На правах рукописи

Андреев Илья Алексеевич

ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ СОЦИАЛЬНЫХ СЕТЕЙ В ЗАДАЧАХ ФОРМИРОВАНИЯ СОЦИАЛЬНОГО ПОРТРЕТА ПОЛЬЗОВАТЕЛЯ

Специальность: 05.13.01 – Системный анализ, управление и обработка информации (информационные технологии и промышленность)

Автореферат

диссертации на соискание ученой степени
кандидата технических наук

Ульяновск – 2022

Работа выполнена на кафедре «Информационные системы» федерального государственного бюджетного образовательного учреждения высшего образования «Ульяновский государственный технический университет».

Научный руководитель: **Мошкин Вадим Сергеевич**
кандидат технических наук, доцент кафедры “Информационные системы” Федерального государственного бюджетного образовательного учреждения высшего образования «Ульяновский государственный технический университет»

Официальные оппоненты: **Куприянов Александр Викторович**, доктор технических наук, заведующий кафедрой технической кибернетики, исполнительный директор института информатики и кибернетики Федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева»

Абрамов Максим Викторович, кандидат технических наук, заведующий лабораторией теоретических и междисциплинарных проблем информатики Федерального государственного бюджетного учреждения науки «Санкт-Петербургского Федерального исследовательского центра Российской академии наук» (СПб ФИЦ РАН)

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича»

Защита диссертации состоится «14» сентября 2022 г. в 15 часов 00 минут на заседании диссертационного совета Д 212.277.04 при Ульяновском государственном техническом университете по адресу: 432027, г. Ульяновск, ул. Северный Венец, 32 (ауд. 211, Главный корпус).

С диссертацией можно ознакомиться в библиотеке Ульяновского государственного технического университета. Также диссертация и автореферат размещены в Internet на сайте УлГТУ – http://www.ulstu.ru.

Автореферат разослан « » _____ 2022 г.

Ученый секретарь
диссертационного совета,
д.т н., доцент Наместников А.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Социальные сети – неотъемлемая часть современного общества. Концепция Web 2.0, основной составляющей которой стало появление и бурное развитие социальных сетей, предполагает формирование контента электронных ресурсов (в том числе и текстового) пользователями через свои профили (сообщения, комментарии, наименования файлов и др.).

Многие бизнес-задачи, которые ранее было невозможно решить из-за недостатка данных, теперь могут быть решены с помощью анализа социальных сетей. Исследовательские центры и коммерческие компании по всему миру используют данные социальных сетей для моделирования экономических, социальных, политических и других процессов различного уровня с целью разработки механизмов воздействия на них.

Мониторинг и анализ слабоструктурированных данных социальных сетей включает сбор и структурирование различной информации: фотографии, заметки, сообщения, связи между пользователями, сообщества, контакты, личная информация. Из количественных и качественных характеристик извлекаются закономерности и строятся необходимые математические модели для их описания. Текстовые данные в социальных сетях обладают следующими особенностями:

- Использование разговорных оборотов, неологизмов, а также различных диалектических форм.
- Наличие односоставных и неполных предложений.
- Наличие речевых и орфографических ошибок.
- Использование авторских символов выражения эмоций (т.н. «смайлов» и «эмоджи») с целью придания сообщению определенной эмоциональной окраски.

Анализ больших данных и поведения пользователей социальных сетей открывает новые возможности для исследования личностных черт, таких как построение и проверка предсказательных моделей о личностных чертах и поведении людей, в том числе и с использованием методов анализа русскоязычных текстовых данных. В работах Widiger T., Costa M.R., Ледовой Я.А., Паничевой П.В. авторы находят прямую взаимосвязь между социальным портретом пользователя социальной сети и тональностью их текстовых сообщений. Исследователи Wiggins J.S., Piedmont R.L., Ozer U., Тулупьев А.Л, Абрамов М.В. в своих работах показали, что результаты применения таких методов анализа текстовых данных пользователей социальных сетей, как сентимент анализ, психолингвистический анализа, методы классификации текстов, могут быть прямыми коррелянтами результатов анализа пользователей согласно различным моделям личностных черт, в том числе и модели «Большой пятерки».

Актуальность темы

Одним из важнейших направлений применения моделей и алгоритмов

анализа неструктурированных данных социальных сетей является построение социального портрета пользователя в рамках подбора кадров организаций. В настоящее время важным критерием, который все чаще приходится учитывать при отборе персонала, - это безопасность организации-работодателя. При проверке работников в ходе отбора приходится иметь в виду материальные, профессиональные и социальные риски. Эффективная работа с социальными сетями может принести значительную пользу при реализации функции системы управления персоналом организации.

В настоящее время сбор и/или содержательный анализ собранной в социальных сетях информации проводится вручную специалистами кадровых служб, что требует больших затрат времени и ограничивает объем обрабатываемой информации. Интеллектуальный анализ структурированных и слабоструктурированных ресурсов социальных сетей предполагает необходимость системного решения ряда научных задач:

- необходимость моделей унификации и агрегации данных из различных социальных сетей с учетом особенностей их представления в разных источниках;
- необходимость в адаптации методов обработки естественного языка к особенностям представления информации в социальных сетях:
- необходимость решения проблемы автоматизированного сопоставления профилей пользователей в разных социальных сетях на основании структурированных и слабоструктурированных данных со страниц профилей;
- необходимость решения задачи формирования обучающей выборки для классификации русскоязычных текстов по семи классам эмоций с учетом особенностей представления текстовых данных в социальных сетях;
- необходимость в разработке подхода к определению психологических характеристик пользователя социальных сетей посредством анализа текстовых сообщений;
- необходимость в решении задачи классификации русскоязычных текстовых сообщений из социальных сетей по эмоциональной составляющей.

Поэтому актуальной является тема диссертации, направленная на снижение трудозатрат при обеспечении полноты информации посредством автоматизации и учета дополнительных факторов в процессе анализа открытых русскоязычных текстовых данных пользователей социальных сетей с использованием интеллектуальных алгоритмов на основе гибридизации семантических подходов и методов машинного обучения.

Цель диссертационной работы

Целью диссертации является снижение трудозатрат на построение социального портрета пользователей социальных сетей посредством автоматизации и учета дополнительных факторов в процессе анализа открытых русскоязычных текстовых данных.

Объектом исследования является набор открытых русскоязычных текстовых данных, извлекаемых со страниц пользователей социальных сетей.

Предметом исследования являются модели и алгоритмы психолингвистического и сентимент-анализа русскоязычных текстовых данных социальных сетей.

