входят как одиночные аккаунты, так и пользователи с несколькими аккаунтами.

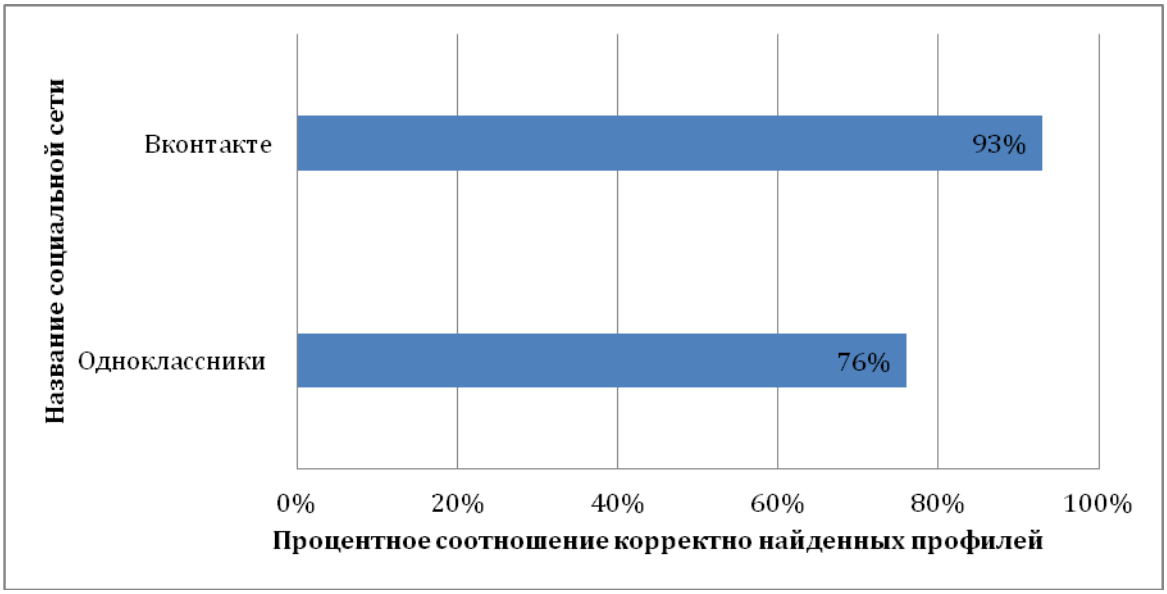


Рисунок 4.2 – Диаграмма процентного соотношения корректно найденных профилей

Исходя из полученных данных, рисунок 4.2 подтверждает не только популярность «Вконтакте» по сравнению с «Одноклассниками», но и удобство использования *VK_API* для извлечения данных. «ВКонтакте», имея более молодую аудиторию и более удобные инструменты для работы с данными, может обеспечивать более высокий процент успешного нахождения профилей

пользователей. 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. Также были рассчитаны результаты для критериев сравнения профилей в рамках одной социальной сети. Эти данные представлены на рисунке 4.3.

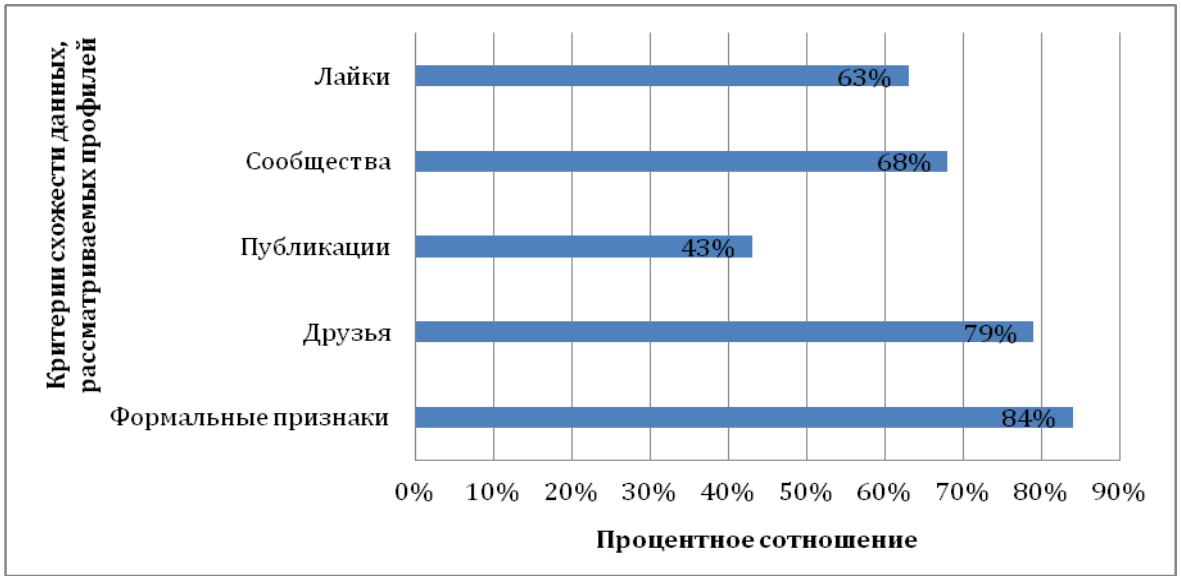


Рисунок 4.3 – Критерии сравнения профилей в одной социальной сети

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 4.4), 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.

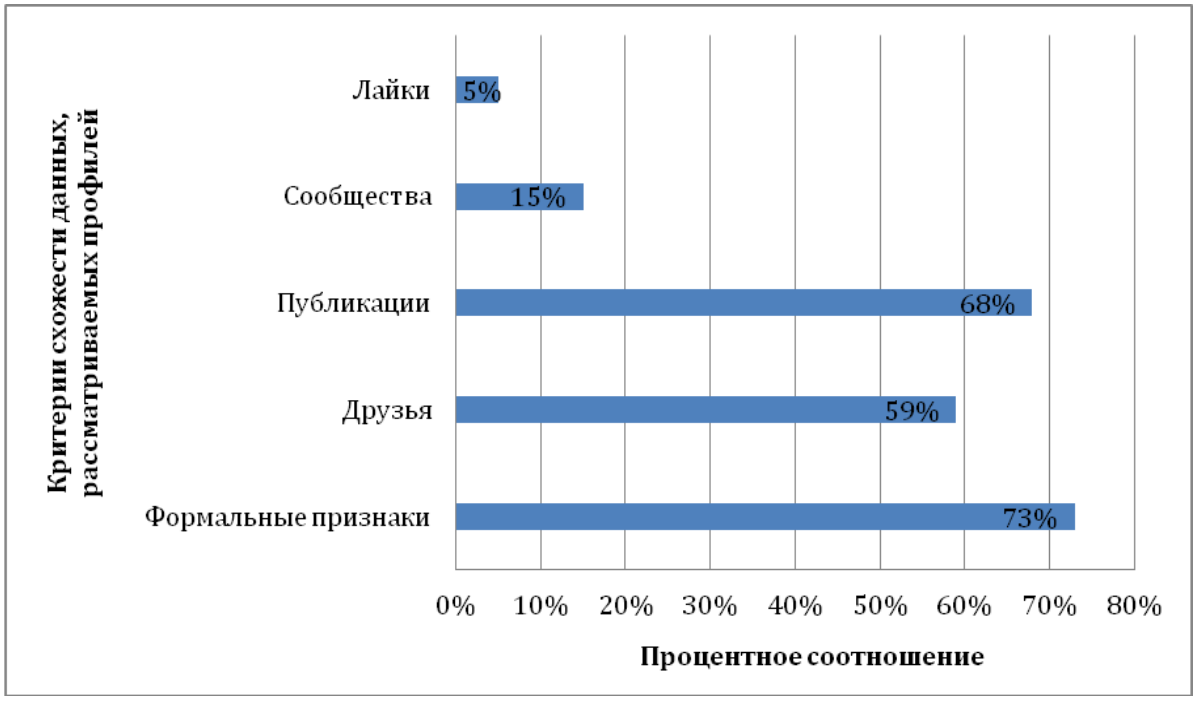


Рисунок 4.4 – Критерии сравнения профилей в двух социальных сетях

Таким образом, анализ и объединение информации из нескольких социальных сетей может значительно улучшить точность и скорость составления психологического портрета пользователя.

