МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

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)

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель – кандидат технических наук
Мошкин Вадим Сергеевич

Ульяновск – 2022

Оглавление

Список сокращений 7
Введение 8
Глава 1. Анализ современного состояния моделей и методов интеллектуального анализа текстовых данных социальных сетей 18
1.1. Процесс подбора персонала 18
1.2. Объединение профилей пользователя социальных сетей. 20
1.2.1. Обзор существующих систем анализа социальных сетей 20
1.2.2. Идентификация пользователя в различных социальных сетях 23
1.3. Методы формирования обучающей выборки для сентимент-анализа текста 26
1.3.1. Группы методов оценки тональности текстовых данных 26
1.3.2. Математическое планирование 28
1.3.3. Программная генерация 28
1.3.4. Применение сэмплирования 28
1.3.5. Закономерная модификация базового объекта 29
1.3.6. Выборка из базы объектов 29
1.3.7. Вероятностные методы 29
1.3.8. Детерминированные методы 30
1.3.9. Методика формирования обучающего множества с использованием меры схожести и матрицы схожести 31
1.3.10. Проблемы формирования обучающей выборки 31
1.3.11. Семантические тезаурусы 32
1.4. Методы сентимент-анализа текста 34
1.4.1. Подходы к определению эмоциональной окраски текстов на русском языке 35
1.4.2. Применение сентимент-анализа текстов для оценки общественного мнения 36
1.4.3. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики 36

1.4.4. Анализ тональности текста на русском языке при помощи графовых моделей 37
1.4.5. Сентимент-анализ коротких русскоязычных текстов в социальных медиа 37
1.4.6. Лексико-грамматические маркеры эмоций в качестве параметров для сентимент-анализа русскоязычных интернет-текстов 38
1.4.7. Выбор топологии нейронных сетей и их применение для классификации коротких текстов 39
1.4.8. Проблемы определения тональности текста 39
1.4.9. Существующие системы определения тональности текста 39
1.5. Построение психологического портрета человека на основе публичной информации социальных сетей 41
1.5.1. Психологическая классификация людей 41
1.5.2. Классификация полученных данных 44
1.6. Существующие решения и аналоги 50
1.7. Выводы по главе 51
Глава 2. Методы и алгоритмы интеллектуального анализа текстовых данных социальных сетей 53
2.1. Унификация извлекаемых данных различных социальных сетей 53
2.2. Алгоритм формирования обучающей выборки 55
2.3. Подход к сопоставлению профилей пользователей с использованием гибридизации различных подходов структурированных и неструктурированных данных. 60
2.3.1. Критерии схожести профилей. 60
2.3.2. Критерий схожести лиц 61
2.3.3. Критерий схожести контактов, мест работы и учебы 62
2.3.4. Критерий схожести сообщений 63
2.3.5. Критерий совпадения социальных графов 65
2.4. Определение психологических характеристик пользователя социальных сетей 67
2.4.1. Классификация психологических характеристик пользователя с

использованием метода «Большой пятерки» 67
2.4.2. Психолингвистический анализ текстовых данных социальных сетей.. 69
2.5. Алгоритм анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей 72
2.6. Выводы по главе..... 83
Глава 3. Реализация программного комлекса интеллектуального анализа текстовых данных социальных сетей на основе интеграции семантических подходов и машинного обучения 84
3.1. Общая концепция программного комплекса 84
3.1.1. Диаграмма развертывания программного комплекса 84
3.1.2. Подходы к извлечению данных из социальных сетей 85
3.2. Проектирование и реализация подсистемы анализа тональности текстов 86
3.2.1. Диаграмма вариантов использования системы анализа тональности текстов 86
3.2.2. Диаграмма последовательности системы анализа тональности текстов 89
3.2.3. Диаграмма классов системы анализа тональности текстов..... 91
3.2.4. Диаграмма «сущность-связь» системы анализа тональности текстов .. 95
3.2.5. Диаграмма компонентов системы анализа тональности текстов 96
3.2.6. Диаграмма развертывания системы анализа тональности текстов 97
3.2.7. Диаграмма потоков данных системы анализа тональности текстов 98
3.2.8. Описание входных и выходных данных системы анализа тональности текстов 98
3.2.9. Описание реализации подсистемы анализа тональности текстов 99
3.3. Проектирование и реализация программного комплекса психолингвистического анализа социальных сетей..... 101
3.3.1. Диаграмма вариантов использования программного комплекса психолингвистического анализа социальных сетей..... 101
3.3.2. Диаграмма классов программного комплекса психолингвистического анализа социальных сетей..... 102

3.3.3. Диаграмма последовательности программного комплекса психолингвистического анализа социальных сетей 105
3.3.4. Диаграмма развертывания программного комплекса психолингвистического анализа социальных сетей 106
3.3.5. Программная реализация клиентской части программного комплекса психолингвистического анализа социальных сетей 107
3.3.6. Программная реализация серверной части программного комплекса психолингвистического анализа социальных сетей 111
3.3.7. Реализация классификатора текстов с целью определения психолингвистических характеристик автора 113
3.3.8. Реализация непрерывной интеграции и доставки в программном комплексе психолингвистического анализа социальных сетей 114
3.4 Выводы по главе 115
Глава 4. Анализ адекватности разработанных моделей и методов на основе вычислительных экспериментов и практики применения 116
4.1. План экспериментов 116
4.2. Эксперименты по объединению профилей пользователей в различных социальных сетях 117
4.3. Эксперименты по сентимент-анализу текстовых данных 119
4.3.1. Эксперименты по оценке алгоритма формирования обучающей выборки 119
4.3.2. Статистика этапов формирования обучающей выборки 120
4.3.3. Оценка разных языковых моделей при формировании обучающей выборки 120
4.3.4. Оценка использования разных словарей формирования обучающей выборки 121
4.3.5. Оценка использования разных языковых моделей, словарей формирования обучающей выборки, количества постов и длин сообщений 122
4.3.6. Оценка разных архитектур нейронных сетей в задаче сентимент-анализа текстовых ресурсов 123

4.4. Эксперименты по оценке алгоритма психолингвистического анализа текста профилей социальных сетей 124
4.5. Внедрение разработанных алгоритмов и методов..... 129
4.6. Выводы по главе..... 133
Заключение 135
Библиографический список 137
Приложение А. Блок-схема процесса подбора персонала. 154
Приложение Б. Таблица сравнения характеристик систем анализа социальных сетей 155
Приложение В. Акты внедрения 157
Приложение Г. Свидетельства о государственной регистрации программ для ЭВМ 164

Список сокращений

БД – база данных.

ИИ – искусственный интеллект.

ПО – программное обеспечение.

ПОДА – поражения опорно-двигательного аппарата.

ПрО – предметная область.

СМИ – средства массовой информации.

СУБД – система управления базами данных.

ЦА – целевая аудитория

API – программный интерфейс приложения.

AUC ROC – площадь под кривой ошибок.

Big5 – модель личности человека «Большая пятерка».

CNN – сверточная нейронная сеть.

CSS – каскадные таблицы стилей.

GRU – управляемый рекуррентный блок.

FPR – ложная положительная оценка.

LSTM – долгая краткосрочная память.

MLP – многослойный перцептрон.

RNN – рекуррентная нейронная сеть.

RF – метод случайного леса

SVM – метод опорных векторов.

TPR – верная положительная оценка.

URL – унифицированный указатель ресурса.

Введение

В современном мире неотъемлемой частью хранилища знаний всего человечества является глобальная сеть Интернет. С каждым годом возрастает количество информации, которая хранится в этой сети. Одним из видов хранения знаний глобальной сети Интернет являются данные человеческой деятельности, содержащиеся в социальных сетях.

Социальные сети занимают особую нишу в социальной жизни современного общества. Самыми популярными социальными сетями в мире являются «Facebook»*, имеющий более миллиарда уникальных посетителей в месяц, и «Twitter», который имеет около 300 миллионов посетителей. В РФ и странах СНГ наиболее популярной социальной сетью явлеся «ВКонтакте», которую посещает около 80 миллионов уникальных посетителей каждый месяц. Большинство посетителей каждый день пишут одно или несколько сообщений, которые так или иначе отражают их личную позицию. Суммарно данную информацию можно отнести к позиции граждан из разных стран и разных слоев общества [168].

Социальные сети можно исследовать с разными целями. Например, можно наблюдать (путем анализа комментариев) общественные мнения по тем или иным событиям, как мелким, так и крупным, выявлять общественно опасных индивидов и проводить иные мероприятия обеспечения безопасности населения. Анализ комментариев, постов и сообщений может помочь оценить изменения в настроениях, что дает большое количество данных для политических и социальных исследований, в том числе и в исследованиях потребительских предпочтений.

---

*Организация Meta, а также ее продукты Instagram и Facebook, 21 марта 2022г. Тверским судом города Москвы признаны экстремистскими и запрещены на территории РФ.

Коммерческие фирмы заинтересованы в получении мнений покупателей о разных продуктах, при этом мнения важны как фирмам, которые производят продукт самостоятельно, так и фирмам, занимающимся перепродажей товаров. Эти данные могут успешно использоваться для следующих целей: повышение качества продукта, определение и изменение ЦА, определения главных достоинств и недостатков своего продукта относительно продуктов конкурентов. При анализе тональности текстов сообщений пользователей сотрудник фирмы может сделать выводы о:

- эмоциональной оценке пользователей различных событий и объектов;
- предпочтениях отдельных пользователей;
- некоторых чертах характера пользователей [123].