Задачи исследования

В соответствии с целью работы актуальными являются следующие задачи диссертационного исследования:

- провести анализ существующих работ по формированию обучающих выборок и сентимент-анализу текстовых постов социальной сети;
- провести сравнение современных интеллектуальных методов анализа текстовых данных, выявления их возможностей и ограничений в рамках психолингвистического и сентимент-анализа данных постов в социальной сети;
- разработать алгоритм формирования обучающей выборки, состоящей из открытых русскоязычных текстовых ресурсов социальных сетей, классифицированных по 7-ми эмоциям;
- разработать алгоритм классификации текстовых сообщений социальной сети по классам тональности на основе семантических подходов и машинного обучения;
- разработать подход к сопоставлению профилей пользователей в разных социальных сетях посредством анализа структурированных и неструктурированных данных анкет, а также социальных графов профилей;
- разработать метод определения психологических характеристик пользователя социальных сетей посредством анализа текстовых сообщений в социальных сетях;
- разработать программную систему психолингвистического и сентимент-анализа открытых текстовых русскоязычных данных профилей пользователей социальных сетей;
- провести вычислительные эксперименты, позволяющие оценить эффективность предложенных методов и алгоритмов;
- внедрить результаты исследований в практику процесса подбора персонала организаций региона.

При решении задачи оценки эффективности предложенных моделей и алгоритмов необходима адаптация условий проведения экспериментов под специфику решаемых задач.

Методы исследования

В диссертационной работе применяются методы онтологического инжиниринга, методы обработки естественного языка, нейросетевые методы, методы машинного обучения (Machine Learning), методы теории анализа социальных сетей (Social Network Analysis, SNA), а также объектно-ориентированного программирования при построении программного комплекса.

Область исследования

Область исследования соответствует паспорту специальности 05.13.01. –

«Системный анализ, управление и обработка информации (технические науки)», а именно:

п. 4 – разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации;

п. 10 – методы и алгоритмы интеллектуальной поддержки при принятии управленческих решений в технических, экономических, биологических, медицинских и социальных системах.

Научная новизна

Научная новизна результатов исследования заключается в следующем:

1. Разработан алгоритм формирования обучающей выборки для обучения моделей классификации в задачах сентимент-анализа текстовых данных, отличающийся совместным использованием словарей авторских символов выражения эмоций и ключевых фраз.
2. Предложен подход к сопоставлению профилей пользователей в разных социальных сетях, отличающийся гибридизацией подходов анализа графической информации, структурированных данных анкет, текстовых данных, а также социальных графов профилей.
3. Разработан метод определения психологических характеристик пользователя социальных сетей, отличающийся гибридизацией алгоритмов обработки естественного языка текстовых данных, машинного обучения и метода «Большой пятерки».
4. Предложен алгоритм анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей, отличающийся интеграцией семантических подходов и методов машинного обучения.

Теоретическая значимость работы

Теоретическая значимость работы заключается в разработке и реализации новых эффективных моделей и алгоритмов психолингвистического и сентимент-анализа открытых русскоязычных текстовых данных пользователей социальных сетей, позволяющих снизить трудозатраты при решении задач подбора персонала организации.

Практическая значимость работы

Разработанный программный комплекс, реализующий предложенные модели и алгоритмы, был применен в рамках проекта «Интеллектуальная платформа формирования социального портрета соискателя на основании семантико-когнитивного анализа профилей в социальных сетях» компании ООО «ФаззиЛаб», поддержанного Фондом содействия инноваций, применен общественной организацией «Федерация бадминтона» в рамках проекта «Парабадминтон: все силы - для победы», поддержанного Фондом Президентских грантов, с целью поиска волонтеров для сопровождения спортсменов с ПОДА, а также в проектной деятельности в рамках проекта «Система интеллектуального поиска и анализа в Интернет-СМИ и социальных сетях», реализуемого совместно с Федеральным научно-производственным

центром АО «Научно-производственное объединение «Марс» (ФНПЦ АО НПО «Марс»).

Основания для выполнения работы

Результаты диссертационной работы использовались в ряде НИОКР, выполненных в Ульяновском государственном техническом университете, направленных на решение научно-технических задач. К наиболее важным результатам следует отнести:

1. Участие в выполнении гранта РФФИ № 18-47-732007 р_мк «Разработка методов семантического анализа слабоформализованных данных проектной документации на основе модели индивидуальных предпочтений проектировщика».
2. Участие в выполнении гранта РФФИ № 18-47-730035 р_а «Исследование и разработка методов извлечения и анализа семантики слабоструктурированных ресурсов социальных сетей на основе онтологического инжиниринга и машинного обучения».
3. Участие в реализации гранта по программе Старт-ЦТ/Искусственный интеллект на тему «Интеллектуальная платформа формирования социального портрета соискателя на основании семантико-когнитивного анализа профилей в социальных сетях» при поддержке Фонда содействия развитию малых форм предприятий в научно-технической сфере.
4. Участие в реализации государственного задания №075-00233-20-05 от 03.11.2020 на выполнение государственных работ в сфере научной деятельности Минобрнауки России по проекту «Исследование интеллектуального предиктивного мультимодального анализа больших данных и извлечения знаний из различных источников».

Достоверность результатов диссертационной работы

Достоверность научных положений, выводов и рекомендаций подтверждена результатами вычислительных экспериментов и результатами практического использования.

Основные положения, выносимые на защиту

1. Разработанный алгоритм формирования обучающей выборки позволяет эффективно решать задачу обучения нейронной сети в процессе сентимент-анализа русскоязычных текстов социальных сетей;
2. Предложенный подход к сопоставлению профилей пользователей в разных социальных сетях реализован в программном комплексе и автоматизирует процесс поиска профилей пользователя в задаче построения социального портрета;
3. Предложенный метод определения психологических характеристик пользователя социальных сетей с применением методов машинного обучения и модели «Большой пятерки» позволяет классифицировать пользователя по пяти основным факторам данной модели;
4. Разработанный алгоритм анализа эмоциональной окраски

русскоязычных текстовых данных, отличающийся интеграцией семантических подходов и методов машинного обучения, повышает точность классификации текстов социальных сетей по классам тональности.

Апробация работы

Основные положения и результаты диссертационной работы докладывались, обсуждались и получили одобрение на следующих конференциях, семинарах и симпозиумах:

- XXIX Международной конференции «Computational Science and Its Applications»-ICCSA-2019 (г.Санкт-Петербург, 2019 г.);
- Международной научно-технической конференции «Автоматизация» - RusAutoConf-2020 (г.Сочи, 2020 г.);
- XII Международной конференции Developments in eSystems Engineering – DESE- 2019 (г.Казань, 2019 г.);
- V Международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» (г. Минск, 2015 г.);
- VIII и IX Международных научно-практических конференциях «Интегрированные модели и мягкие вычисления в искусственном интеллекте» (г. Коломна, 2015, 2020 гг.);
- I Международной научной конференции «Интеллектуальные информационные технологии в технике и на производстве» (г. Сочи, 2016 г.);
- III, V, VI, VII Международных конференциях и молодежных школах «Информационные технологии и нанотехнологии» (г.Самара, 2017, 2019, 2020, 2021);
- VIII Международной конференции «Системный анализ и информационные технологии» САИТ – 2019 (г.Иркутск, 2019 г.);
- I Международной Поспеловской летней школе-семинаре для студентов, магистрантов и аспирантов «Методы и технологии гибридного и синергетического искусственного интеллекта» (г. Светлогорск, 2014 г.);
- V Всероссийской Поспеловской конференции с международным участием «Гибридные и синергетические интеллектуальные системы» (г. Светлогорск, 2020 г.);
- XVII национальной конференции по искусственному интеллекту с международным участием «КИИ-2019» (г. Ульяновск, 2019 г.);
- IV Всероссийской научно-практической мультиконференции с международным участием «Прикладные информационные системы»-ПИС-2017 (г. Ульяновск, 2017 г.);
- 6-й Всероссийской научно-технической конференции аспирантов, студентов и молодых ученых ИВТ-2014 (г. Ульяновск, 2014 г.).