4.3 Эксперименты алгоритмов классификации

Обучение нейронной сети для определения психологического портрета и вероятности отклоняющегося поведения пользователя происходило в течение 100 эпох. В каждой эпохе использовался набор данных, состоящий из 29000 записей из социальных сетей, включая публикации, сообщества, комментарии, лайки и другие активности пользователей. Общая выборка была разделена на обучающую и тестовую в соотношении 70/30.

Для контроля обучаемости нейронной сети применялись функции потерь и метрики, оценивающие сходство между истинными и предсказанными

данными. В данной задаче использовалась комбинация двух функций потерь: кросс-энтропия [130] и коэффициент Дайса [131].

Пусть y – истинная метка (0 или 1 для двоичной классификации), а y' – это предсказанная моделью вероятность принадлежности к классу 1. Тогда формула кросс-энтропии будет выглядеть следующим образом:

$$CE(y, y') = -y \times \ln(y') - (1 - y) \times \ln(1 - y') \quad (4.1)$$

Коэффициент Дайса также применяется для оценки сходства между двумя множествами данных, в данном случае текстовыми данными. Он помогает модели учитывать не только конкретные слова или фразы, но и их контекст и структуру в текстах из социальных сетей.

Пусть A – множество элементов, которые предсказаны как принадлежащие к классу 1, а B – множество истинных элементов, принадлежащих к классу 1. Тогда формула коэффициента Дайса выглядит следующим образом:

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|}, \quad (4.2)$$

где $|A \cap B|$ – количество элементов, которые присутствуют как в множестве A, так и в множестве B, а $|A|$ и $|B|$ – количество элементов в множествах A и B, соответственно.

Статистика обучения нейронной сети *LSTM* представлена в таблице 4.2.

Таблица 4.2 – Статистика обучения нейронной сети *LSTM*.

№ эпохи	Кросс–энтропия (обучение)	Коэффициент Дайса (обучение)	Кросс-энтропия (тестовый)	Коэффициент Дайса (тестовый)
1	**2**	**3**	**4**	**5**
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

1	2	3	4	5
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
11	0,442	0,889	0,458	0,88
12	0,418	0,89	0,446	0,881
13	0,419	0,95	0,435	0,889
14	0,409	0,963	0,425	0,93
15	0,399	0,972	0,416	0,956
16	0,39	0,968	0,417	0,957
17	0,382	0,968	0,408	0,957
18	0,374	0,968	0,391	0,957
19	0,367	0,992	0,384	0,973
20	0,368	0,995	0,392	0,975
21	0,369	0,982	0,394	0,963
22	0,347	0,985	0,373	0,965
23	0,348	0,975	0,375	0,952
24	0,335	0,972	0,387	0,95
25	0,33	0,971	0,35	0,953
26	0,324	0,975	0,342	0,955
27	0,319	0,978	0,337	0,958
28	0,32	0,98	0,338	0,96
29	0,31	0,981	0,332	0,961
30	0,311	0,982	0,325	0,962
31	0,301	0,983	0,326	0,963

1	2	3	4	5
32	0,297	0,985	0,319	0,965
33	0,31	0,981	0,345	0,961
34	0,289	0,9573	0,307	0,923
35	0,286	0,965	0,312	0,93
36	0,287	0,965	0,307	0,945
37	0,279	0,972	0,298	0,953
38	0,283	0,98	0,3	0,954
39	0,3	0,975	0,291	0,955
40	0,27	0,985	0,288	0,963
41	0,253	0,986	0,285	0,964
42	0,264	0,99	0,289	0,967
43	0,268	0,99	0,28	0,968
44	0,259	0,992	0,277	0,969
45	0,257	0,989	0,275	0,97
46	0,254	0,992	0,272	0,971
47	0,255	0,993	0,27	0,972
48	0,253	0,995	0,272	0,95
49	0,254	0,985	0,268	0,96
50	0,246	0,985	0,264	0,961
51	0,244	0,982	0,262	0,962
52	0,243	0,986	0,265	0,965
53	0,241	0,988	0,26	0,962
54	0,24	0,983	0,263	0,963
55	0,239	0,984	0,257	0,964
56	0,237	0,995	0,256	0,97
57	0,239	0,992	0,27	0,971
58	0,235	0,993	0,254	0,972
59	0,234	0,994	0,253	0,973

1	2	3	4	5
60	0,236	0,996	0,252	0,974
61	0,231	0,995	0,254	0,971
62	0,23	0,992	0,25	0,972
63	0,24	0,991	0,253	0,977
64	0,228	0,989	0,248	0,979
65	0,227	0,995	0,247	0,98
66	0,226	0,996	0,253	0,976
67	0,225	0,995	0,245	0,991
68	0,224	0,995	0,255	0,991
69	0,223	0,995	0,249	0,991
70	0,222	0,996	0,249	0,992
71	0,221	0,998	0,25	0,993
72	0,22	0,998	0,241	0,995
73	0,219	0,998	0,243	0,99
74	0,218	0,992	0,239	0,989
75	0,218	0,995	0,238	0,985
76	0,217	0,991	0,239	0,984
77	0,225	0,992	0,237	0,982
78	0,215	0,995	0,236	0,985
79	0,214	0,997	0,235	0,987
80	0,218	0,995	0,241	0,988
81	0,213	0,996	0,234	0,99
82	0,212	0,997	0,236	0,991
83	0,211	0,998	0,237	0,992
84	0,22	0,998	0,232	0,99
85	0,21	0,998	0,231	0,99
86	0,209	0,998	0,232	0,989
87	0,208	0,998	0,24	0,989

1	2	3	4	5
88	0,207	0,998	0,235	0,988
89	0,21	0,995	0,229	0,991
90	0,206	0,995	0,23	0,99
91	0,207	0,994	0,24	0,985
92	0,205	0,996	0,228	0,985
93	0,204	0,997	0,23	0,985
94	0,203	0,9981	0,228	0,987
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

Рисунки 4.5 и 4.6 демонстрируют ключевые результаты обучения нейронной сети.