Социальный портрет – это набор различных ценностей, присущих определенному человеку. Установки могут быть социальными, психологическими, экономическими, политическими или культурными, однако не существует однозначного научно обоснованного определения социального портрета [131], однако большинство авторов сходится во мнении, что социальный портрет, независимо от того, какой набор данных он хранит – это комплекс обобщенных характеристик человека или группы людей. Социальный портрет в данной диссертационной работе формируется путем агрегации результатов психолингвистического анализа текста и сентимент-анализа к объектам, событиям и персоналиям реального мира.

Тональностью текста называется эмоциональное отношение, выраженное в тексте автором или группой авторов к какому-либо событию, объекту или персоналии. На сегодняшний день уже разработан определенный набор различных методов анализа тональности текста. Их можно разделить на два класса: методы, основанные на словарях и методы, основанные на машинном обучении с учителем.

Согласно концепции Web 2.0, основной составляющей которой являются социальный сети, большая часть контента всех электронных ресурсов формируется пользователями. Одним из вариантов подобного

заполнения являются профили, которые состоят из постов, комментариев, файлов и др. Исходят из этого можно выделить следующие особенности данных социальных сетей:

- Текст можно отнести к разговорной стилистике. Это сопровождается, в том числе, использованием сленговых выражений, неологизмов, а также различных диалектических форм.
- Текст может содержать односоставные и неполные предложения.
- Текст часто содержит речевые и орфографические ошибки.
- Текст может содержать авторские символы выражения эмоций (т.н. «смайлов», «эмоджи»). Пользователи указывают их для придания сообщению определенной эмоциональной окраски или пояснения двусмысленных текстов.

Одним из направлений использования результатов анализа слабоструктурированной информации социальных сетей с привязкой к конкретному человеку, является процесс отбора персонала. Одним из важных критерием, учитываемым при отборе персонала, является безопасность работодателя. Работодатель хочет обеспечить в своей организации как социальную безопасность – не брать на работу людей, склонных к воровству, алкоголизму или связанным с криминальными структурами, так и информационную безопасность. В современном мире при отборе кандидатов приходится учитывать множество рисокв: материальные, профессиональные, социальные и другие. Для их снижения работодатели прибегают к проверке сведений, представленных кандидатом различными способами, а так же ищут дополнительную информацию о кандидате, в том числе и в социальных сетях.

Работа с социальными сетями может улучшить работы специалистов отдела кадров (HR- специалистов). Из социальных сетей можно извлечь знания как о профессиональных, так и личностных качествах соискателя на конкретную должность. Зачастую эта информации более подробна, чем резюме соискателя. Сейчас работа с социальными сетями у HR-специалистов

– это ручной труд, который требует больших затрат времени и имеет ограничения по объему информации, которые можно обработать.

Наличие моделей и методов формирования и описания социального портрета соискателя позволит компания-нанимателям получить объективное представление о различных качествах соискателя: личностных, психических деловых, при этом данную информацию можно получить на основании семантико-когнитивного анализа профилей социальных сетей.

Существует целый ряд различных моделей личностных черт. Ряд исследований показал, что личностные черты могут выступать в качестве предикторов и коррелянтов различных психических отклонений [97]. Одной из таких моделей является модель «Большой пятерки». Ее, наряду с другими, можно использовать в части диагностики личностных и психических расстройств. Ряд исследователей считает, что данная модель может быть интегрирована в современные психиатрические модели [96, 66, 60].

Анализ больших данных социальных сетей – это возможность исследования личностных характеристик, построение и проверка предсказательных моделей о личностных чертах и поведении людей. Сбор данных может быть осуществлен как онлайн, так и офлайн. Такая процедура сбора данных позволяет значительно увеличить размеры выборки [38]. Однако для анализа этих данных необходимо использовать разные методы в зависимости от языка [62].

На основании вышеизложенного можно сделать вывод о том, что исследования в области обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя являются важной и актуальной задачей.

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.

Научная новизна

Научная новизна результатов исследования заключается в следующем:

- 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.
- 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.
- Разработан метод определения психологических характеристик пользователя социальных сетей, отличающийся гибридизацией алгоритмов обработки естественного языка текстовых данных, машинного обучения и метода «Большой пятерки».
- Предложен алгоритм анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей, отличающийся интеграцией семантических подходов и методов машинного обучения.

Достоверность результатов диссертационной работы

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

Теоретическая значимость диссертационной работы

Теоретическая значимость диссертационной работы заключается в

разработке новых алгоритмов, подходов и методов обработки текстовой информации социальных сетей для решения задачи подбора персонала.

Практическая значимость диссертационной работы

Практическая значимость диссертационной работы заключается в разработке программного комплекса, позволяющего упростить подбор персонала в организации посредством разработки социального портрета, полученного путем анализа профилей человека в социальных сетях.

Основные положения, выносимые на защиту

- Разработанный алгоритм формирования обучающей выборки позволяет эффективно решать задачу обучения нейронной сети в процессе сентимент-анализа русскоязычных текстов социальных сетей;
- 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;
- 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;
- 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.

Соответствие паспорту специальности

Область исследования соответствует паспорту специальности 05.13.01 «Системный анализ, управление и обработка информации (информационные технологии и промышленность)», 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.

Реализация и внерение результатов работы

Основные теоретические и практические результаты диссертационной работы были использованы в рамках проекта «Интеллектуальная платформа формирования социального портрета соискателя на основании семантико-когнитивного анализа профилей в социальных сетях», поддержанного Фонда содействия инновациям по программе «Старт-Цифровые технологии» для компании ООО «Центр программной инженерии и аналитики «ФаззиЛаб».

Кроме того, разработанные алгоритмы и подходы были применены УОСОО «Федерация бадминтона» в рамках проекта «Парабадминтон: все силы – для победы», поддержанного Фондом Президентских грантов для отбора волонтеров, обеспечивающих сопровождение лиц с ПОДА (проект № 18-2-009220).

Также предложенные в рамках диссертационной работы алгоритмы и методы интеллектуального анализа неструктурированных данных социальных сетей были использованы при разработке системы интеллектуального поиска и анализа в Интернет-СМИ и социальных сетях в рамках совместного проекта с ФНПЦ АО «НПО «Марс».

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.

Структура и объем диссертации

Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы и приложений. Текст работы изложен на 166 страницах, включает в себя 50 рисунков, 11 таблиц. Список использованной литературы – 170 наименований.

В приложение входят:

А. Блок-схема процесса подбора персонала.

Б. Таблица сравнения характеристик систем анализа социальных сетей.

В. Акты внедрения.

Г. Свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад

Представленные в данной работе результаты получены автором самостоятельно. Подготовка к публикации некоторых результатов проводилась совместно с соавторами, причем вклад диссертанта был определяющим.

Глава 1. Анализ современного состояния моделей и методов интеллектуального анализа текстовых данных социальных сетей

1.1. Процесс подбора персонала

Подбор персонала в современном мире постоянный процесс, в котором задействованы миллионы человека. Это деятельность, выполняемая профессиональными подборщиками, направлена на поиск и подбор кандидатов для выполнения каких-либо задач или на вакантные места в организации. Эта деятельность может осуществляться штатными сотрудниками организации или внешними специалистами по подбору персонала, что используется, например, в малом бизнесе.

Первым этапом подбора персонала является анализ заявки. Если заказчиком выступает внешняя организация, то необходимо провести интервью с заказчиками для понимания: персонал какого типа и характеристик требуется данной организации. Если заказчиком является отдел организации, то аналогичное интервью или заявка по внутренней форме оформляется между отделами. Кроме того, на этом этапе определяется метод поиска персонала – будет поиск активным или пассивным. Активный поиск предполагает обзванивание кандидатов отделом подбора персонала, пассивный – размещение вакансии с ожиданием отклика со стороны кандидатов.

Вторым этапом происходит первоначальный отбор кандидатов по резюме. Обычно при этом задействовано большое количество персонала, которое просматривает и оценивает резюме и анкеты в ручном режиме. При этом отбираются кандидаты, подходящие под установленные на прошлом этапе рамки. Данный процесс можно частично автоматизировать, применив определенные правила фильтрации.

Процесс подбора персонала может быть представлен в виде блок-

схемы. Упрощённая блок-схема представлена на рисунке 1.1. Полная блок-схема представлена в приложении А.