Научные публикации

По результатам работы было опубликовано 32 статьи, 4 из которых в журналах из перечня ВАК, 11 статей в изданиях, индексируемых в Scopus и/или

Web Of Science, а также 1 монография. Получены 3 свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад

Постановка задач исследования осуществлялась совместно с научным руководителем. Все результаты, составляющие содержание диссертации, получены автором самостоятельно. Подготовка к публикации некоторых результатов проводилась совместно с соавторами, причем вклад соискателя был определяющим.

Структура и объем диссертации

Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы и приложений. Основное содержание работы изложено на 166 страницах, включая 50 рисунков и 11 таблиц. Список использованных источников состоит из 170 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрена актуальность выполненного исследования, формулируются цель и задачи работы, определяются теоретическая значимость и практическая ценность результатов исследования, а также положения, выносимые на защиту.

В **первой главе** приводится описание бизнес-процесса подбора персонала в современных организациях, а также областей данного бизнес-процесса, которые могут быть оптимизированы с точки зрения трудозатрат посредством интеллектуализации анализа данных социальных сетей. Выявлена необходимость разработки методов унификации данных извлеченных из различных социальных сетей, разработка методов сопоставления профилей людей в разных социальных сетях, а также методы психолингвистического и сентимент-анализа неструктурированных данных.

Также в **первой главе** приводится описание современных методов формирования обучающей выборки для сентимент-анализа текста. Описаны такие методы, как программная генерация, применение сэмплирования, вероятностные и детерминированные методы. Приведены тенденций сентимент-анализа текстовой информации, упор сделан на методы обработки русского языка, в том числе с применением машинного обучения. Рассмотрены различные варианты применения сентимент-анализа, в том числе для оценки общественного мнения и классификации текстов. Рассмотрены существующие методы построения психологического портрета человека на основе публичной информации социальных сетей. Выявлена необходимость гибридизации методов инженерии знаний и машинного обучения для улучшения качества сентимент-анализа текста и применение сентимент-анализа в рамках построения психологического портрета. Рассмотрены архитектуры, применяемые методы систем и алгоритмы, выполняющих схожие задачи.

Во **второй главе** приводится описание унифицированной онтологической модели профилей социальных сетей, которую можно представить следующим

образом:

$$O^{SN} = \{N^{SN}, R^{SN}, F^{SN}\},$$

где N^{SN} – множество узлов (объектов и классов) онтологии;
R^{SN} – множество отношений онтологии, $R^{SN} \in N^{SN} \times N^{SN}$;
F^{SN} – множество функций интерпретации (аксиом) онтологии;

$$N^{SN} = N^{B} \cup N^{COM} \cup N^{DOM};$$

$N^{B} = \{n_1^{B}, n_2^{B}, \dots, n_m^{B}\}$ –Узловые объекты – пользователи социальной сети
$N^{COM} = \{n_1^{COM}, n_2^{COM}, \dots, n_l^{COM}\}$ – внутренние объекты-сущности социальных сетей (Группа, Пост, Комментарий, Вложение).
$N^{DOM} = \{n_1^{DOM}, n_2^{DOM}, \dots, n_k^{DOM}\}$ – объекты материального мира: в/ч, школа, ВУЗ, город, государство, музыкальная группа, книга и пр.).
Типы отношений:

$$R^{SN} = R^{OP} \cup R^{DTP} \cup R^{CONT}$$

$R^{OP} = \{r_1^{OP}, r_2^{OP}, \dots, r_s^{OP}\}$ –Object Properties (hasFriend, hasFollower etc.), т.е. отношения между объектами онтологии;
$R^{DTP} = \{r_1^{DTP}, r_2^{DTP}, \dots, r_h^{DTP}\}$ – DataType Properties, т.е. отношения между объектами онтологии и значениями встроенного типа (Boolean, String, Number).
R^{CONT} – Annotation Properties, это свойства аннотации, необходимые для определения контекста.

В рамках предложенной модели, было определено два типа контекста:

- R^{CSN} –отношение аннотации, в которой отражена социальная сеть, из которой извлечены данные.

- R^{T}- отношение аннотации, в которой отражен временной промежуток, на протяжении которого данное отношение было актуально, при этом

$$(\forall r_i \in R^{OP}, R^{DTP}) \, , \exists \, r_i^{CONT} \in R^{CONT}, r_i^{CONT} = \{ r_i^{T}, r_i^{CSN}\}.$$

Схематично учет временного, а также контекста источника данных в рамках предложенной модели приведен на рисунке 1.

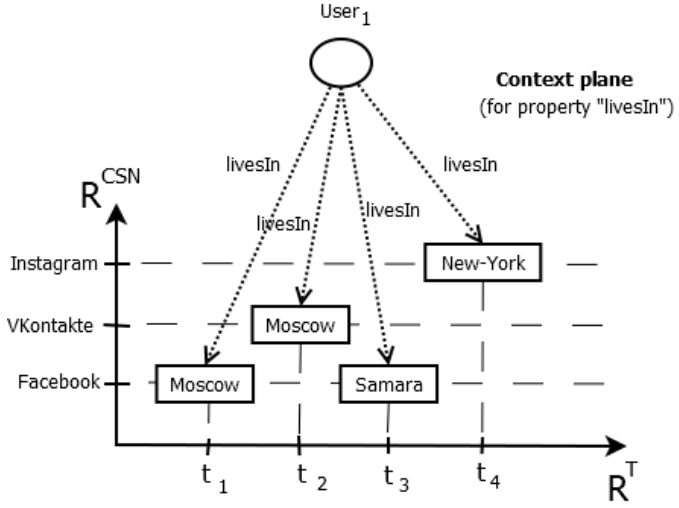

Рисунок 1 Соотношение контекстов времени и источника

Также во **второй главе** приводится описание алгоритма формирования обучающей выборки, который состоит из следующих этапов:

Шаг 1. Формирование словаря авторских символов выражения эмоций.

1.1. Отбор среди символов Unicode массива символов выражения эмоций.

1.2. Выбор агрегирующих символов в полученном массиве путем группировки по признакам, не влияющим на семантику символа (цвет символа, его «род» или «возраст»).

1.3. Классификация агрегирующих символов на 7 классов:

$$D = \{D_{joy}, D_{sad}, D_{surp}, D_{anger}, D_{disg}, D_{cont}, D_{fear}\}$$

где D_{joy} – класс объектов с эмоцией «радость», D_{sad} – класс объектов с эмоцией «грусть», D_{surp} – класс объектов с эмоцией «удивление», D_{anger} – класс объектов с эмоцией «злость», D_{disg} – класс объектов с эмоцией «отвращение», D_{cont} – класс объектов с эмоцией «презрение», D_{fear} – класс объектов с эмоцией «страх». Примеры таких символов представлены на рисунке 2.