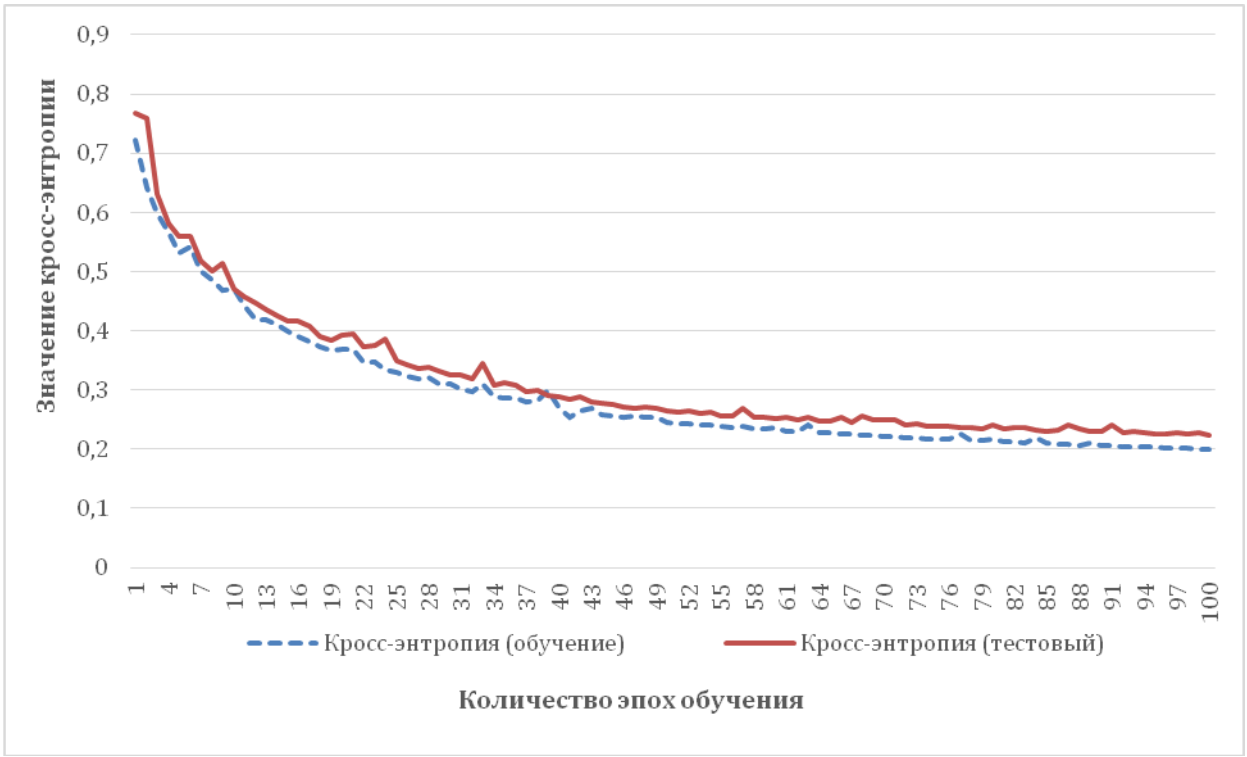


Рисунок 4.5 – График изменения функции потерь при обучении

Из графика видно, что в ходе обучения кросс-энтропия, предназначенная для оценки разницы между распределениями вероятностей истинных и предсказанных классов, постепенно уменьшается пока не доходит до некоторого постоянного значения, при этом разница составляет 12%, что является приемлемым значением. Эти данные указывают на то, что модель хорошо обучается и не слишком сильно изменяет распределение вероятностей между истинными и предсказанными классами.

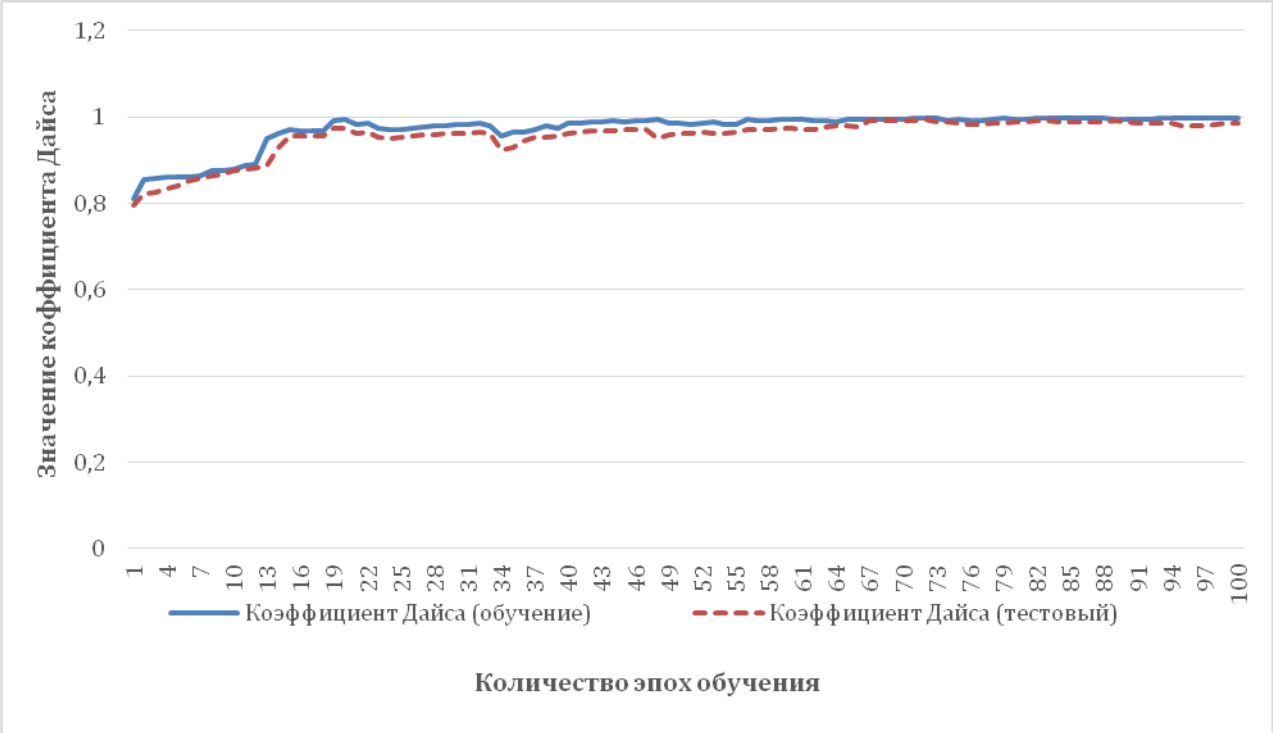


Рисунок 4.6 – График изменения функции сходства в процессе обучения

Аналогичным образом ведет себя коэффициент Дайса, который оценивает сходства между двумя множествами текстовых данных, учитывая их контекст и структуру. Разница между тестовой и обучающей выборкой незначительная и составляет 1%.

В ходе проведенных экспериментов было установлено, что выбор конкретного набора данных оказывает значительное влияние на конечные результаты анализа. Это свидетельствует о важности тщательного подбора данных для достижения высококачественных результатов. Одним из

перспективных направлений для улучшения работы данной сети является объединение различных наборов данных. Это позволит увеличить вариативность образцов и более точно имитировать реальные данные, что, в свою очередь, может способствовать повышению точности и надежности предсказаний модели. Комбинирование данных из разных источников также может помочь учесть разнообразие пользовательского поведения и характеристик, что важно для более глубокого анализа и понимания психологического портрета пользователей.

4.4 Метрики оценки результатов классификации

В рамках эксперимента использована предварительно подготовленная выборка из 153 пользователей, каждый из которых имел аккаунты в различных социальных сетях. Общее количество аккаунтов у этих пользователей составляло 261.

В ходе экспериментов оценивалась эффективность и точность алгоритмов классификации текстов для определения психологического портрета пользователя социальной сети. Параметры оценки качества работы классификатора представлены в таблице 4.3.