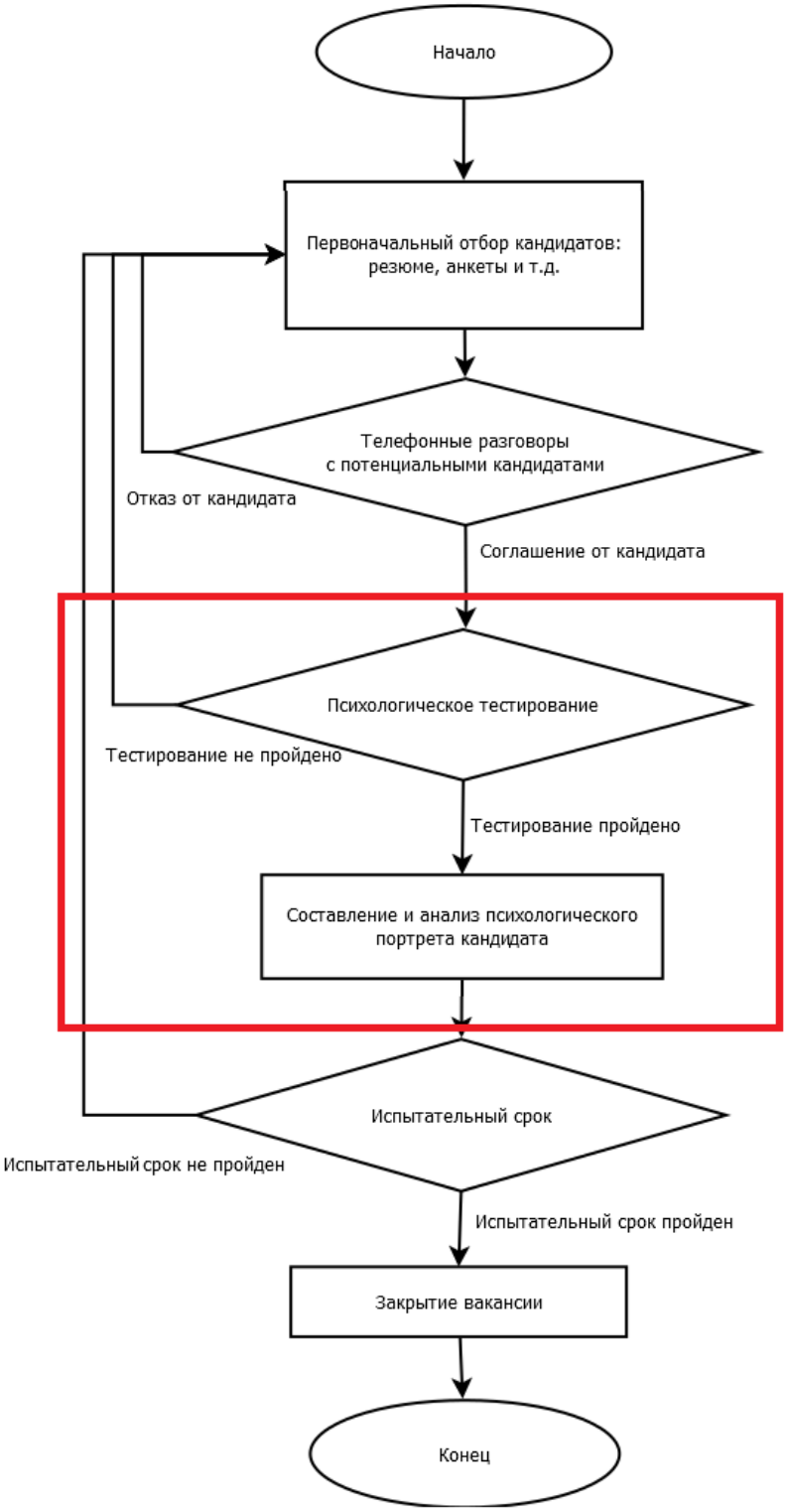


Рисунок 1.1. Блок-схема процесса подбора персонала.

Кандидаты, прошедшие первоначальный отбор, третьим и четвертым этапами приглашаются на первичное собеседование и, если оно пройдено, отправляются на психологическое тестирование. Психологическое

тестирование направлено на проверку эмоциональной стабильности кандидата, предрасположенности к обучению и другие психологические аспекты, которые могут быть важны для работодателя.

Психологическое тестирование может быть автоматизировано посредством проверки анкет, заполняемых кандидатами. Анкета также может быть заменена на ручную проверку постов кандидата в социальных сетях. Некоторые работодатели ставят особое условие – не рассматривать кандидатов, ведущих скрытный образ «жизни в сети» [162]. С другой стороны, при подборе в силовые структуры, отсутствие страниц в социальных сетях рассматривается как положительная сторона кандидата.

Если психологичекое тестирование успешно пройдено, то кандидату предлагается подготовить документы для приема на работу и пройти испытательный срок. При успешном прохождении испытательного срока вакансия считается закрытой.

1.2. Объединение профилей пользователя социальных сетей

1.2.1. Обзор существующих систем анализа социальных сетей

Современные социальные сети являются практически безграничным источником как личных данных пользователей, так и данных о пользовательских взаимодействиях, интересах, сообществах и многом другом. Обилие таких данных открывает новые возможности для анализа и структурирования информации, полученной из сети, с целью извлечения новых знаний.

Многие бизнес-задачи, которые ранее было невозможно решить из-за недостатка данных, теперь могут быть решены с помощью анализа социальных сетей. Повышенный интерес к данной тематике проявляют как исследовательские центры, так и различные компании по всему миру. Они

используют данные социальных сетей для моделирования экономических, социальных, политических и других процессов различного уровня с целью разработки механизмов воздействия на них [137]. Основными задачами систем анализа данных в социальных сетях являются анализ происходящих процессов, мониторинг, прогнозирование и управление.

Анализ может строиться различными способами и обычно делится на несколько частей. Первым делом извлекаются количественные характеристики, например, количество заметок, фотографий и так далее. Затем из количественных характеристик пытаются извлечь некоторые закономерности и построить необходимые математические модели для их описания. Например, это может быть распределение суточной активности пользователя.

Мониторинг включает сбор и структурирование различной информации: фотографии, заметки, сообщения, связи между пользователями, сообщества, контакты, личная информация. Способности систем во многом зависят от того, какое количество информации они имеют и каким образом они ее получают: в режиме реального времени или с использованием некоторой индексации ресурсов сети. Стоит заметить, что первые системы зачастую сложнее в реализации, в то время, как вторые, позволяют быстрее обрабатывать требуемые запросы, но в таких системах многое в том числе зависит от качества и количества индексаций ресурсов.

Некоторые системы позволяют на основе созданных математических моделей строить прогнозы относительно поведения пользователей в сети. Такая функциональность зачастую может быть востребована с целью определения спроса на определенные товары и услуги у некоторых групп пользователей.

Управление заключается в том или ином воздействии на поведение в социальной сети с целью достижения поставленных целей. Возможны, как простые рекомендации, так и конкретные количественные установки в соответствии с разработанными моделями. Этот механизм может использоваться уже, следуя примеру предыдущего абзаца, для увеличения

пользовательского охвата или конверсии охвата в покупки для какого-то бренда.

В анализе социальных сетей могут быть заинтересованы различные структуры такие, как органы государственной власти различных уровней, коммерческие и некоммерческие организации, средства массовой информации и физические лица. В зависимости от заказчика, могут использоваться различные варианты и способы мониторинга, анализа и пр. [118].

На текущий момент существует достаточно большое количество систем анализа социальных сетей, реализующих те или иные этапы с помощью различных алгоритмов.

Широко представлены различные массовые системы анализа, реализующие в основном этапы мониторинга и анализа текстовой информации. Примерами таких систем являются:

- people.yandex.ru – сервис Яндекса по поиску страниц пользователей в различных социальных сетях;
- blogsearch.google.com – поиск в блогах от Google;
- Google Trends – агрегатор информации о поисковых запросах.

Можно привести еще множество подобных систем, однако, всех их объединяют общие черты. Системы являются удобными и не требующими особых навыков от конечного пользователя для получения некоторых обобщенных результатов мониторинга или анализа активности в социальных сетях.

Достоинствами систем данного типа можно считать целенаправленность анализа социальных сетей, который в основном связан с продвижением брендов или продуктов, а также взаимодействием с потребителем. Однако, хотя и данные системы могут включать некоторые методы анализа конкретных пользователей, это не является их основной задачей. Примерами данных систем являются: Social Studio, ALTERIAN REAL-TIME CX, youscan.io.

Также существуют системы с более персонализированным подходом,

которые зачастую применяются в государственных, общественных и силовых структурах. Можно выделить следующие основные цели данных систем:

- Обнаружение, предупреждение и предотвращение информационных атак;
- Обнаружение и отслеживание злоумышленников, а также их сообществ;
- Определение значимости тех или иных событий;
- Политическая повестка;
- Оценка общественного мнения;
- Взаимодействие государства и гражданского общества.

Примерами систем данного класса являются: RecordedFuture, информационно-аналитическая система «Призма», Palantir.

Отличительные характеристики систем разных классов с использованием предложенной классификации на примере конкретных представителей, а также характеристики разрабатываемой системы представлены в виде таблицы в приложении Б.

На основании проведенного сравнения можно сделать вывод, что характерные представители различных классов делают упор на анализ сети в целом, а также анализ информационной повестки, мнений и упоминаний. В то же время данные системы уделяют меньше внимания конкретным пользователям сети, ограничиваясь необходимыми метриками для анализа продвижения брендов. Это обусловлено направленностью данных систем на потребности широкого круга заказчиков.

1.2.2. Идентификация пользователя в различных социальных сетях

Зачастую один и тот же пользователь имеет страницы на различных ресурсах, при этом размещаемая там информация может как отличаться, так и быть продублированной. Для построения наиболее полного

пользовательского портрета требуется как можно больше информации, которая может быть разбросана по страницам различных сетей. Однако, не существует точного алгоритма нахождения пользовательских страниц в других сетях, если не указаны прямые ссылки самим пользователем.