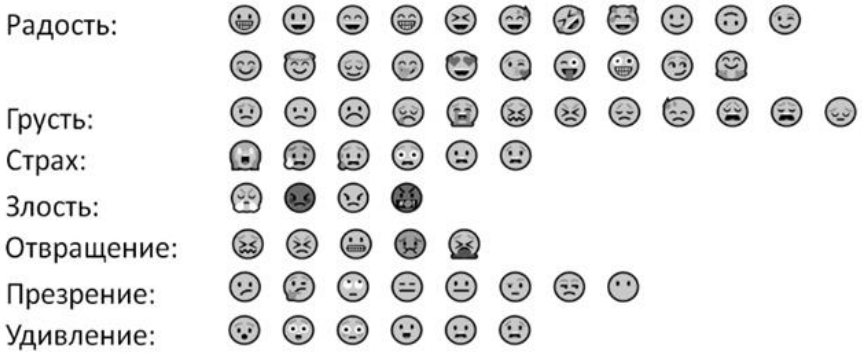

Рисунок 2 Словарь авторских символов выражения эмоций

Шаг 2. Формирование словаря ключевых фраз путем экспертного расширения тезауруса WordNet-Affect.

2.1 Добавление ключевых фраз в тезаурус.

2.2 Классификация добавленных слов/словосочетаний на 7 классов, определяемых в п.1.3 алгоритма.

Шаг 3. Автоматическое извлечение открытых текстовых данных сообщений из профилей социальных сетей.

Шаг 4. Предобработка извлеченных текстовых данных:

Шаг 5. Отбор текстовых сообщений из предобработанных текстовых данных, содержащих элементы словаря авторских символов выражения эмоций, сформированного на Шаге 1.

Шаг 6. Отбор из полученного на Шаге 5 текстового массива набора текстовых сообщений, содержащих элементы сформированного на Шаге 2 словаря ключевых фраз.

Также во **второй главе** приводится описание оригинального подхода к сопоставлению профилей пользователей в разных социальных сетях посредством гибридизации различных подходов анализа структурированных и неструктурированных данных страниц социальных сетей, который основан на применении следующих критериев:

1. Критерий схожести анкет профилей.

2. Критерий наличия схожих лиц на фотографиях.
3. Критерий наличия схожих контактов.
4. Критерий наличия схожего места работы и учебы.
5. Критерий наличия схожих сообщений со страниц профиля с посредством нахождения расстояния Левенштейна и использования алгоритма шинглов.
6. Критерий совпадения социальных графов.

Также во **второй главе** описан новый метод определения психологических характеристик пользователя социальных сетей с применением методов машинного обучения и метода «Большой пятерки».

Схематично общий процесс психолингвистического анализа данных социальных сетей с использованием машинного обучения и модели Большой Пятерки представлен на рисунке 3.

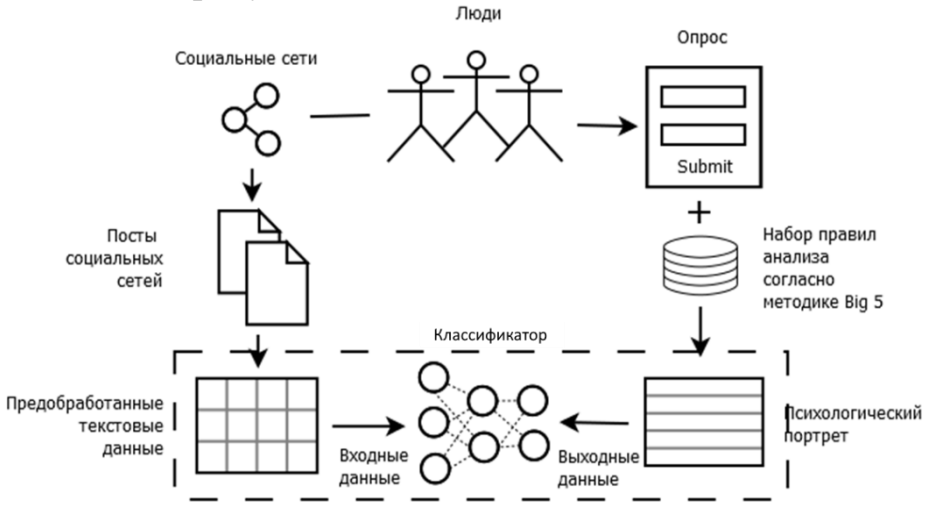

Рисунок 3 Подход к психолингвистическому анализу данных социальных сетей с использованием машинного обучения и модели Большой Пятерки

Шаг 1. Обучение классификатора.

1.1 Проведение опроса среди пользователей социальных сетей по методике «Большой пятерки».

1.2 Получение психологического портрета пользователей по 5 вторичным факторам путем оценки результатов прохождения опроса по методике «Большой пятерки».

1.3 Автоматическое извлечение открытых текстовых данных сообщений из профилей социальных сетей пользователей, участвующих в опросе по методике «Большой пятерки».

1.4 Предобработка извлеченных текстовых данных.

1.5 Обучение классификатора путем использования в качестве входных значений – предобработанных текстовых данных, выходных – результатов получения психологического портрета пользователей по 5 вторичным факторам.

Шаг 2. Определение психологических характеристик пользователя социальных сетей

2.1 Извлечение, предобработка и векторизация текстовых данных

социальных сетей с использованием подходов, примененных при формировании обучающей выборки.

2.2 Классификация предобработанных и векторизированных текстовых данных с применением обученного на Шаге 3 с получением 5 значений вторичных факторов психологического портрета пользователя в соответствии с методикой «Большой пятерки».

Помимо этого, **во второй главе** представлен оригинальный алгоритм анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей на основании интеграции семантических подходов и методов машинного обучения:

1. Формирование обучающей выборки для обучения нейронной сети с использованием разработанного алгоритма совместного применения расширенного тезауруса WordNet-Affect и словаря авторских символов выражения эмоций.
2. Векторизация обучающей выборки посредством применения языковой модели BERT.
3. Обучение нейронной сети эффективной архитектуры на полученной обучающей выборке.
4. Извлечение, предобработка и векторизация текстовых данных социальных сетей с использованием подходов, примененных при формировании обучающей выборки.
5. Классификация предобработанных и векторизированных текстовых данных с использованием обученной на Шаге 3 нейронной сети на 7 классов тональности.