Таблица 4.3 – Оценка качества работы классификатора

Документ		Оценка эксперта	
		Positive	*Negative*
Оценка системы	*Positive*	*TP (True Positive)*	*FP (False Positive)*
	Negative	*FN (False Negative)*	*TN (True Negative)*

В таблице приняты следующие условные обозначения:

- *TP* (*True Positive*) – истинно-положительные, то есть классифицированные и системой, и экспертом как положительные.

- *FP* (*False Positive*) – ложно-положительные, то есть классифицированные экспертом как отрицательные, а системой как положительные.
- *FN* (*False Negative*) – ложно-отрицательные, то есть классифицированные экспертом как положительные, а системой, как отрицательные.
- *TN* (*True Negative*) – истинно-отрицательные, то есть классифицированные и системой, и экспертом как отрицательные.

На основе введенных понятия рассчитываются показатели точности и полноты.

$$precision = \frac{TP}{TP + FP} \tag{4.3}$$

$$recall = \frac{TP}{TP + FN} \tag{4.4}$$

Очевидно, что чем выше показатели точности и полноты, тем качественнее результат классификации. Однако, в реальной жизни максимальные значения точности и полноты практически недостижимы, поэтому приходится искать компромисс. *F*-мера представляет собой средне между показателями плотности и точности.

$$F = (\beta^2 + 1)\frac{precision \times recall}{\beta^2 precision + recall}, \tag{4.5}$$

где β – коэффициент, который отвечает за вес, отдаваемый точности в полноте. При $0 < \beta < 1$ приоритет будет отдан показателю точности, при $\beta > 1$ – полноте. При $\beta = 1$ показателю полноты и точности будут иметь одинаковые веса.

$$F = 2\frac{precision \times recall}{precision + recall}, \tag{4.6}$$

F–мера будет использоваться для метрики с $\beta = 1$.

Для сравнения были следующие классификаторы: *LSTM, KNN, Random Forest*. Выборка была разделена на обучающую и тестовую в пропорции 70/30.

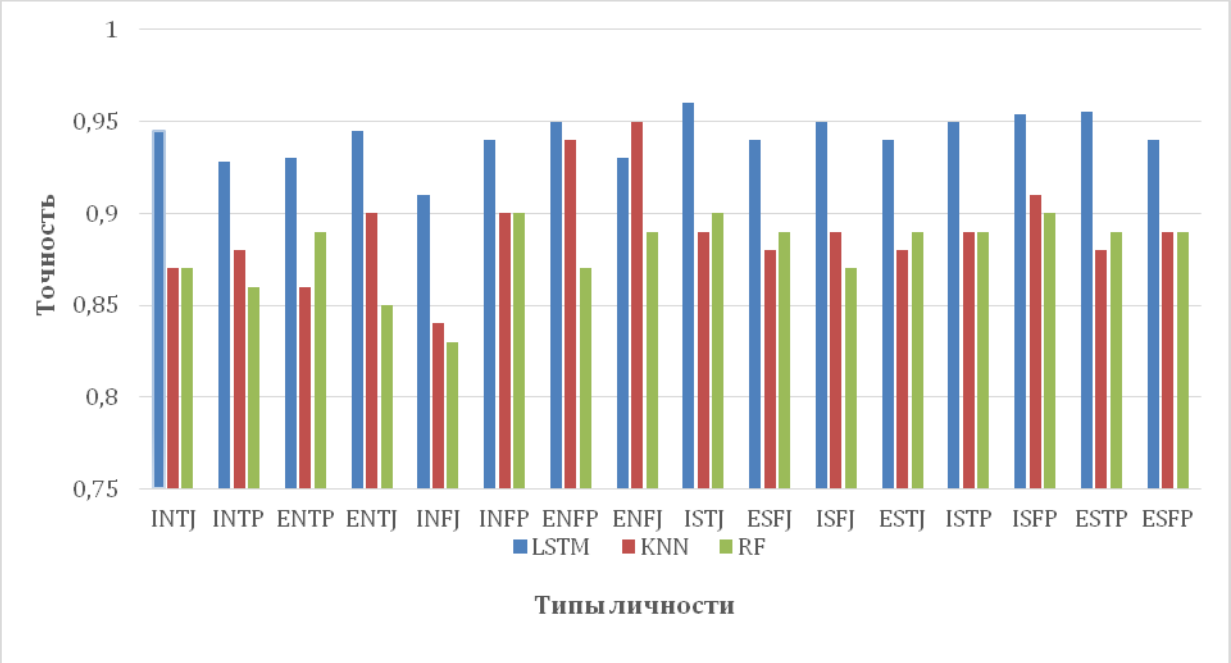


Рисунок 4.7 – Результаты экспериментов

LSTM показывает наилучшие результаты точности *INTJ*, *ENTJ*, *INFP*, *ENFP*, *ISTJ*, *ISFJ*, *ISTP*, *ISFP*, *ESTP*, *ESFP*. Это указывает на то, что *LSTM* справляется лучше всего с классификацией этих типов личности (рисунок 4.7).

Для следующих типов личности, таких как *INTP*, *INFJ*, *ESFJ*, *ESTJ*, результаты классификации *LSTM* также выше, хотя не настолько 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.

Результаты экспериментов сравнения различных методов классификации представлены в таблице 4.4.

Таблица 4.4 – 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. В данном случае *F*-мера для *LSTM* составляет 0,945, что означает баланс между точностью и полнотой, превосходя показатели *KNN* и *Random Forest*.

Предлагаемая модель *IbDA-LSTM-CRF* позволяет лучше понимать контекст данных из различных социальных сетей и более точно распознавать тексты, соответствующие различным типам личности согласно *MBTI*, и проводить оценку вероятности отклоняющегося поведения. Комбинирование этого метода позволяет учитывать не только сами слова, но и их контекст, а также взаимосвязи между различными элементами текста. Это приводит к более высокой точности классификации и более глубокому пониманию текстовых данных, что особенно важно при анализе психологических характеристик и поведения пользователей в социальных сетях [132,133].

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 определен экспертом.

Выводы по главе

В четвертой главе изложены результаты экспериментов, проведенных с применением предложенных алгоритмов, реализованных в виде программного комплекса. Исследование включает анализ различных методов поиска аккаунтов одного и того же человека как в различных социальных сетях, так и на одной платформе.

В процессе экспериментов были протестированы различные архитектуры нейронных сетей, такие как *LSTM (Long Short-Term Memory), KNN (k-Nearest Neighbors)* и *Random Forest*. Все они использовали модель *IbDA-LSTM-CRF*, что позволило получить более точные результаты. Наилучшие показатели продемонстрировала архитектура *LSTM*, достигнув точности 0,95. Сравнительный анализ показал, что *LSTM* превосходит другие методы, такие как *KNN* и *Random Forest*, в плане точности и надежности. Это может быть объяснено тем, что *LSTM* лучше справляется с контекстной информацией и временными зависимостями, что критически важно при анализе данных из социальных сетей, где информация может изменяться со временем. Применение кросс-энтропии и коэффициента Дайса для оценки сходства между истинными и предсказанными значениями подтвердило надежность модели. Использование комбинированных функций потерь привело к более точным и стабильным результатам.

Эти результаты подчеркивают потенциал нейронных сетей для анализа и интерпретации данных социальных сетей, а также их применимость в решении задач, связанных с формированием психологического портрета пользователя.

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 ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ И ПЕРВОИСТОЧНИКОВ

1. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живѐм, работаем и мыслим = Big Data. A Revolution That Will Transform How We Live, Work, and Think / пер. с англ. Инны Гайдюк. – М.: Манн, Иванов, Фербер, 2014. – 240 с.