В работах [10, 105, 54, 68, 90] проводятся исследования на основании данных социальных сетей «MySpace» и «StudiVZ». Предложенный авторами подход основывается на построении векторов признаков пользовательских профилей, а затем данные вектора сравниваются различными способами. Авторами были разработаны алгоритмы, показавшие точность около 80% на отобранной выборке аккаунтов. Однако, данные сети практически не распространены в русскоязычном интернете.

В публикациях [68, 90] описаны методы сопоставления пользователей на основании текстовой информации, публикуемой на страницах пользователей.

В работе [68] авторы утверждают, что пользователя можно идентифицировать по уникальному стилю письменной речи. В работе [90] используются не только текстовая информации, но и связанная с ней метаинформации: время, геолокация, теги, специальные знаки и др.

В статье [136] предложена методика идентификации аккаунтов в социальных сетях «ВКонтакте» и «Одноклассники». Авторы предлагают использовать для идентификации анкетные данные пользователей, а также способы предобработки и унификации полученных данных. Однако, авторами не предоставлены результаты применения данной методики на реальной системе.

Исходя из этого, можно сделать вывод, что проблема идентификации пользователя на различных ресурсах не решена в должной мере. Решение такой проблемы позволило бы повысить точность систем анализа информации в социальных сетях, а также многих других агрегаторов информации.

Возможность определения личностных характеристик пользователя на основе слабоструктурированной информации рассматривалась в

различных научных работах. В статье [100] авторы провели большое исследование личностей авторов различных блогов, предлагая пройти им психологический опрос. Было показано, что использование некоторых слов может быть связано с личностными характеристиками автора.

В работе [36] было доказано, что используемые слова и структура текста могут отражать те или иные черты личности автора блога.

В исследовании [58] авторы опросили 71 пользователя, имеющих страницы в социальных сетях, а также провели анализ текстовых заметок этих пользователей. При помощи метода SVM и выделения N-грамм из текстов авторы показали возможность определения личностных характеристик.

Авторы работы [31] проанализировали пользователей социальной сети «Twitter» и показали, что можно оценить личность автора по их текстовым заметкам, а также с помощью дополнительной информации, такой как количество слов в сообщении, количество подписчиков страницы и т.д.

В научной работе [83] авторы проанализировали пользователей социальной сети «Facebook». Ста пользователям было предложено пройти опрос на основе пятифакторной модели личности. Далее с помощью Facebook API была получена информация со страниц этих пользователей. На основе методов интеллектуального анализа данных авторам удалось достигнуть точности 82,2% при определении личностных характеристик пользователей.

Также проводились исследования по анализу не только текстовой информации, но и изображений. В работах [84, 21, 76, 77] авторы показали возможность применения новейших методов интеллектуального анализа данных при работе с изображениями с целью получения личностных характеристик пользователя.

Большинство работ в этой области проводится на основе англоязычной текстовой информации, тем не менее, существуют работы [62, 38] по анализу данных из социальных сетей, популярных в русскоязычных сообществах.

На основании данных из вышеупомянутых исследований в рамках диссертационной работы было решено использовать пятифакторный метод оценки личности, также называемый «Большая пятерка» [167].

1.3. Методы формирования обучающей выборки для сентимент-анализа текста

В настоящее время нейросетевой подход применяют для обработки естественного языка, генерации текстов, классификации текстов и т.д. Разработку нейронной сети можно разделить на следующие этапы:

- выбор архитектуры нейронной сети;
- выбор типа формирования обучающей выборки;
- обучение нейронной сети.

Важно заметить, что этап формирования обучающей выборки может занять большой промежуток времени, так как выборка в большинстве случаев формируется экспертом в ручном режиме. Задача эксперта заключается в анализе текстов и классификации их по нескольким категориям.

На сегодняшний день существует множество различных способов формирования обучающих выборок. Подходы, предложенные как российскими, так и зарубежными исследователями, будут описаны далее.

1.3.1. Группы методов оценки тональности текстовых данных

Можно выделить две основные группы методов, которые применяются при решении задачи определения тональности:

Статистические методы, в основе которых лежит создание машинного классификатора обучающегося на заранее размеченных текстах. Краткий алгоритм состоит в следующем:

- Собирать некоторый набор текстов для обучения;
- Преобразование текста в набор векторов признаков, по которым

он будет исследоваться;

- Указание типа тональности для каждого текста обучающей выборки;
- Выбор алгоритма классификации и метод обучения классификатора;
- Использование полученной модели для определения тональности текстов, не входящих в обучающую выборку.

Недостатком подобных методов является потребность в большом количестве данных для обучения.

Анализ литературы показал, что при статистическом подходе для решения задачи определения тональности текстов часто используются методы: опорных векторов (SVM) [72], Байесовы модели [23], регрессии [20], методы Word2Vec, Doc2Vec [19], CRF [117], а также нейронные сети, при этом лучшие результаты показали сверточные и рекуррентные нейронные сети [43].

При использовании методов, основанных на словарях, составляются специализированные словари тональности – тональные словари. Эти словари используеются для создания лингвистических правил, по которым производится поиск тональной лексики, которая после оценкивается по шкале негативной и позитивной лексики [75]. Использование методов, основанных на словарях, предполагают наличие эксперта в области лингвистики, задачей которого будет составить наиболее словарь эмоционально окрашенных слов и выражений, называемых маркерами. При нахождении маркера эмоция учитывается по заданному алгоритму, итогом выполнения которого является эмоциональная окраска текста [61,161].

Краткий алгоритм выглядит следующим образом:

- Присвоить каждому слову в тексте значение тональности из словаря;
- Вычислить общую тональность целого текста, путем суммирования тональностей отдельных слов [132].

Недостатками данного метода являются необходимость в привлечении эксперта и большой объем трудозатрат на создание правил.

Ряд авторв также использует смешанный метод, которые является комбинацией статистических и словарных методов [64].

1.3.2. Математическое планирование

Идея данного подхода состоит в уменьшении объема обучающей выборки до минимума, при условии, что выборка будет удовлетворять всем необходимым условиям. Чтобы уменьшить количество обучающих данных, предлагается генерировать каждую пару с помощью математической модели, которая на определенный вход выдает определенный выход [118].

Модель позволяет сгенерировать обучающую выборку достаточного объема. Модель представляет собой нейронную сеть, которая на определенный вход дает определенный результат. В работе вводится такое понятие как ошибка обобщения, означающее отклик нейронной сети на входное значение, которое не было в обучающей выборке. Для ее получения проводится анализ части примеров из имеющейся базы данных, для которых известны отклики системы, но которые не использовались при обучении.

1.3.3. Программная генерация

Идея подхода программной генерации состоит в формировании обучающей выборки по заранее известному алгоритму, на выходе которого будет множество объектов. Объекты распределены в пространстве на основе признаков. Как правило, в процессе формирования обучающей выборки варьируются все параметры. Данный метод малоэффективный, так как выборка получается очень большой по объему [118].

1.3.4. Применение сэмплирования

Использование сэмплирования исключает недостатки программной генерации обучающей выборки. На первом этапе все объекты представляются в виде распределения в пространстве. На втором этапе

алгоритм генерирует обучающую выборку на основе построенного распределения. В результате алгоритм выбирает объекты из скоплений, в которых находятся объекты близкие по признакам. Из каждого скопления берутся несколько объектов, соответственно, получается выборка меньшего размера [118].

1.3.5. Закономерная модификация базового объекта

Идея алгоритма модификация базового объекта заключается в последовательном изменении базового набора объектов. Варьируется все параметры. Таким образом, на выходе получается множество случайных объектов. Данные метод применяют, когда нет возможности построить распределение объектов в пространстве, описанном в предыдущем пункте [118].

1.3.6. Выборка из базы объектов

Идея подхода выборки из базы объектов заключается в разбиении объектов на группы по некоторым признакам. Как правило, объекты внутри группы похожи, а объекты из разных групп различаются. Такой подход применяют, когда количество объектов ограничено и нет объектов с указанными параметрами. Тогда будут возвращены объекты, наиболее похожие на запрошенный [118].

1.3.7. Вероятностные методы

Вероятностные методы предполагают случайное извлечение набора экземпляров объектов из исходной выборки. В выборке каждый объект имеет некоторую вероятность, что он принадлежит конкретной группе [160].

При простом случайном отборе обучающая выборка формируется на основе случайного выбора определенного числа объектов из некоторого множества.

При использовании систематического отбора исходное множество объектов упорядочивается по определенным правилам и затем разбивается на

несколько групп. После этого выбираются объекты с заданными порядковыми номерами в группе и добавляются в обучающую выборку.

При использовании стратифицированного отбора исходное множество разбивается на несколько подмножеств. Затем для каждого подмножества применяется случайный или систематический отбор

Достоинствами вероятностных методов является их простота реализации и наличие возможности оценить ошибку выборки по отношению к исходной совокупности.

Главным недостатком вероятностных методов является то, что они не гарантируют, что сформированная обучающая выборка будет в полной мере отражать свойства исходного множества.

1.3.8. Детерминированные методы

Детерминированные методы формирования выборок предполагают формирование выборки на основе предположений о полезности экземпляров исходной выборки [160].

При использовании удобного отбора формируется выборка очень низкого качества из наиболее доступных для исследования объектов исходного множества.