Схематично данный алгоритм можно представить следующим образом (рисунок 4):

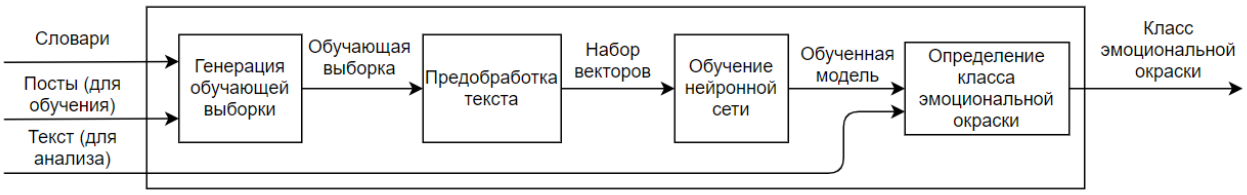

Рисунок 4 Процесс сентимент-анализа текстов социальных сетей

В **третьей главе** рассмотрена архитектура и функциональные возможности разработанного в рамках диссертационной работы программного комплекса. Диаграмма развертывания программного комплекса представлена на рисунке 5.

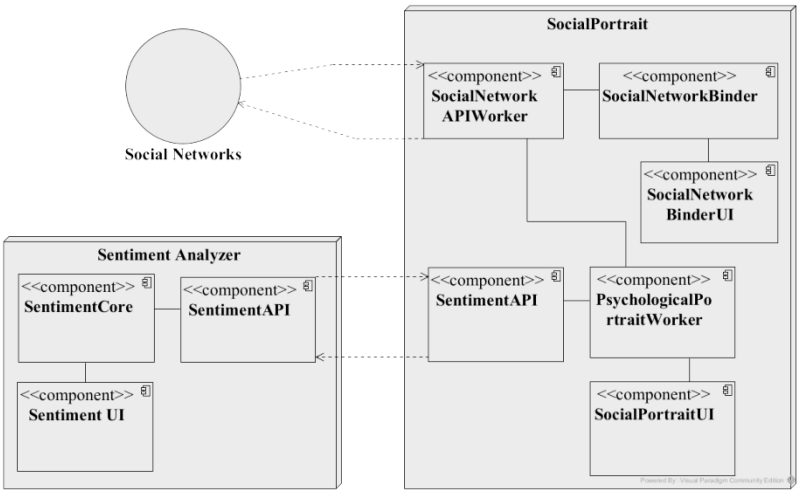

Рисунок 5. Диаграмма развертывания разработанного программного комплекса

Программный комплекс состоит из двух систем, реализован с использованием микросервисной архитектуры, реализует разработанные модели и алгоритмы и обеспечивает анализ профилей пользователей социальных сетей для построения психологического портрета и сентимент-анализа русскоязычных текстов.

Отдельные системы программного комплекса могут быть использованы независимо – так, сервис определения тональности текста может быть использован при помощи API или автономно посредством веб-интерфейса.

В **четвертой главе** представлены подтверждающие эффективность результаты проведенных экспериментов по применению разработанных алгоритмов:

1. Эксперименты по объединению профилей пользователей в различных социальных сетях.

В качестве экспериментальной базы была использована заранее подготовленная выборка из 100 пользователей, имеющих профили в различных социальных сетях. Все эти пользователи имели 204 аккаунта. Для каждого из этих аккаунтов были подобраны похожие профили с помощью разработанного алгоритма. Также были подсчитаны результаты для критериев сравнения профилей, результат показан на рисунке 6.

Рисунок 6 Результаты экспериментов по объединению профилей пользователей в различных социальных сетях

2. Эксперименты по сентимент-анализу текстовых данных, включали в себя эксперименты по оценке алгоритма формирования обучающей выборки; оценке использования языковых моделей Word2Vec и BERT; оценке использования разных языковых моделей, словарей формирования обучающей выборки, количества постов и длин сообщений; оценке использования разных словарей формирования обучающей выборки; оценке архитектур нейронных сетей.

Для проведения данного набора экспериментов было обработано около 2,5 миллионов текстовых сообщений социальной сети. Текстовые сообщения были загружены через API «ВКонтакте» из открытых групп. Результаты экспериментов приведены в таблице 1 и на рисунке 8.

Таблица 1 – Результаты экспериментов по оценке использования разных языковых моделей, словарей формирования обучающей выборки, количества постов и длин сообщений

№	Языковая модель	Кол-во постов	Словари при получении обучающей выборки	Выборка сбалансирована	Веса классов	Длина сообщения	Точность на тестовой выборке
1	word2vec	1042	смайлы, ключевые слова	нет	нет	40-50 слов	0,77
2	word2vec	1042	смайлы, ключевые слова	нет	да	40-50 слов	0,79
3	BERT	556	смайлы, ключевые слова	нет	нет	90-110 сим.	0,86
4	BERT	556	смайлы, ключевые слова	нет	да	90-110 сим.	0,87
5	BERT	726	смайлы	нет	да	90-110 сим.	0,82
6	BERT	2100	ключевые слова	да	да	90-110 сим.	0,83
7	BERT	513	смайлы, ключевые слова, без стоп слов	нет	да	90-110 сим.	0,82

Проведенные эксперименты показали: наиболее эффективным алгоритмом сентимент-анализа русскоязычных текстовых данных социальных сетей стал подход, включающий в качестве классификатора многослойный персептрон, языковую модель BERT, а также алгоритм формирования обучающей выборки, использующий авторские символы выражения эмоций и расширенный словарь WordNet-Affect.

Рисунок 7 Результаты экспериментов по оценке архитектур различных нейронных сетей в задаче сентимент-анализа русскоязычных текстов социальных сетей

Наивысшая точность классификации постов на тестовой выборке была достигнута при использовании многослойного персептрона – 0,87.

3. Эксперименты по оценке алгоритма психолингвистического анализа текста профилей социальных сетей.

Было проведено 3 множества экспериментов с разбивкой множества классифицируемых объектов на обучающую и тестовую выборку в соотношениях: 70/30, 60/40 и 50/50. Результаты экспериментов представлены на рисунке 8.

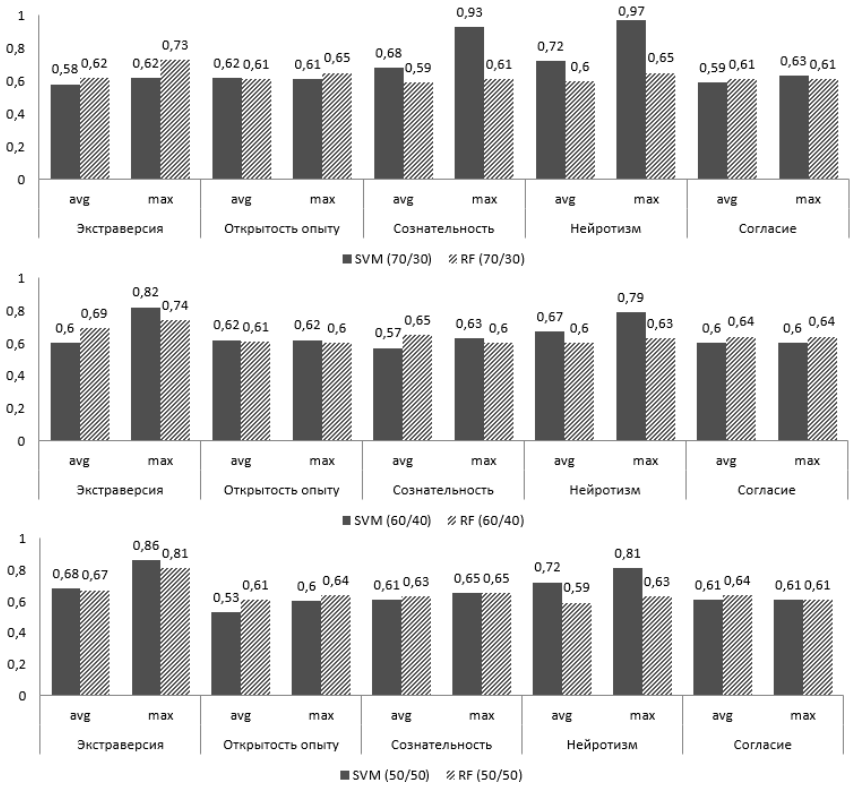

Рисунок 8 Результаты экспериментов по оценке алгоритма психолингвистического анализа текста профилей социальных сетей