2. Силен Д. Основы Data Science и Big Data. Python и наука о данных / Д. Силен, А. Мейсман, М. Али. – СПб.: Издательский дом «Питер», 2017. – 336 с.

3. Eileen McNulty. Understanding Big Data: The Seven V’s // Dataconomy. – [Электронный ресурс] – http://dataconomy.com/2014/05/seven-vs-big-data/. – Режим доступа-свободный. (дата обращения 14.01.2020).

4. V Kalyani, Big Data and Social Science Data Science Methods and Tools for Research and Practice, Journal of the Royal Statistical Society Series A: Statistics in Society, Volume 187, Issue 2, April 2024, Pages 542–543, https://doi.org/10.1093/jrsssa/qnad109.

5. Идигова Л. М. Data Science как новый тренд. Исследование методов работы с большим объемом данных в организации / Л. М. Идигова, А. Х. Абубакаров // Влияние новой геополитической реальности на государственное управление и развитие Российской Федерации: материалы II Всероссийской научно-практической конференции, Грозный, 20-21 сентября 2019 г. – Грозный: Чеченский государственный университет, 2019. − 275−280.

6. Губанов Д. А. Социальные сети: модели информационного влияния, управления и противоборства / Д. А. Губанов, Д. А. Новиков, А. Г. Чхартишвили. – М.: Изд-во физико-математической литературы, 2010. – 228 с.

7. Когда и как родились «Большие данные» – краткая история. – Текст: электронный // Цифровой инжиниринг ВИШ МИФИ: [сайт]. – URL: https://dzen.ru/a/YGoNRrIHhgN5y8Ab (дата обращения: 21.01.2022).

8. Big Data = Большие данные: учеб. пособие / И. Б. Тесленко [и др.]; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир: Изд-во ВлГУ, 2021. – 123 с.

9. Герман Холлерит – изобретатель первой электрической вычислительной машины – Текст: электронный // VXI - информационно-измерительные технологии: [сайт]. – URL: http://www.vxi.ru/history/german-hollerit/ (дата обращения: 21.01.2022).

10. С немецким акцентом: краткая история создания магнитной ленты— Текст: электронный // Хабр: [сайт]. – URL: https://habr.com/ru/companies/onlinepatent/articles/792356/ (дата обращения: 21.01.2022).

11. ЭВМ: ЧТО? ГДЕ? КОГДА? – Текст: электронный // ЭВМ history: [сайт]. – URL: https://evmhistory.ru/history/colossus.html (дата обращения: 21.01.2022).

12. История электронных компьютеров, часть 4: электронная революция – Текст: электронный // Хабр: [сайт]. – URL: https://habr.com/ru/articles/447916/ (дата обращения: 21.01.2022).

13. Базенков Н. И. Обзор информационных систем анализа социальных сетей / Н. И. Базенков, Д. А. Губанов. // Управление большими системами. – 2013. – 41. – С. 357-394.

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

15. Domo – Текст: электронный // Morning Dough: [сайт]. – URL: https://www.morningdough.com/ru/ai-tools/domo/ (дата обращения: 21.01.2022).

16. Управление большими данными с помощью приложений On-demand— Текст: электронный // Qlik Help: [сайт]. – URL: https://help.qlik.com/ru-RU/sense/February2024/Subsystems/Hub/Content/Sense_Hub/DataSource/Manage-big-data.htm (дата обращения: 21.01.2022).

17. Tableau: визуализация данных для каждого – Текст: электронный // IBS Training Center: [сайт]. – URL: https://ibs-training.ru/about/news/Tableau_vizualizaciya_dannih_dlya_kajdogo/ (дата обращения: 21.01.2022).

18. Sisense – Текст: электронный // Morning Dough: [сайт]. – URL: https://www.morningdough.com/ru/ai-tools/sisense/ (дата обращения: 21.01.2022).

19. Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.

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

21. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann.

22. Zikopoulos, P., Eaton, C., deRoos, D., Deutsch, T., & Lapis, G. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media.

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

24. Marz, N., & Warren, J. (2015). Big Data: Principles and Best Practices of Scalable Realtime Data Systems. Manning Publications.

25. 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.

26. 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

27. Social media listening with salesforce social studio— Текст: электронный // Real Consulting: [сайт]. – URL: https://www.realconsulting.de/articles/social-media-listening (дата обращения: 21.01.2022).

28. IQBuzz – Текст: электронный // Startpack: [сайт]. – URL: https://startpack.ru/application/iqbuzz-smm (дата обращения: 21.01.2022).

29. Brand Analytics - система мониторинга и анализа – Текст: электронный // brandanalytics: [сайт]. – URL: https://brandanalytics.ru/ (дата обращения: 21.01.2022).

30. Hadoop – Текст: электронный // ВикипедиЯ: [сайт]. – URL: https://ru.wikipedia.org/wiki/Hadoop (дата обращения: 21.01.2022).

31. Apache Spark – Текст: электронный // ВикипедиЯ: [сайт]. – URL: https://ru.wikipedia.org/wiki/Apache_Spark (дата обращения: 21.01.2022).

32. White, T. (2015). Hadoop: The Definitive Guide (4th ed.). O'Reilly Media.

33. Lublinsky, B., Smith, K. T., & Yakubovich, A. (2013). Professional Hadoop Solutions. Wrox.

34. Schuler, D. (1994). "Social Computing." Introduction to Social Computing special edition of the Communications of the ACM, Volume 37, Issue 1, Pages 28-108.

35. boyd, d.m. and Ellison, N.B. (2007), Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication, 13: 210-230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

36. Аудитория восьми крупнейших соцсетей в России в 2023 году: исследования и цифры [Электронный ресурс]. – URL: https://ppc.world/articles/auditoriya-vosmi-krupneyshih-socsetey-v-rossii-issledovaniya-i-cifry/ (дата обращения: 21.01.2022).

37. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2007. – 384 с.: ил. + CD-ROM

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

39. Golbeck J., Robles C., Turner K. Predicting personality with social media // CHI’11 extended abstracts on human factors in computing systems. – ACM. 2011. – С. 253–262.

40. Adali S., Golbeck J. Predicting personality with social behavior // Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. – IEEE. 2012. – С. 302–309.

41. Youyou W., Kosinski M., Stillwell D. Computer-based personality judgments are more accurate than those made by humans // Proceedings of the National Academy of Sciences. – 2015. – Т. 112, № 4. – С. 1036–1040.

42. Personality, gender, and age in the language of social media: The openvocabulary approach / H. A. Schwartz [и др.] // PloS one. – 2013. – Т. 8, № 9. – e73791.

43. Пермские ученые научились определять психотип пользователя по его комментариям в соцсетях – Текст: электронный // COMNEWS: [сайт]. – URL: https://www.comnews.ru/content/202675/2019-10-31/2019-w44/permskie-uchenye-nauchilis-opredelyat-psikhotip-polzovatelya-ego-kommentariyam-socsetyakh (дата обращения: 21.01.2022).

44. Ученые РФ запатентовали программу для психолингвистического анализа пользователей соцсетей – Текст: электронный // МИНОБРНАУКИ РОССИИ: [сайт]. – URL: https://www.minobrnauki.gov.ru/press-center/news/main/23262/ (дата обращения: 21.01.2022).

45. Разработка алгоритма идентификации факторов риска безопасности пользователей социальных сетей на основе анализа контента и психологических характеристик его потребителей – Текст: электронный // frpss.tilda.ws: [сайт]. URL: http://frpss.tilda.ws/page20092272.html (дата обращения: 21.01.2022).