При использовании квотного отбора исходное множество разбивается на несколько групп, которые различны по свойствам. После этого происходит пропорциональный выбор объектов из каждой группы.

При целевом отборе обучающая выборка формируется на основе мнения исследователя, который определяет объекты, пригодные для исследования.

Главным недостатком детерминированных методов является отсутствие возможности оценивания ошибки сформированной выборки по отношению к исходной совокупности. Главным достоинством детерминированных методов является то, что они могут выявить наиболее значимые объекты для решения поставленной задачи.

1.3.9. Методика формирования обучающего множества с использованием меры схожести и матрицы схожести

Помимо этого, применяется метод формирования обучающего множества с использованием меры схожести и матрицы схожести при антивирусном эвристическом анализе файлов [128].

При решении задачи формирования максимально разнообразной обучающей выборки вводится мера схожести файлов, с помощью которого появляется возможность сравнения файлов. Файл – это последовательность байт. Таким образом, введенная мера схожести позволяет построить матрицу схожести для множества файлов.

Матрица схожести состоит из набора мер схожести для набора файлов. На этапе обучения классификатора данная матрица схожести может быть использована для формирования разнообразной обучающей выборки, состоящей из файлов.

В работе [128] описан классификатор, предназначенный для решения задачи анализа файлов на вирусы. Классификатор был обучен на выборке, содержащей около двух тысяч файлов.

Обучение классификатора проводилось на двух выборках. Первая выборка содержала файлы без вирусов, второй обучающий набор содержал файлы без вирусов и несколько файлов с вирусами. При проведении экспериментов было подтверждено, что качество обнаружения вирусов повысилось на 6%. Количество ложных срабатываний на файлах без вирусов увеличилось на 1%.

1.3.10. Проблемы формирования обучающей выборки

Из рассмотренных ранее подходов к формированию обучающей выборки видно, что задача не является сложной, но в процессе реализации возникают проблемы.

Первая проблема заключается в том, что может быть множество очень похожих объектов. Вторая проблема заключается в том, что в процессе

формирования выборки некоторые признаки исходной совокупности могут быть пропущены. Так же в самом алгоритме формирования обучающей выборки могут быть ошибки.

Третья проблема состоит в том, что при формировании обучающей выборки могут отсутствовать объекты определенного типа, следовательно, определенная область пространства не будет покрыта. А значит, классификатор не сможет правильно обучиться и классифицировать объекты данного типа.

При формировании обучающей выборки может возникнуть разбалансировка. Это значит, что данных одного вида в несколько раз больше, чем данных другого вида. Соответственно наблюдается ситуация завышения влияния на результат определенного класса. При обучении будут возникать неверные связи, из-за которых классификатор будет давать не правильную оценку.

При создании выборки с помощью варьирования различных параметров может возникнуть следующая проблема. Если задать слишком малый диапазон значений, то разброс значений будет минимальный и выборка получится однородной. Тогда классификатору будет сложно отличить объекты друг от друга.

1.3.11. Семантические тезаурусы

В настоящее время создаются и поддерживаются различные тезаурусы – словари, слова в которых размечены в соответствии с эмоциональной принадлежностью. Такие словари имеют семантическую или графовую структуру, что позволяет отследить связи между словами, их принадлежность к какой-либо группе и тд. Такие тезаурусы необходимы компьютерным программам при обработке естественного языка или при анализе тональности текста.

Наиболее популярным семантическим тезаурусом является тезаурус «**WordNet**», разработанный для английского языка и считающийся самым качественным. Для русского языка тоже существует ряд тезаурусов,

некоторые из них наследуют структуру WordNet для удобной работы, другие имеют свою структуру и свои метки.

WordNet-Affect – семантический тезаурус, в котором каждому слово сопоставляется эмоциональная метка. Так же в WordNet-Affect используются дополнительные эмоциональные метки для четырех категорий: позитивная, негативная, неоднозначная и нейтральная [99].

Словарь содержит слова четырех частей речи: существительные, глаголы, прилагательные и наречия. Основой WordNet-Affect является не отдельное слово, а синсет – синонимичный ряд.

WordNet-Affect представляет собой электронный словарь-тезаурус и набор семантических сетей только для английского языка.

SentiWordNet – это лексический семантический тезаурус, полученный в результате автоматического аннотирования набора синонимов в соответствии с его степенью позитивности, негативности и объективности [80]. Тезаурус SentiWordNet используется для обработки только английского языка.

SenticNet – семантический тезаурус, в основе которого лежат наборы эмоциональных понятий [78]. SenticNet применяют при разработке систем анализа тональности текстов. SenticNet в отличие от SentiWordNet и WordNet-Affect позволяет выявить смысловую составляющую.

Тезаурус SenticNet представлен в виде интернет сервиса с открытым API для взаимодействия. Данный тезаурус поддерживает множество языков, в том числе и русский.

Лингвистическая онтология **РуТез** – русскоязычный тезаурус, состоящий из иерархии понятий. В РуТез используются специальные типы отношений между объектами. РуТез создан для автоматической обработки текстов [141].

Объем тезауруса РуТез составляет около 160 тысяч слов и выражений, между которыми вручную установлено более 200 тысяч связей. Особенностью тезауруса является то, что он используется в реальных проектах на протяжении многих лет.

Тезаурус русского языка **RuWordNet** был создан на основе автоматизированной трансформации тезауруса РуТез в формат WordNet – наиболее популярный формат в области обработки естественного языка [163].

Тезаурус RuWordNet состоит из синсетов для 3 частей речи: существительные, глаголы и прилагательные. RuWordNet содержит более 120 тысяч слов и выражений на русском языке.

РуСентиЛекс – словарь оценочных слов и выражений русского языка, каждая запись которого ссылается на понятия РуТез. Структура РуСентиЛекс – это упорядоченный по алфавиту список тональных слов и выражений [157], каждая из которых имеет три поля:

- полярность слова, т.е. какую оценку несет слово – позитивную или нейтральную;
- источник тональности;
- тональные различия (при условии, что слово многозначное).

Текущая версия словаря распространяется авторами свободно и содержит окло двенадцати тысяч слов и выражений.

1.4. Методы сентимент-анализа текста

Анализ тональности текста – задача компьютерной лингвистики, заключающаяся в определении категории текста или документа на основе его содержимого.

В последнее время данная предметная область развивается очень быстро, а интерес исследователей стремительно растет. Сентимент-анализ русского языка является непростой задачей, по сравнению с английским языком. Одной из сложностей обработки русского языка является свободный порядок слов, когда в английском языке порядок слов фиксированный.

Так же в русском языке большой словарный запас, по сравнению с английским, что делает задачу составления словарей и тезаурусов достаточно трудоемкой.

1.4.1. Подходы к определению эмоциональной окраски текстов на русском языке

Анализ текста методом векторного анализа. Данный метод работает достаточно быстро, и требует предварительно размеченного корпуса слов, на основе которого происходит обучение алгоритма сравнения. Другими словами, каждое предложение преобразуется в вектор на основе корпуса слов и далее происходит обучение модели [169].

Главными недостатками такого подхода является увеличение трудоемкости и ограничение разнородности корпуса, что приводит к потере точности. К тому же данный метод не позволяет провести глубокий анализ текста, то есть определить эмоциональную окраску текста на уровне предложения.

Тональные словари. Данный метод не менее трудоемкий, чем предыдущий, но в сочетании с синтаксическим и морфологическим анализом более гибок. Данный подход позволяет показать цепочки тональной лексики и получить эмоциональные выражения. При хорошем наполнении тональных словарей этот метод позволяет достичь хорошего результата оценивания эмоциональной окраски.

Недостаток этого метода в том, что с помощью него сложно дать количественную оценку эмоциональной окраски текста.

Чтобы избежать недостатков предыдущего и текущего метода, используют смешанный подход, частично включающий в себя два первых.

Подход, основанный на правилах, использует набор правил, которые были составлены экспертами для конкретной предметной области. При анализе из текста извлекаются биграммы или триграммы. Данный подход сопоставим по трудоемкости с подходом, основанном на тональном словаре.

При обучении модели с учителем модель обучают на размеченном корпусе смайлов. Каждый текст в обучающем наборе представляют в виде пары – вектора признаков и тональности. Далее происходит обучение модели для дальнейшей классификации текстов.

При обучении модели без учителя используется неразмеченная обучающая выборка. При таком подходе наибольший вес будет у самых часто встречающихся слов, наименьший вес – у редко встречающихся слов. Классическим примером обучения без учителя является автоматическая кластеризация документов.

1.4.2. Применение сентимент-анализа текстов для оценки общественного мнения

В работе [152] описывается прототип, предназначенный для определения тональности текстов. Процесс обработки состоит из нескольких этапов. На первом этапе производится токенизация – разбиение текста на слова. Далее выполняется лемматизация каждого слова и определение части речи. После этого производится поиск слов в тональном словаре.

В данной работе задействован тональный словарь русского языка, состоящий из трицати пяти тысяч слов. Каждое слово имеет свою тональную оценку: крайне отрицательный, отрицательный, нейтральный, положительный, крайне положительный. Если в словаре отсутствует определенное слово, то его тональность – нейтральная. На последнем этапе происходит вычисление общей тональности текста.

1.4.3. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики

В работе [138] описывается прототип системы определения тональности текстов на основе тонального словаря, составленного для конкретной предметной области.