Как видно из графика, наилучшие результаты были получены при использовании SVM и разбиении обучающей и тестовой выборки соотношением 70/30 (от 0,58 до 0,93 для различных показателей Big5).

В рамках проверки правильности выбора метода классификации был проведен эксперимент по сравнению различных методов бинарной классификации пользователей по всем пяти классам. Для сравнения были добавлены наивный Байесовский классификатор и линейная регрессия. Во всех случаях оценке подлежали аккаунты, принадлежащие 100 добровольцам. Выборка была разделена на обучающую и тестовую в пропорции 70/30. Результаты по 5 факторам личности в каждом случае были приведены к среднему значению. Результаты эксперимента представлены в таблице 2.

Таблица 2 – Сравнение различных методов классификации в рамках эксперимента по психолингвистическому анализу данных социальных сетей

№	Алгоритм классификации	Средняя точность метода (%)
1	Наивный Байесовский классификатор	57,50
2	Метод опорных векторов (SVM)	65,12
3	Линейная регрессия	51,44
4	Метод случайного леса (RF)	59,92

Лучший результат показал метод опорных векторов, достигнув средней точности 65,12%. Необходимо учесть тот факт, что минимальная точность в экспериментах 58%, а максимальная – 93% в зависимости от исследуемого психологического показателя. Такой разброс данных связан с недостаточно большой обучающей выборкой, а также с тем фактом, что обучающая выборка была построена на основе данных добровольцев, которые изначально имеют особый психологический портрет – так, например, среди анкетных данных оказалось 74% людей, которые по шкале «экстраверсия – интроверсия» оказались в категории «экстраверсия».

Кроме того, в 4 главе представлены примеры практического применения разработанных алгоритмов и реализующего программного комплекса в рамках проекта «Интеллектуальная платформа формирования социального портрета соискателя на основании семантико-когнитивного анализа профилей в социальных сетях» компании ООО «ФаззиЛаб», поддержанного Фондом содействия инноваций.

Также разработанный в рамках диссертационной работы программный комплекс был апробирован в рамках проекта «Парабадминтон: все силы - для победы», поддержанного Фондом Президентских грантов при отборе волонтеров, обеспечивающих сопровождение лиц с ПОДА для УОСОО «Федерация бадминтона».

Условия отбора кандидатов:

- возраст: от 15 до 40 лет;
- место проживания: Ульяновская область, г. Ульяновск;
- положительная эмоциональная окраска оригинальных текстов профилей социальных сетей относительно терминов «инвалиды», «помощь», «волонтер»;
- положительная эмоциональная устойчивость.

Условия и результаты проведения экспериментов: проанализировано 10116 профилей в социальной сети «ВКонтакте»; удовлетворили условиям поиска – 17 человек. После личного собеседования из них было отобрано 9 человек. Было сэкономлено порядка 14 человеко-часов на проведение данного эксперимента.

Также эффективность использования научно-технических результатов подтверждена экспериментальными исследованиями, целью которых являлось определение количественной оценки временной сложности нахождения

профилей пользователей с учетом установленных параметров в сравнении с экспертным отбором, проведенным сотрудниками ФНПЦ АО НПО «Марс» в рамках проекта «Система интеллектуального поиска и анализа в Интернет-СМИ и социальных сетях».

Отбор пользователей проводился в социальной сети «ВКонтакте» с учетом параметров таких, как возраст, место проживания, области интересов, эмоциональная окраска комментариев, и т.д. и был направлен на:

- выявление пользователей социальной сети «ВКонтакте», распространяющих материалы на определенную тему;
- мониторинг необходимых тем в пространстве социальной сети «ВКонтакте»;
- поиск текстовых сообщений в социальной сети «ВКонтакте», имеющих отношение к возникшей ситуативной задаче, формируемой в виде множества ключевых слов;
- поиск пользователей социальной сети «ВКонтакте» по неполной информации в профиле.

Достигнуто среднее сокращение времени поиска профилей, удовлетворяющих условиям отбора, на 40%.

В **заключении** сформулированы основные выводы и результаты диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В ходе диссертационного исследования получены следующие результаты:

1) проведен анализ существующих работ по формированию обучающих выборок и сентимент-анализу текстовых постов социальной сети;

2) проведено сравнение современных интеллектуальных методов анализа текстовых данных, их возможностей и ограничений в рамках психолингвистического и сентимент-анализа в социальной сети;

3) разработан алгоритм формирования обучающей выборки, состоящей из открытых русскоязычных текстовых ресурсов социальных сетей, классифицированных по 7-ми эмоциям;

4) разработан метод классификации текстовых постов социальной сети по классам тональности на основе семантических подходов и машинного обучения;

5) разработан подход к сопоставлению профилей пользователей в разных социальных сетях посредством анализа структурированных и неструктурированных данных анкет, а также социальных графов профилей;

6) разработан подход к определению психологических характеристик пользователя социальных сетей посредством анализа текстовых сообщений в социальных сетях;

7) разработан программный комплекс психолингвистического и сентимент-анализа открытых текстовых русскоязычных данных профилей пользователей социальных сетей;

8) проведены вычислительные эксперименты, основными результатами

которых являются:

- наиболее эффективным алгоритмом сентимент анализа русскоязычных текстовых данных социальных сетей стал подход, включающий в качестве классификатора многослойный персептрон, в качестве языковой модели – модель BERT, а также предполагающий в качестве алгоритм формирования обучающей выборки – алгоритм, использующий авторские символы выражения эмоций и расширенный словарь WordNet-Affetct (87% точности);
- наилучшие результаты по классификации пользователей социальных сетей по психологическим характеристикам были получены при использовании в качестве классификатора алгоритм SVM и разбиении обучающей и тестовой выборки соотношением 70/30 (от 0,58 до 0,93 для различных показателей Big5).
- использование подхода к построению психологического портрета пользователя социальных сетей, включающего разработанные алгоритмы психолингвистического и сентимент-анализа русскоязычных структурированных и неструктурированных ресурсов социальных сетей, позволило сократить трудозатраты на поиск волонтеров, обеспечивающих сопровождение лиц с ПОДА, в рамках проекта «Парабадминтон: все силы - для победы» для УОСОО «Федерация бадминтона» на 14 часов на 10116 пользователей.
- применение разработанных подходов в проектной деятельности ФНПЦ «АО «НПО Марс» в рамках проекта «Система интеллектуального поиска и анализа в Интернет-СМИ и социальных сетях» позволило сократить время поиска профилей, удовлетворяющих условиям отбора, в среднем на 40%.