46. Мацута В.В., Мундриевская Ю.О., Сербина Г.Н., Пешковская А.Г. Identification Strategy of Deviant Communities in Social Media (as Exemplified by School Shooting) Social and Behavioral Sciences, – (год публикации – 2020)

47. R. B. Tareaf, P. Berger, P. Hennig, and C. Meinel, “Personality Ex-ploration System for Online Social Networks: Facebook Brands Asa Use Case,” 2018 IEEE/WIC/ACM International Conference onWeb Intelligence (WI), 2018 (PDF) Personality Prediction from Social Media Text: An Overview. Available from: https://www.researchgate.net/publication/341873172_Personality_Prediction_from_Social_Media_Text_An_Overview [accessed Jun 03 2024].

48. M. Vaidhya, B. Shrestha, B. Sainju, K. Khaniya, and A. Shakya,“Personality Traits Analysis from Facebook Data”, 21stInternational Computer Science and Engineering Conference(ICSEC), 2017 (PDF) Personality Prediction from Social Media Text: An Overview. Available from: https://www.researchgate.net/publication/341873172_Personality_Prediction_from_Social_Media_Text_An_Overview [accessed Jun 03 2024].

49. Branitskiy A., Doynikova E., Kotenko I. Technique for classification of social network users by psychological scales of Ammon’s test on the basis of artificial neural networks. Proceedings of the conference “Information Technologies in Control” (ITC 2020). 2020.

50. Liu F., Perez J., Nowson S. A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts // arXiv preprint arXiv:1610.04345. – 2016.

51. Y. Neuman, Y. Cohen, A Vectorial Semantics Approach to Personality Assessment, Scientific Reports. 4 (2014) 4761. https://doi.org/10.1038/srep04761.

52. F Alam, E A Stepanov, and Giuseppe Riccardi. 2013. Personality traits recognition on social network-facebook. WCPR (ICWSM-13), Cambridge, MA, USA (2013).

53. Gozde Ikizer, Marta Kowal, Ilknur Dilekler Aldemir, Alma Jeftic, Aybegum Memisoglu-Sanli, Arooj Najmussaqib, David Lacko, Kristina Eichel, Fidan Turk, Stavroula Chrona, Oli Ahmed, Jesper Rasmussen, Raisa Kumaga, Muhammad Kamal Uddin, Vicenta Reynoso-Alcantara, Daniel Pankowski, and Tao Coll-Martín, "Big Five traits predict stress and loneliness during the COVID-19 pandemic: Evidence for the role of neuroticism", Elsevier,2022.

54. Mr. R. Valanarasu, "Comparative Analysis for Personality Prediction by Digital Footprints in Social Media", Journal of Information Technology and Digital World, 2021, Pages: 77-91.

55. Yang Li, Amirmohammad Kazameini, Yash Mehta, and Erik Cambria, "Multitask Learning for Emotion and Personality Detection", IEEE, 2021.

56. Fatemeh Mohades Deilami, Hossein Sadr, and Mozhdeh Nazari, "Using Machine Learning-Based Models for Personality Recognition", arXiv, 2022.

57. Ghina Dwi Salsabila and Erwin Budi Setiawan, "Semantic Approach for Big Five Personality Prediction on Twitter", Rumah Jurnal Elektronik Ikatan Ahli Informatika Indonesia, 2021

58. Олпорт Г. Становление личности. Избранные труды. – М.: Смысл, 2002

59. Гордон Олпорт: диспозициональная теория личности [Электронный ресурс]. – URL: https://psychojournal.ru/psychologists/146-gordon-olport-dispozicionalnaya-teoriya-lichnosti.html (дата обращения: 21.01.2022).

60. Baranovskaya, M. S. (2005). Pyatifaktornaya model’ lichnosti P. Kosta i R. MakKreya i ee vzaimosvyaz’ s faktornymi teoriyami lichnosti G. Aizenka i R. Kettella [Five Factor model of personality by P. Costa and R. McCrae and its relations with the factor theories of personality by H. Eysenck and R. Cattell]. Psikhologicheskii Zhurnal, 26(4), 52–57. (in Russian)

61. Hassan, H., Asad, S., & Hoshino, Y. (2016). Determinants of Leadership Style in Big Five Personality Dimensions. Universal Journal of Management, 4(4), 161-179.https://doi.org/10.13189/ujm.2016.040402

62. Шмелев А. Г., Взорин Г. Д., Рыбникова М.К. Шестифакторная модель личности на базе психосемантического исследования русскоязычной лексики личностных черт // Организационная психология. – 2021. – №3. – С. 92-105.

63. Модель личности HEXACO – Текст: электронный // Наш ум прекрасен: [сайт]. URL: https://isurv.ru/model-lichnosti-hexaco/ (дата обращения: 01.05.2020).

64. Amirhosseini M. H., Kazemian H. Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator® // Multimodal Technologies and Interaction. – 2020. – Mar. – Vol. 4, no. 1. – P. 9. – ISSN 2414-4088. – DOI: 10.3390/mti4010009. – URL: https://www.mdpi.com/2414-4088/4/1/9.

65. Hernandez R., Knight I. Predicting Myers-Bridge Type Indicator with text classification// Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA. – 2017. – P. 4-9.

66. Rushton S., Morgan J., Richard M. Teacher’s Myers-Briggs personality profiles: Identifying effective teacher personality traits // Teaching andTeacher Education. – 2007. – Т. 23, № 4. – С. 432–441.

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

68. 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.

69. Можно ли использовать данные из соцсетей – Текст: электронный // Vc.ru: [сайт]. – URL: https://vc.ru/legal/59184-mozhno-li-ispolzovat-dannye-iz-socsetey (дата обращения: 21.01.2022).

70. Гражданский кодекс Российской Федерации (часть первая) от 30.11.1994 N 51-ФЗ (ред. от 11.03.2024) — Текст: электронный // КонсультантПлюс: [сайт]. – URL:https://www.consultant.ru/document/cons_doc_LAW_5142/9c307a0f2164645c15ca4e3146ff5f6e56060b23/ (дата обращения: 11.03.2024).

71. Федеральный закон "О персональных данных" от 27.07.2006 N 152-ФЗ (последняя редакция) – Текст: электронный // КонсультантПлюс: [сайт]. – URL: https://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения: 11.03.2024).

72. Правила защиты информации о пользователях сайта VK.com – Текст: электронный // Vk.com: [сайт]. – URL: https://vk.com/privacy (дата обращения: 11.03.2024).

73. 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.

74. Saha A., Sindhwani V. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization // Proceedings of the

fifth ACM international conference on Web searchand data mining. - ACM. 2012. - С. 693–702.

75. Gräßer F. et al. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning //Proceedings of the 2018 International Conference on Digital Health. – 2018. – С. 121-125

76. Poria S. et al. A rule-based approach to aspect extraction from product reviews // Proceedings of the second workshop on natural language processing for social media (SocialNLP). – 2014. – С. 28-37.

77. Al-Smadi M. et al. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews //International Journal of Machine Learning and Cybernetics. – 2019. – Т. 10. – №. 8. – С. 2163-2175.

78. Giannakopoulos A. et al. Unsupervised aspect term extraction with b-lstm & crf using automatically labelled datasets // Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. – 2017. – С. 180-188.

79. 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.

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

81. 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.

82. Yarushkina N. G., Moshkin V. S., Andreev I. A. The sentiment-analysis algorithm of social networks text resources based on ontology // Информационные технологии и нанотехнологии (ИТНТ-2020). – 2020. – pp. 226-232.