Тональный словарь содержит около 260 ключевых слов, каждому из которых поставлено значение от -5 до +5, для отрицательных и положительных слов соответственно.

В данной работе для определения тональности текста использовались оценочные слова. В каждом тексте производился поиск оценочных слов, для которых в тональном словаре есть оценка. После вычислялся средний вес

всех оценочных слов.

Чтобы повысить точность, из этого множества исключались те слова, которые находились дальше всех от кластеров положительных или отрицательных слов. После исключения вычислялась общая тональность.

Если общая оценка положительная – то эмоциональная окраска положительная. Если общая оценка отрицательная – то эмоциональная окраска отрицательная.

1.4.4. Анализ тональности текста на русском языке при помощи графовых моделей

В работе [143] описывается метод определения тональности текста с помощью графовых моделей.

На первом этапе строится взвешенный граф, где вершины – это слова в тексте, а ребра – расстояния между словами, получаемые путем итерационного перебора слов в тексте [165].

На втором этапе выполняется ранжирование вершин, то есть определение ранга вершины с учетом связей. Далее выполняется определение классов найденных слов на основе тонального словаря. В тональном словаре 2 типа тональности: положительная и отрицательная. Сила тональности определялась от 1 до 5.

Далее оценивался итоговый результат по полученной положительной оценке и полученной отрицательной оценке. Положительная или отрицательная тональности вычислялись путем суммирования произведений оценки составляющей текста и силы его тональности. Итоговая тональность вычислялась как отношение этих величин.

Нейтральными считаются те тексты, у которых итоговое значение около единицы. Если итоговое значение больше единицы, то текст считается положительным, если меньше единицы, то отрицательным.

1.4.5. Сентимент-анализ коротких русскоязычных текстов в социальных медиа

В работе [121] проводилась работа с социальной сетью Twitter. Для разметки текста был составлен корпус смайлов для двух категорий – положительный и отрицательной. Если пост содержит негативный смайл – то он относится к негативной группе, и наоборот, если содержит позитивный смайл – то относится к позитивной группе.

Далее было загружено около 113 тысяч записей, из которых около 47 тысяч негативных записей и около 66 тысяч – позитивных. При скачивании постов сразу происходила разметка текста на основе корпуса смайлов.

Далее для текстов была выполнена токенизация – разбиение на отдельные слова. Затем все слова были приведены к нижнему регистру и затем преобразованы в векторную форму с помощью алгоритма «мешок слов».

В качестве классификаторов использовались 3 модели: логистическая регрессия, дерево решений, многослойный персептрон. Наивысшая точность классификации была у многослойного персептрона – 76%.

1.4.6. Лексико-грамматические маркеры эмоций в качестве параметров для сентимент-анализа русскоязычных интернет-текстов

В работе [135] описывается прототип, предназначенный для классификации текстов по 8 группам эмоциональной принадлежности.

Материалом для формирования обучающей выборки для классификатора послужили анонимные текстовые записи в жанре «интернет-откровения» пользователей в социальной сети «ВКонтакте».

В основе классификатора лежит метод опорных векторов. На вход классификатору подаются различные характеристики текста: частота слов, знаков препинания, наречий. На выходе получается одна из 8 групп эмоциональной принадлежности. В результате экспериментов была определена точность классификации. Точность определения эмоциональной окраски эмоций не более 48%.

1.4.7. Выбор топологии нейронных сетей и их применение для классификации коротких текстов

В работе [158] рассматривается классификация текстов как подход, применяющийся для определения тональности текстовых сообщений социальных сетей. Сентимент-анализа построен на двух моделях нейронных сетей, построенных при помощи фреймворка Python Keras:

- рекуррентный слой и сверточный слой,
- два рекуррентных слоя.

Для обучения была сформирована обучающая выборка. Тесты были размечены на 2 группы вручную. Тексты содержали посты до 140 символов. В обучающей выборке было около 20 тысяч постов. Перед обучением модели была проведена лемматизация всех текстов, а также удаление стоп слов. Далее выборка была разделена на обучающую и тестовую. Точность классификации при использовании первой модели составила 71%, а второй – 69%.

1.4.8. Проблемы определения тональности текста

При разработке подхода к сентимент-анализу текстов возникает множество проблем [169], которые представлены ниже.

Тональность текста зависит от конкретной предметной области. Например, если слово «большой» описывает монитор, то это положительная тональность. А если слово «большой» описывает телефон, то это уже отрицательная тональность.

Сентимент-анализ плохо справляется с отрицаниями. Например, в предложении «Мне нравится ездить на природу, там свежий воздух и цветы. К сожалению, у меня аллергия» много положительных слов. Использование отрицания в конце предложения полностью меняет окраску текста на противоположную. Также сентимент-анализ плохо справляется с текстами, в которых присутствует сарказм.

1.4.9. Существующие системы определения тональности

текста

В ходе диссертационного исследования были рассмотрены системы и модули, предназначенные для сентимент-анализа текстов. Далее будут рассмотрены самые популярные системы или компоненты для сентимент-анализа. Некоторые из описанных систем распространяются бесплатно, а другие являются коммерческими [144].

Eureka Engine – система, предназначенная для лингвистического анализа текстов. Система позволяет новые знания из данных больших объемов. Система является коммерческой [25].

Система включает в себя различные модули, основными из которых считаются: модуль определения языка сообщения, модуль автоматического определения тональности документа, модуль морфологического анализа.

Система определяет три вида тональности – положительная, отрицательная и нейтральная. Система может определять общую тональность для предложения или всего документа. Объектом тональности может быть как отдельное слово, так и словосочетание. Средняя точность определения тональности около 86%.

Система **«SentiStrength».** Система используется для сентимент-анализа неструктурированных текстов на английском языке. В системе имеется возможность использовать ее для других языков, в том числе и для русского [79].

Результат анализа получается в виде двух оценок. Оценка позитивной составляющей текста от +1 до +5. Оценка негативной составляющей от -1 до -5. Так же есть возможность выбрать иной вид оценки: бинарную (позитивный или негативный), тернарную (позитивный, негативный или нейтральный) и оценку по единой шкале (от -4 до +4).

Система может быть использована и для русского языка, но алгоритмы системы не учитывают специфику русского языка, что может привести к ряду проблем.

Система **«Аналитический курьер»** использует подход, основанный

на семантических словарях и правилах [156].

На вход системе подается текст, а на выходе получается размеченный текст. Каждое слово и словосочетание имеет эмоциональную метку. Так же подсчитывается общая тональность всего текста. На первом этапе выполняется предобработка текста и выделение значимых слов. На втором этапе выполняется объединение найденных слов в цепочки. И затем определение тональности.

Недостаток системы – система не определяет количественную оценку тональности. Система является коммерческой.

Проект «Ваал». Данная система разработана российскими разработчиками. Алгоритм работы системы следующий. На первом этапе составляется частотный словарь. На втором этапе происходит определение тональности для самых часто встречающихся слов. После этого вычисляется общая тональность текста [153].

На выходе генерируется отчет, который состоит из набора оценок для всего текста и для каждого слова по отдельности.

Недостаток системы в том, что система не анализирует семантическую составляющую текста из-за чего систему трудно применить.

Система **«Fact Extractor».** В текущей версии системы используется набор правил для определения тональности текста. В системе анализируется синтаксическая структура текста и связи между словами [70].

1.5. Построение психологического портрета человека на основе публичной информации социальных сетей

1.5.1. Психологическая классификация людей

Существует несколько видов классификаций людей по их психологическому портрету, однако самым популярным является разделение по типу темперамента. При этом каждый тип обладает своими чертами поведения, характера, суждениями, стремлениями и желаниями. Всего

существует четыре типа темперамента – холерики, меланхолики, сангвиники и флегматики, однако абсолютное большинство людей не являются ни одним из типов в крайней его мере. Так, например, если задать шкалу «эмоциональная нестабильность – «эмоциональная стабильность» от 1 до -1, и шкалу «экстраверсия – интроверсия» от 1 до -1, то человек, обладающий параметрами [0,3; -0,3] будет являться меланхоликом аналогично человеку с параметрами [1;-1], однако его реальный характер будет сильно отличаться от «эталонного» примера [162].

Рисунок 1.2. Компас темпераментов

В настоящее время популярность получила типология Майерс–Бриггс (Myers-Briggs Type Indicator), поскольку предоставляет более подробный портрет человека, позволяя глубже изучить и понять личность. По классификации этой типологии, любой человек может быть отнесен к одному из 16 типов, который зашифрован набором из 4 букв. Типы представлены в таблице 1.1

Таблица 1.1. Типология Майерс–Бриггс

ISTJ	Инспектор
ISTP	Изобретатель
ISFJ	Хранитель
ISFP	Посредник
INFJ	Гуманист

INFP	Романтик
INTJ	Аналитик
INTP	Критик
ESTP	Командир
ESTJ	Администратор
ESFP	Политик
ESFJ	Энтузиаст
ENFP	Инициатор
ENFJ	Наставник
ENTP	Новатор
ENTJ	Предприниматель

При этом каждый шифр можно расшифровать по 4 шкалам.

Шкала E–I (Экстраверсия – Интроверсия). Экстраверты – это люди, которые ориентированы на социальные взаимодействия, они «заряжаются энергией» после общения с другими людьми. Интроверты, наоборот, на размышления, эскапизм и «заряжаются энергией» после проведения времени в одиночестве.