9) результаты исследований внедрены в практику процесса подбора персонала организаций региона.

СПИСОК ПУБЛИКАЦИЙ

Статьи, опубликованные в журналах, рекомендованных Перечнем **ВАК** России:

1. Комбинирование статистического и лингвистического методов для извлечения двухсловных терминов из текста / И. А. Андреев, В. А. Башаев, В. В. Клейн, Н. Г. Ярушкина // Автоматизация процессов управления. – 2013. – № 4 (34). – С. 61–70.

2. Применение способа интеграции нечетких временных рядов и нечѐтких онтологий в задачах диагностики технических систем / [Н. Г. Ярушкина, В. С. Мошкин, И. А. Андреев и др.] // Онтология проектирования. – 2018. – Т. 8, № 4 (30). – С. 594–604.

3. Семантическая метрика терминологичности на основе онтологии предметной области / [И. А. Андреев, В. А. Башаев, Н. Г. Ярушкина и др.] // Автоматизация процессов управления. – 2014. – № 4 (38). – С. 76–84.

4. Ярушкина, Н. Г. Алгоритм психолингвистического анализа текстовых данных социальных сетей с применением модели «Большая пятёрка» / Н. Г. Ярушкина, В. С. Мошкин, И. А. Андреев // Онтология проектирования. – 2022. – Т. 12, №1 (43). – С. 82–92.

проиндексированные в SCOPUS и/или WebOfScience

5. Development of a Mobile System for Interactive Forecasting of Statistical Graph Data / V. Moshkin, D. Averin, I. Moshkina, I. Andreev // International Conference on Information Technology and Nanotechnology : (ITNT-2021). – P. 1–5.

6. An Intelligent Search Algorithm for Extremist Texts / V. Moshkin, D. Fadeev, D. Kurilo, I. Andreev, //International Conference on Information Technology and Nanotechnology (ITNT-2021) – P. 1–4.

7. Moshkin, V. The Extending the Knowledge Base of the Intelligent CAD of a Design Organization Using Semantic Analysis of Wiki-Resources / V. Moshkin, N. Yarushkina, I. Andreev // Advances in Automation II, Proceedings of the International Russian Automation Conference (RusAutoConf2020) – P. 1–4.

8. Yarushkina, N.G. Solving the problem of determining the author of text data using a combined assessment / Yarushkina N.G., Moshkin V.S., Andreev I.A. // Proceedings of 8th International Conference «Fuzzy Systems, Soft Computing and Intelligent Technologies 2020 (FSSCIT 2020)», CEUR WS Proceedings – Vol. 2782 – P. 112–118.

9. An integrated approach to mapping user profiles on social networks / Vladimir Belov, Dmitriy Drozdov, Roman Shakurov, Vadim Moshkin, Ilya Andreev // Proceedings of the Data Science Session at the VI International Conference on Information Technology and Nanotechnology (DS-ITNT 2020) – Vol. 2667. – P. 225–228.

10. Yarushkina, Nadezhda Approaches to sentiment analysis of the social network text data / Yarushkina Nadezhda, Moshkin Vadim, Andreev Ilya // Proceedings of the Data Science Session at the VI International Conference on Information Technology and Nanotechnology (DS-ITNT 2020) – Vol. 2667. – P. 198–202.

11. Moshkin, V. The Sentiment Analysis of Unstructured Social Network Data Using the Extended Ontology SentiWordNet / V. Moshkin, N. Yarushkina, I. Andreev IEEE, 12th International Conference on Developments in eSystems Engineering (DeSE) – Kazan, Russia – 2019 – P. 576-580.

12. Yarushkina, N. Hybridization of fuzzy time series and fuzzy ontologies in the diagnosis of complex technical systems / Nadezhda Yarushkina, Vadim Moshkin, Ilya Andreev, Gelya Ishmuratova // Proceedings of the Data Science Session at the V International Conference on Information Technology and Nanotechnology (DS-ITNT 2019) – Vol. 2416 – P. 252-259.

13. Yarushkina, N. Integration of Fuzzy OWL Ontologies and Fuzzy Time Series in the Determination of Faulty Technical Units / Yarushkina N., Andreev I., Moshkin V., Moshkina I. // Computational Science and Its Applications : (ICCSA 2019) – Lecture Notes in Computer Science, – 2019 – Vol. 11619 – Springer, Cham, P. 545-555.

14. Andreev, I. Attacking the problem of continuous speech segmentation into basic units / Ilya A. Andreev, Andrey I. Armer, Natalia A. Krasheninnikova, Vadim S. Moshkin // IPGTIS-ITNT 2017 : Proceedings of the International conference

Information Technology and Nanotechnology. Session Data Science, Samara, Russia, 24–27 April. – Samara, 2017. – P. 6–9.

15. Yarushkina, N. Hybridization of Fuzzy Inference and Self-learning Fuzzy Ontology-Based Semantic Data Analysis. / Yarushkina N., Moshkin V., Klein V., Andreev I, Beksaeva E.: // Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’16), P. 277–285 (2016)

в иных изданиях

16. Интеллектуальный предиктивный мультимодальный анализ слабоструктурированных больших данных / Ярушкина Н. Г., Андреев И. А., Гуськов Г. Ю. [и др.]. – Ульяновск : УлГТУ, 2020. – 220 с. – Библиогр.: с. 201–220 (202 назв.).

17. Yarushkina, N. G. The sentiment-analysis algorithm of social networks text resources based on ontology / Yarushkina N. G., Moshkin V. S., Andreev I. A. // Сборник трудов по материалам VI Международной конференции и молодежной школы / под редакцией В. А. Фурсова. – 2020. – С. 226–232.

18. Андреев, И. А. Подход к интеграции нечетких временных рядов и нечетких онтологий / И. А. Андреев // Вузовская наука в современных условиях : сборник материалов 55-й научно-технической конференции (25–30 янв.). – Ульяновск : УлГТУ, 2021. – Ч. 2. – С. 233–236.

19. Фадеев, Д. О. Интеллектуальный алгоритм поиска текстов экстремистской направленности / Д. О. Фадеев, В. С. Мошкин, И. А. Андреев // Информационные технологии и нанотехнологии : (ИТНТ-2021) : сборник трудов по материалам VII Международной конференции и молодежной школы. – Самара, 2021. – С. 30612.

20. Мошкин, В. С. Разработка мобильной системы интерактивного прогнозирования данных статистических графиков / В. С. Мошкин, И. А. Андреев, Д. С. Аверин // Информационные технологии и нанотехнологии : (ИТНТ-2021) : сборник трудов по материалам VII Международной конференции и молодежной школы. – Самара, 2021. – С. 30622.