83. 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.

84. Токенизация в Python с использованием NLTK — Текст: электронный // pythobyte.com: [сайт]. – URL: https://pythobyte.com/tokenization-in-python-using-nltk-96642092/ (дата обращения: 11.03.2021).

85. Нормализация данных в Python – Текст: электронный // PYTHONIST: [сайт]. – https://pythonist.ru/normalizacziya-dannyh-v-python/ (дата обращения: 25.09.2022).

86. Обработка текстов на естественных языках – Текст: электронный // Хабр: [сайт]. – URL: https://habr.com/ru/company/vk/blog/358736/ (дата обращения: 11.03.2021).

87. Подходы лемматизации с примерами на Python – Текст: электронный // Еще один блог веб-разработчика: [сайт]. – URL: https://webdevblog.ru/podhody-lemmatizacii-s-primerami-v-python/ (дата обращения: 20.09.2022).

88. Краткий обзор техник векторизации в NLP – Текст: электронный // Хабр: [сайт]. – URL: https://habr.com/ru/articles/778048/ (дата обращения: 11.03.2021).

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

90. Horev R. BERT Explained: State of the art language model for NLP [Электронный ресурс]. — 11/2018. — URL: https://towardsdatascience.com/bert-explained-state-of-the-art-language-modelfor-nlp-f8b21a9b6270.

91. Word2Vec Project — Текст: электронный // Google Code: [сайт]. – URL: https://code.google.com/archive/p/word2vec/ (дата обращения: 11.03.2021).

92. Практическое руководство по NLP: изучаем классификацию текстов с помощью библиотеки fastText— Текст: электронный // Proglib: [сайт]. – URL: https://proglib.io/p/prakticheskoe-rukovodstvo-po-nlp-izuchaem-klassifikaciyu-tekstov-s-pomoshchyu-biblioteki-fasttext-2021-08-28 (дата обращения: 11.03.2021).

93. Руководство по NLTK с использованием Python — Текст: электронный // BI CONSULT: [сайт]. – URL: https://datafinder.ru/products/rukovodstvo-po-nltk-s-ispolzovaniem-python (дата обращения: 11.03.2021).

94. Обработка естественного языка. Python и spaCy на практике. — СПб.: Питер, 2021. – 256 с.: ил. – (Серия «Библиотека программиста»). ISBN 978-5-4461-1506-8

95. 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.

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

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

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

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

100. 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.

101. Знакомство с API ВКонтакте — Текст: электронный // VK.com: [сайт]. – URL: https://vk.com/dev/first_guide/ (дата обращения: 11.03.2021).

102. 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.

103. 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.

104. 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.

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

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

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

108. LSTM – нейронная сеть с долгой краткосрочной памятью — Текст: электронный // Neurohive: [сайт]. – URL: https://neurohive.io/ru/osnovy-data-science/lstm-nejronnaja-set/ (дата обращения: 11.03.2021).

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

110. 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

111. 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.

112. 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.

113. 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.

114. Чистяков С.П. Случайные леса: обзор // Труды Карельского научного центра РАН. – 2013. – № 1. – С. 117–136

115. Cover T. M., Hart P. E. Nearest neighbor pattern classification //Information Theory, IEEE Transactions on. – 1967. – Т. 13. – №. 1. – С. 21-27.

116. 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.

117. Malkov Y. et al. Approximate nearest neighbor algorithm based on navigable small world graphs //Information Systems. – 2014. – Т. 45. – С. 61-68.

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

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

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

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

122. 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.

123. 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.

124. Khan, Chandler, and Mahfuzul Hasan. *Anomaly Detection Principles and Algorithms*. Springer, 2019.

125. Томас Марк Тиленс React в действии. СПБ.: Питер, 2019. – 368 с.: ил. – (Серия «Для профессионалов»). ISBN 978-5-4461-0999-9

126. Дронов В. А. Django 3.0. Практика создания веб-сайтов на Python. СПб.: БХВ-Петербург, 2021. – 704 с. ил. (Профессиональное программирование) ISBN 978-5-9775-6691-9

127. (MBTI) Myers-Briggs Personality Type Dataset – Текст: электронный // kaggle: [сайт]. – URL: https://www.kaggle.com/datasets/datasnaek/mbti-type (дата обращения: 11.03.2023).

128. Personality-prediction – Текст: электронный // kaggle: [сайт]. – URL: https://www.kaggle.com/datasets/qatask/personalityprediction (дата обращения: 11.03.2023).

129. MBTI Personality Types 500 Dataset – Текст: электронный // kaggle: [сайт]. – URL: https://www.kaggle.com/datasets/zeyadkhalid/mbti-personality-types-500-dataset (дата обращения: 11.03.2023).

130. Функция потерь перекрестной энтропии: Обзор – Текст: электронный // weights biases: [сайт]. – URL: https://wandb.ai/wandb_fc/russian/reports/---VmlldzoxNDI4NjAw (дата обращения: 11.10.2023).

131. Manning, C. D., Raghavan, P., & Schütze, H. (2008). An Introduction to Information Retrieval. Cambridge University Press.

132. Miller, H., & Smith, R. (2018). "Cross-Domain Analysis of User Behavior: A Multi-Aspect Approach." Proceedings of the 2018 International Conference on Machine Learning and Data Mining, 102-113.

133. Wang, H., & Liu, J. (2020). "A Cross-Domain Framework for Analyzing User Behavior and Personality Traits." International Journal of Computational Social Science, 12(1), 89-105.

ПРИЛОЖЕНИЯ

ПРИЛОЖЕНИЕ 1. Свидетельства о государственной регистрации программ для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2022662518

Программа для анализа архетипов пользователей социальных сетей с использованием открытых данных профиля

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Пензенский государственный технологический университет» (RU)***

Авторы: ***Зоткина Алена Александровна (RU), Мартышкин Алексей Иванович (RU), Данилов Евгений Александрович (RU)***

Заявка № **2022662150**

Дата поступления **01 июля 2022 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***05 июля 2022 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023682329

Программа для автоматизированной очистки базы гетерогенных данных

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Пензенский государственный технологический университет» (RU)***

Авторы: ***Зоткина Алена Александровна (RU), Мартышкин Алексей Иванович (RU)***

Заявка № **2023681640**

Дата поступления **20 октября 2023 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***24 октября 2023 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

ПРИЛОЖЕНИЕ 2. Акты внедрения результатов кандидатской диссертации

«УТВЕРЖДАЮ»
Ректор ФГБОУ ВО «Пензенский государственный технологический университет»
_____ Л.В. Пащенко
«30» _____ 2024 г.

АКТ
о внедрении результатов диссертационной работы
Зоткиной Алены Александровны

Комиссия в составе:

председатель комиссии – к.т.н., доцент Сёмочкина И.Ю. – начальник учебно-методическое управление ФГБОУ ВО «Пензенский государственный технологический университет»;

члены комиссии:

к.т.н., доцент Ремонтов А.П. – декан факультета автоматизированных информационных технологий ФГБОУ ВО «Пензенский государственный технологический университет»;

д.т.н., профессор Курносов В.Е. – профессор кафедры «Программирование» ФГБОУ ВО «Пензенский государственный технологический университет»;

к.т.н., профессор Бершадская Е.Г. – профессор кафедры «Программирование» ФГБОУ ВО «Пензенский государственный технологический университет»,

составила настоящий акт о том, что результаты диссертационной работы Зоткиной А.А. на тему «Методы и алгоритмы формирования психологического портрета пользователя социальной сети для эффективного подбора кадров», представленной на соискание ученой степени кандидата технических наук, внедрены в учебный процесс кафедры «Программирование» ФГБОУ ВО «Пензенский государственный технологический университет».