Шкала S–N (Ощущение – Интуиция). Люди, относящиеся к категории «ощущение», уделяют больше внимания реальности. Решения, принимаемые такими людьми, как правило, основаны на фактах, деталях и практическом опыте. Те, кто относится к категории «интуиция», как правило, принимают решения, основываясь на внутреннем состоянии и впечатлениях. Им нравится думать о возможностях, воображать будущее и принимать решения, опираясь на абстрактные теории.

Шкала T–F (Мышление – Чувство) фокусируется на том, как люди принимают решения, основываясь на информации, полученной ими в результате своих ощущений или интуитивных функций. Категория «мышления» – те, кто предпочитает думать над фактами. Склонны быть последовательными, логичными и беспристрастным. Категория «чувства» чаще принимают решения, исходя из эмоций и мнения других людей.

Шкала J–P (Суждение – Восприятие) включает в себя то, как люди взаимодействуют с внешним миром. Те, кто склоняется к суждению,

предпочитают структуру и твердые решения. Люди, склонные к восприятию, более открытые, гибкие и адаптируемые.

Существует классификатор психотипов по речевому поведению человека. Его data set сформирован записями речи людей, переведенной в текстовый формат с помощью алгоритмов распознавания речи. Data set состоит из 16 каталогов по числу психотипов. Каждый каталог содержит корпусы текстов из социальных сетей, аудиозаписи речи знакомых людей, переведенные в текстовый формат. Значительная часть этого data set – записи социальных сетей, поскольку в социальных сетях люди общаются простым языком, больше похожим на реальную речь, а сообщения имеют небольшую длину, удобную для анализа. В данной системе применяется технология Text Mining, на основе которой формируется вектор параметров P. Вектор P представляет собой словарь часто встречаемых слов, построенный в соответствии с показателем TF-IDF.

Для автоматизации анализа записанной речи и дальнейшей классификации психотипа и, соответственно, построения психологического портрета человека применяется алгоритм глубокого обучения на базе нейронной сети LSTM. Параметры для алгоритма обучения были определены экспериментальным путем. Использование данного метода позволило достичь точности 83% при бинарной классификации [162].

1.5.2. Классификация полученных данных

На основании данных из вышеупомянутых исследований в текущей работе было решено использовать пятифакторный метод оценки личности, также называемый «Большая пятерка» [167]. Пятифакторная модель личности является диспозициональной моделью личности и включает в себя пять различных черт (шкал): нейротизм, экстраверсия, открытость опыту, добросовестность и сотрудничество.

Описание шкал оценки личностных характеристик согласно пятифакторному анализу данных представлено в таблице 1.2.

Таблица 1.2. Пятифакторный анализ личности

№	Черта характера	Уровень	Значение	Характеристика
1	Нейротизм	Высокий	Эмоциональность	Возможность испытания негативных состояний: страх, тревога, грусть
		Низкий	Спокойствие	Высокая стрессоустойчивость
2	Экстраверсия	Высокий	Энергичность	Энергичность, оптимистичность настроя
		Низкий	Сдержанность (интроверсия)	Эмоциональная сдержанность, склонность к одиночеству и независимости
3	Открытость опыту	Высокий	Оригинальность	Возможность подвергать сомнения устои, законы, нормы, принимать новые идеи
		Низкий	Практичность	Консервативность во взглядах
4	Добросовестность	Высокий	Контролирование	Пунктуальность в поведении, надежность
		Низкий	Импульсивность	Преобладание гедонистических устремлений
5	Сотрудничество	Высокий	Привязанность	Альтруизм и желание сотрудничать
		Низкий	Отделенность	Эгоцентризм в поведении, соперничество, жажда конкуренции

Задача психолингвистического анализа – это задача классификации. Формально данную задачу можно описать следующим образом. Пусть X – это некое множество описаний объектов, Y – это некое конечное множество классов. Существует неизвестное отображение $y^*: X \to Y$, значения которого известны только на объектах конечной обучающей выборки

$$X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}.$$

Необходимо построить алгоритм $a\colon X \to Y$, способный классифицировать произвольный объект $x \in X$.

Классификация текстов, как и многие другие задачи интеллектуального анализа данных, состоит из нескольких последовательных этапов. Эти этапы представлены на рисунке 1.3, рассмотрим их подробнее.

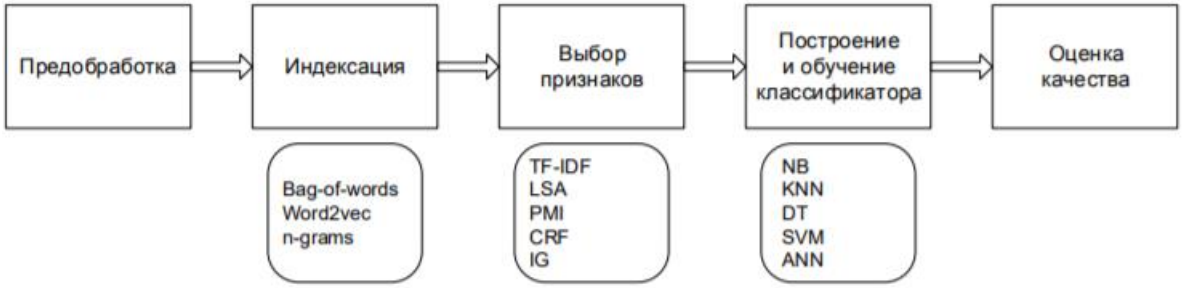


Рисунок 1.3. Этапы классификации текстов

Предварительная обработка текста необходима для выделения значимой информации из текста.

В первую очередь из текста выделяют слова (токены), то есть производят токенизацию [125]. Знаки препинания зачастую не несут важной информации об исходном тексте, поэтому от них избавляются.

Также избавляются от стоп-слов, в которые входят предлоги, союзы, местоимения и многие другие слова, также не несущие значимой смысловой нагрузки.

Обычно текст приводят к нижнему регистру, чтобы слова с буквами в разных регистрах не являлись различными.

Далее осуществляется морфологический анализ, приведение слов к начальной форме с помощью стемминга или лемматизации [111].

Стемминг является более простым и быстрым, но менее точным способом, так как слова с разным началом, но одинаковые по смыслу не могут быть корректно обработаны. Существует и более продвинутый подход лемматизации текста. Он основан на объединение языковых словарей и методов машинного обучения. В этом случае все слова, которые содержатся в документе, помечаются как признаки данного документа. В разработанном

алгоритме классификации применяются оба указанных метода.

Индексация электронных документов заключается в построении некоторой числовой модели текста, которую может обработать некоторое программное обеспечение. Одним из вариантов индексации является модель «мешок слов» [91], которая преобразовывает документ в многомерный вектор, содержащий все слова в документе и веса этих слов для данного документа. Модель индексации учета n-грамм [127] основана на учете последовательностей соседних слов. Моделью индексации Word2vec [71], в отличии от предыдущей модели, представляет каждое слово как вектор, который содержит информацию о контексте данного слова (путем ссылки на соседние слова). Важно учесть, что выбрать метод индексации нужно заранее и далее должен применяться один и тот же метод индексации, поскольку сменить его будет трудоемкой задачей.

Кроме индексации существует еще несколько способов определения веса признаков документа. Одним из наиболее простых и при этом качественных методов является статистический метод TF-IDF [69]. Основная идея метода – если слово встречается часто в одном документе, но редко во всем массиве документов, то слово имеет большой вес для данного документа. В методе считаются две величины – частота термина TF (term frequency) – встречаемость слова в пределах одного документа:

$$TF = n_{t,d} / n_d,$$

где $n_{t,\,d}$ – количество употреблений слова t в документе d; n_d – общее число слов в документе d. Второй величиной является обратная частота документа IDF (inverse document frequency) – инверсия встречаемости данного слова в текущем документ относительно всех документов коллекци. IDF уменьшает вес общеупотребительных слов по формуле

$$IDF = log(|D| / D_t),$$

где |D| – общее количество документов в коллекции; D_t – количество всех документов, в которых встречается слово t. Итоговый вес термина в документе относительно всей коллекции документов вычисляется по формуле

$$V_{t,d} = TF \cdot IDF.$$

Необходимо отметить, что данный метод никак не учитывает лексическую значимость термина, лексическую сочетаемость и порядок терминов в документе, учитывается только частоты встречаемости. Для учета лексических и семантических признаком термина может применяться латентно-семантический анализ (LSA), основанный на сингулярном разложении матриц [47].

Из других методов уменьшения размерности пространства признаков можно выделить поточечную взаимную информацию (PMI) [47, 67]) и условные случайные поля (CRF) [85].

Можно выделить следующие методы классификации:

- вероятностные (NB [18]);
- метрические (KNN [82]);
- логические (DT [73]);
- линейные (SVM [17]; логистическая регрессия [34]);
- методы на основе искусственных нейронных сетей (RNN [7], CNN [92]).

Рассмотрим алгоритмы, которые были выбраны в качестве классификаторов для решения задачи определения личностных характеристик пользователя социальных сетей. Основанием для выбора данных методов послужили успешные применения данных методов в работе [159].