21. Комплексный подход к маппингу профилей пользователей в социальных сетях / [В. А. Белов, В. С. Мошкин, И. А. Андреев и др.] // Информационные технологии и нанотехнологии : (ИТНТ-2020) : сборник трудов по материалам VI Международной конференции и молодежной школы..... / под редакцией В. А. Фурсова. – Самара, 2020. – С. 220–225.

22. Мошкин, В. С. Алгоритмы машинного обучения для определения авторства текстовых фрагментов / В. С. Мошкин, И. А. Андреев, И. М. Шигабутдинов // Нечеткие системы, мягкие вычисления и интеллектуальные технологии : НСМВИТ-2020 : труды VIII Международной научно-практической конференции. – Смоленск : Универсум, 2020. – Т. 1. – С. 173–181.

23. Мошкин, В. С. Алгоритм анализа эмоциональной окраски текстовых ресурсов социальных сетей на основе онтологии / В. С. Мошкин, И. А. Андреев, Н. Г. Ярушкина // Семнадцатая Национальная конференция по

искусственному интеллекту с международным участием : КИИ-2019 (21–25 окт.) : сборник научных трудов. – Ульяновск : УлГТУ, 2019. – Т. 1. – С. 171–184.

24. Мошкин, В. С. Сравнение эффективности применения алгоритмов сентимент-анализа неструктурированных ресурсов социальных сетей / В. С. Мошкин, И. А. Андреев // Системный анализ и информационные технологии : САИТ-2019 : труды Восьмой Международной конференции, Иркутск-Листвянка, 8–14 июля. – Москва : Информатика и управление, 2019. – С. 534–540.

25. Гибридизация нечетких временных рядов и нечетких онтологий при диагностике сложных технических систем / В. С. Мошкин, Н. Г. Ярушкина, Г. Р. Ишмуратова, И. А. Андреев // Информационные технологии и нанотехнологии : ИТНТ-2019 : сборник трудов V Международной конференции и молодежной школы, 21–24 мая. – Самара : Новая техника, 2019. – Т. 4. – 399–407.

26. Автокорреляционные портреты произнесений звуков в задаче определения границ речевых единиц / И. А. Андреев, А. И. Армер, В. С. Мошкин, Н. А. Крашенинникова // Нечеткие системы и мягкие вычисления. Промышленные применения : сборник научных трудов IV Всероссийской научно-практической мультиконференции с международным участием "Прикладные информационные системы (ПИС-2017) (29–31 мая). – Ульяновск : УлГТУ, 2017. – С. 121–125.

27. Подход к решению задачи членения слитной речи на речевые единицы / И. А. Андреев, А. И. Армер, Н. А. Крашенинникова, В. С. Мошкин // Информационные технологии и нанотехнологии : ИТНТ-2017 : сборник трудов III Международной конференции и молодежной школы. – Самара : 2017. – С. 473–476.

28. Разработка системы анализа эмоциональной окраски текста с использованием расширенного словаря SENTIWORDNET 3.0 / И. А. Андреев, К. П. Мулянова, М. О. Синдюкова, А. В. Шараборина // Гибридные и синергетические интеллектуальные системы : материалы V Всероссийской Поспеловской конференции с международным участием, / под редакцией А. В. Колесникова. – Калининград : Изд-во БФУ им. И. Канта, 2020. – С. 498–505.

29. Лингвистический подход к автоматизированному построению предметной онтологии / И. А. Андреев, Е. А. Бексаева, В. В. Клейн, В. С. Мошкин, И. П. Серков. – Текст : электронный // Прикладные информационные системы : сборник научных трудов Третьей Всероссийской научно-практической конференции (30 мая – 12 июня). – Ульяновск : УлГТУ, 2016. – С. 256–263. – URL: http://venec.ulstu.ru/lib/disk/2016/141.pdf (дата обращения: 20.06.2022).

30. Андреев, И. А. Определение вероятности терминологичности словоупотреблений в текстах конкретной предметной области / И. А. Андреев, В. А. Башаев, В. В. Клейн, В. С. Мошкин // Интегрированные модели и мягкие вычисления в искусственном интеллекте : сборник научных трудов VIII Международной научно-практической конференции (Коломна, 18–20 мая). – Москва : Физматлит, 2015. – Т. 2. – С. 764–773.

31. Оценка терминологичности лексических единиц на основе онтологии предметной области / [Андреев И. А., Башаев В. А., Ярушкина Н. Г. и др.] // Открытые семантические технологии проектирования интеллектуальных систем : (OSTIS-2015) : материалы V Международной научно-технической конференции, 19–21 февр. – Минск : БГУИР. – 2015. – С. 395–400.

32. Семантическая метрика «термин/не термин» на основе онтологии проблемной области / В. С. Мошкин, И. А. Андреев, В. А. Башаев, В. В. Клейн // Методы и технологии гибридного и синергетического искусственного интеллекта : материалы I Международной Поспеловской летней школы-семинара для студентов, магистрантов и аспирантов. – Калининград : Изд-во БФУ им. И. Канта, 2014. – С. 67–73.

33. Использование семантической метрики для решения задачи извлечения терминологии из текста проблемной области / В. А. Башаев, В. С. Мошкин, И. А. Андреев, В. В. Клейн // Информатика и вычислительная техника : ИВТ-2014 : сборник научных трудов 6-й Всероссийской научно-технической конференции аспирантов, студентов и молодых ученых, 26–28 мая. – Ульяновск : УлГТУ, 2014. – С. 72–78.

Свидетельства на регистрацию программы для ЭВМ:

1. Свидетельство № 2021610716 Российская Федерация. Интеллектуальная система для Opinion Mining социальных медиа: свидетельство о государственной регистрации программы для ЭВМ / Ярушкина Н.Г., Мошкин В.С., Андреев И.А., Константинов А.А. ; заявитель и правообладатель Ульян. гос. техн. ун-т. – № 2021610025, заявл. 11.01.2021; зарегистр. 19.01.2021.

2. Свидетельство № 2014612325 Российская Федерация. Программа для автоматизированного извлечения терминологии из текста на основе лингвистических и статистических методов: свидетельство о государственнойрегистрации программы для ЭВМ / Ярушкина Н.Г., Башаев В.А., Клейн В.В., Андреев И.А. ; заявитель и правообладатель Ульян. гос. техн. ун-т. – № 2013662129, заявл. 25.12.2013; зарегистр. 25.02.2014.

3. Свидетельство № 2014616248 Российская Федерация. Онтологически-ориентированная система извлечения терминологии: свидетельство о государственной регистрации программы для ЭВМ / Ярушкина Н.Г., Башаев В.А., Мошкин В.С.; Клейн В.В., Андреев И.А., заявитель и правообладатель Ульян. гос. техн. ун-т. – № 2014613574, заявл. 21.04.2014; зарегистр. 18.06.2014.

Андреев Илья Алексеевич

Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя

Автореферат

Подписано в печать. 24.06.2022. Формат 60x84/16
Усл. печ. л. 1,43
Тираж 100 экз. Заказ
ИПК «Венец» УлГТУ, 432027, г. Ульяновск, Северный Венец, 32.