Автором получены новые научные результаты:

1. Проведен анализ 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 объединения информации, 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.

Указанные результаты внедрены в учебный процесс кафедры «Программирование» по направлениям подготовки 09.03.01 «Информатика и вычислительная техника» (профили подготовки «Системы искусственного интеллекта», «Информационные технологии и искусственный интеллект в инженерии») и 09.03.04 «Программная инженерия» (профили подготовки «Программирование», «Программирование систем искусственного интеллекта», «Программное обеспечение систем искусственного интеллекта») при проведении лекционных и лабораторных работ по дисциплинам: «Методы машинного обучения и искусственного интеллекта», «Сбор и управление большими данными», «Технологии больших данных».

Внедрение полученных автором научных результатов позволило повысить качество учебного процесса.

Председатель комиссии _____ И.Ю. Сёмочкина

Члены комиссии _____ А.П. Ремонтов

_____ В.Е. Курносов

_____ Е.Г. Бершадская

ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ

ТОРГОВЫЙ ДОМ "ПЕНЗЕНСКИЙ ЗАВОД ЭНЕРГЕТИЧЕСКОГО МАШИНОСТРОЕНИЯ"

ISO 9001-2015

440028, г. Пенза, ул. Германа Титова, д. 5; тел. (841-2) 99-16-01, -02, -03, -04; факс (841-2) 99-16-05, -06, -07, -08
E-mail: marketing@pzem.ru; http://www.pzem.ru/; ИНН 5837022182, КПП 583501001, ОГРН 1045803504251

УТВЕРЖДАЮ
Управляющий ООО «ТД «ПЗЭМ»
_____ О.А. Калюжный
«07» 07 2022 г.

070722 № 144076

АКТ

внедрения результатов исследований, полученных в диссертационной работе аспиранта Пензенского государственного технологического университета (ПензГТУ) Зоткиной Алены Александровны

Комиссия в составе:

председатель комиссии –

управляющий ООО «ТД «ПЗЭМ» Калюжный Олег Александрович;

члены комиссии:

руководитель отдела управления персоналом Аброськина Светлана Сергеевна;

специалист по кадрам Пономарева Светлана Геннадьевна

настоящим актом подтверждает, что результаты диссертационной работы Зоткиной А.А. внедрены в деятельность ООО «ТД «ПЗЭМ» в рамках проекта «Кадры для цифровой экономики».

Использование результатов диссертационной работы позволило повысить эффективность управления кадровой системы на 13%.

Председатель комиссии / Калюжный О.А.
Члены комиссии: / Аброськина С.С.
/ Пономарева С.Г.

«07» 07 2022 г.

УТВЕРЖДАЮ
Генеральный директор
АО «Научно-производственное
предприятие «Рубин»

А.В. Данилов

АКТ

о внедрении результатов диссертационной работы на тему
«Методы и алгоритмы формирования психологического портрета
пользователя социальной сети для эффективного подбора кадров»

Комиссия АО «НПП «Рубин» в составе:
- начальника научно-технического центра к.т.н. Кузнецова В.Е.,
- ученого секретаря, д.т.н., профессора Бутаева М.М.;
- главного специалиста научно-технического центра, д.т.н., доцента Бабича М.Ю.,

составила настоящий акт о том, что результаты диссертационной работы Зоткиной А.А., представленной на соискание учёной степени кандидата технических наук, используются в АО «НПП «Рубин» в деятельности предприятия в рамках выполнения составной части научно-исследовательской работы «Метрика-Р».

Использование результатов диссертационной работы обеспечивает более глубокий и детализированный анализ состояния операторов системы охраны объектов особой важности, что в свою очередь способствует более осознанному и обоснованному принятию решений в процессе пресечения попыток проникновения нарушителей в охраняемую зону, снижая риски, связанные с эмоциональной и психологической нагрузкой должностных лиц системы.

Результаты работы аспиранта А.А. Зоткиной использованы в процессе проектных исследований создания системы охраны объектов с замкнутой границей.

Настоящий акт не является основанием для материальных претензий сторон.

Начальник НТЦ, к.т.н. Кузнецов В.Е.

Учёный секретарь, д.т.н., профессор Бутаев М.М.

Главный специалист НТЦ, д.т.н., доцент Бабич М.Ю.

SECON
ассоциация разработчиков

АССОЦИАЦИЯ РАЗРАБОТЧИКОВ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ
ПЕНЗЕНСКОЙ ОБЛАСТИ «СЕКОН»

440600, г. Пенза, ул. Суворова, стр 66, 6 этаж
org@secon.ru | secon.ru

АКТ

внедрения результатов диссертационной работы аспиранта Пензенского государственного технологического университета (ПензГТУ) Зоткиной Алены Александровны

Ассоциация разработчиков программного обеспечения Пензенской области «Секон»

Подтверждаем, что указанные результаты диссертационного исследования Зоткиной А.А. были использованы при разработке решений для систем подбора кадров ряда организаций входящих в Ассоциацию разработчиков программного обеспечения Пензенской области (CodeInside, Tortuga) и позволяют повысить качество процесса подбора сотрудников в корпоративных структурах.

В частности, внедрение:

- методологии и алгоритмов анализа психологических профилей пользователей социальных сетей;
- метода объединения данных из разных социальных сетей для формирования комплексного психологического портрета;
- методики кросс-доменного аспектно-ориентированного анализа тональности текстов;

предоставляет дополнительный инструмент для более глубокой оценки личных качеств и профессиональных склонностей кандидатов. Использование данных материалов диссертационной работы Зоткиной А.А. способствует улучшению точности подбора кандидатов, более точному соответствию требованиям вакансий и оптимизации рекрутинговых процессов.

С уважением,

исполнительный директор Ассоциации «СЕКОН» Белов С.Е.

«23» сентября 2024г.

УТВЕРЖДАЮ
Заместитель генерального директора
АО «ППО ЭВТ им. В.А. Ревунова»
_____ /А.В. Володин
«27» 06 2024г.

АКТ
реализации результатов кандидатской диссертации на тему
«Методы и алгоритмы формирования психологического портрета пользователя социальной сети для эффективного подбора кадров»
Зоткиной Алены Александровны

Комиссия в составе: Бражников Александр Олегович – начальник конструкторского отдела №4, к.т.н., Кузнецова Екатерина Николаевна – начальник отдела по управлению персоналом, - составила настоящий акт о том, что результаты диссертационной работы Зоткиной А.А., представленной на соискание учёной степени кандидата технических наук, используются в АО «ППО ЭВТ им. В.А. Ревунова» г. Пенза в деятельности предприятия.

В частности, внедрение нейросетевой методики для определения психологических характеристик кандидата, с использованием типологии *MBTI*, способствует более осознанному и обоснованному принятию решений при подборе сотрудников, повышая качество соответствия кандидатов требованиям вакансий и снижая риски, связанные с эмоциональной и психологической совместимостью в коллективе.

Настоящий акт не является основанием для материальных претензий сторон.

Члены комиссии:

Начальник конструкторского отдела №4
АО «ППО ЭВТ им. В.А. Ревунова», к.т.н. А.О. Бражников

Начальник отдела по управлению
персоналом
АО «ППО ЭВТ им. В.А. Ревунова» Е.Н. Кузнецова