Метод опорных векторов [17] – это метод линейной классификации, относящийся к классу алгоритмов обучения с учителем. Его можно применять для классификации какого-либо набора документов. Основная идея метода – это перевод существующих векторов, разделяющих точки текущего пространстава в пространство следующей размерности с последующим поиском гиперплоскости с наибольшим зазором. Выборка точек называется линейно разделимой, если точки, принадлежащие разным классам, можно разделить с помощью гиперплоскости (в двумерном случае

гиперплоскость представляет собой прямую линию). Любые точки, которые находятся с одной стороны от выбранной гиперплоскости, могут быть классифицированы как относящиеся или не относящиеся к классу. Однако таких гиперплоскостей в общем случае может быть проведено бесконечное множество.

Метод опорных векторов отличен тем, что в нем учитывается расстояние между самой ближней к точкам гиперплоскостью и набором этих точек, которое, в свою очередь, выбирается как координаты ближайшей точки к этой гиперплоскости. Гиперплоскость, которая максимизирует расстояние между двумя параллельными гиперплоскостями, называется разделяющей гиперплоскостью (на рисунке 1.4 обозначена буквой L). На основе расстояния, разделяющего гиперплоскости вычисляется средняя ошибка классификатора, чем больше расстояние – тем ниже средняя ошибка.

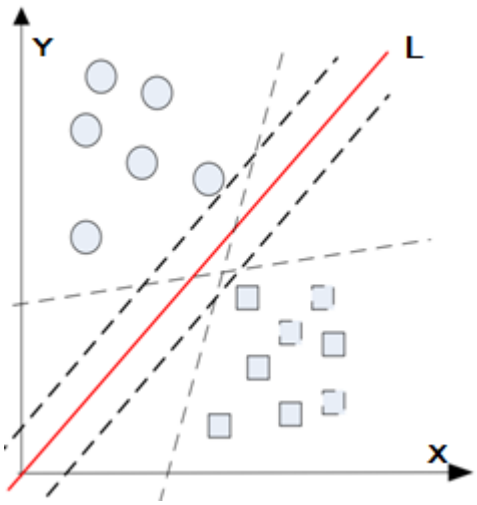


Рисунок 1.4. Разделяющая гиперплоскость в методе опорных векторов

Преимущества метода:

- можно использовать маленькую обучающую выборку;
- можно свести к задаче выпуклой оптимизации;

Недостатки метода:

- мало подходят в задачах небинарной классификации;
- неустойчив к аномалиям.

1.6. Существующие решения и аналоги

Коммерческий рынок автоматизированных систем анализа социальных сетей насыщен системами, обеспечивающими анализ данных с привязкой к конкретному бренду и помогающими принимать бизнес-решения на основании этих данных. К таким системам можно отнести YouScan, BrandAnalytics, IQBuzz, IQSocial и др.

Помимо этого существует набор инструментов, обеспечивающих сбор данных с профилей конкретного пользователя, от имени которого запускается приложение. Подобные системы предназначены в первую очередь для анализа вовлеченности, популярности бренда (через рейтинг соответствующего сообщества или профиля пользователя). Примерами подобных систем являются Cyfe, KUKU.io, Popsters, Postee др.

ForsMedia – система для извлечения из социальных сетей разнообразной информации о существующих и потенциальных клиентах.

В основе решения инструменты и технологии для работы с большими данными. ForsMedia предназначена для анализа ЦА на основе текстовой информации, которая собирается из различных источников. Система рассчитана на оценку эффективности взаимодействия с действующими клиентами и привлечением новых клиентов.

YouScan – система, предназначенная для мониторинга именно русскоязычных социальных медиа. Данная система может отслеживать упоминания брендов и продуктов в блогах, форумах, социальных сетях (Facebook, ВКонтакте). Результаты мониторинга представлены в специальном аналитическом интерфейсе.

BuzzLook – это русскоязычный сервис мониторинга социальных медиа: Facebook, «ВКонтакте», Livejournal, Flickr, YouTube и Twitter. Данная система позволяет отследить как бренд, так и деятельность конурентов.

Возможности:

- Поиск и работа над упоминаниями;
- Просмотр статистики внутри системы;

- Выгрузка графиков и упоминаний (excel, png);
- Формирование отчетов и автоматическая отправка на почту.

IQBuzz – это профессиональный инструмент анализа и управления репутацией в Интернете, сервис для мониторинга социальных медиа. Особенностью сервиса является адапитивность – по запросу клиента подключаются новые социальные сети и другие медиа для мониторинга.

BrandAnalytics – система анализа бренда в социальных медиа. Отслеживает упоминания в социальных сетях, блогах, форумах, сайтах отзывов, мессенджерах, а также онлайн СМИ. Высокая степень автоматизации – определение тональности с точностью 90%, автоматическое тегирование, фильтрация спама и нерелевантных сообщений, оповещение об угрозе для репутации, персонализированные отчеты. Адаптирован только для поиска отзывов об ограниченном числе брэндов.

Meople. Данный сервис позволяет собрать на одной странице контент более десятка наиболее популярных социальных сетей. Среди них – Facebook, Twitter, Linkedin, Одноклассники, ВКонтакте, Flickr, Google+, Instagram и еще несколько других.

Контент из каждой социальной сети представлен в виде вертикальной колонки, в которой отображается лента обновлений пользователя, его друзей и сообществ. Сервис является бесплатным.

1.7. Выводы по главе

В рамках первой главы был проведен анализ современного состояния моделей и методов интеллектуального анализа текстовых данных социальных сетей.

1. Был рассмотрен процесс подбора персонала в организацию и выделены места в алгоритме, которые можно автоматизировать. Социальные сети – это источник информации о кандидате, информацию из которого можно агрегировать в автоматизированном режиме. Была рассмотрена классификация систем анализа социальных сетей, а так же проведен обзор

существующих систем анализа социальных сетей.

Были сделаны выводы, что в данный момент времени процесс подбор персонала недостаточно автоматизирован и затрачиваемое на подбор время может быть сокращено при автоматизации процессов первоначального отбора и построения психологического портрета кандидата на основе открытых данных о нем.

2. Были рассмотрены существующие методы формирования обучающей выборки для сентимент-анализа текста, среди которых можно выделить такие, как математическое планирование, программная генерация, применение сэмплирования, вероятностные методы и детерминированные методы. Рассмотрены различные методы сентимент-анализа текста, как разделяющего текст на две эмоции – положительную и отрицательную, так и на семь эмоций. Приведены аналоги и рассмотрены проблемы определения тональности текста на русском языке и наиболее популярные тезаурусы. Сделан вывод, что не существует решений, позволяющих учитывать авторские символы выражения эмоций в контексте русского языка.

3. Были рассмотрены способы построения психологического портрета человека на основе публичной информации социальных сетей. Приведены примеры и рассмотрены различные способы классификации людей. Рассмотрены существующие решения и аналоги, в некоторых случаях [83] удалось достигнуть до 82,2% точности, однако приведенные алгоритмы применимы только к текстам на английском языке, показывая не очень высокие результаты для текстов на других языках, особенно – на имеющих другую структуру, как, например, русский, японский или казахский.

4. Были рассмотрены коммерческие системы анализа данных социальных сетей. Среди продуктов, выполняющих не только мониторинг, но и анализ данных, из положительных сторон можно выделить точность, достигающую 90%, однако большинство из них построены на основе статистических методов встречаемости брендов и адаптированы только для поиска отзывов об ограниченном числе брендов.

Глава 2. Методы и алгоритмы интеллектуального анализа текстовых данных социальных сетей

2.1. Унификация извлекаемых данных различных социальных сетей

При создании социальных сетей их разработчики обычно не опираются на какую-либо общую модель хранения данных, поэтому структура данных каждой социальной сети уникальна. Поэтому одной из задач при работе с несколькими социальными сетями является унификация структурированных и неструктурированных извлеченных данных [120].

Формально модель онтологии профилей социальных сетей выглядит 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 онтологии;

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

где $N^{B} = \{n_1^{B}, n_2^{B}, \dots, n_m^{B}\}$ –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:

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

где $R^{OP} = \{r_1^{OP}, r_2^{OP}, \dots, r_s^{OP}\}$ – 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. Примеры отношений

предлагаемой модели представлены в таблице 2.1.

Таблица 2.1. Примеры отношений онтологической модели базы знаний

	Поле профиля	Область	Отношение	Тип
Datatype Properties				
1	Имя	Пользователь	#имеетИмя	строка
2	Фамилия	Пользователь	#имеетФамилию	строка
3	Дата	Пользователь	#имеетДатуРождения	дата
Object Properties				
4	Школа	Пользователь	#обучалсяВ	объектШкола
5	Город	Пользователь	#живетВ	объектГород
6	Аудиозапись	Звук	#имеетЗвук	двоичноеЗвук
7	Автор аудиозаписи	Пользователь	#имеетАвтора	пользователь
8	Пост	Пользователь / Сообщество	#имеетПост	объектПост
9	Имеет друга	Пользователь	#имеетДруга	пользователь

R^{CONT} – Annotation Properties, это свойства аннотации, необходимые для определения контекста.

Было определено, что в рамках данной модели можно выделить следующие типы контекстов:

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

- R^{T} – это отношение аннотации описывающее некоторый промежуток времени, во время которого данное отношение существовало

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

Схематично соотношения контекстов времени и источников данных для одного пользователя представлено на рисунке 2.1.

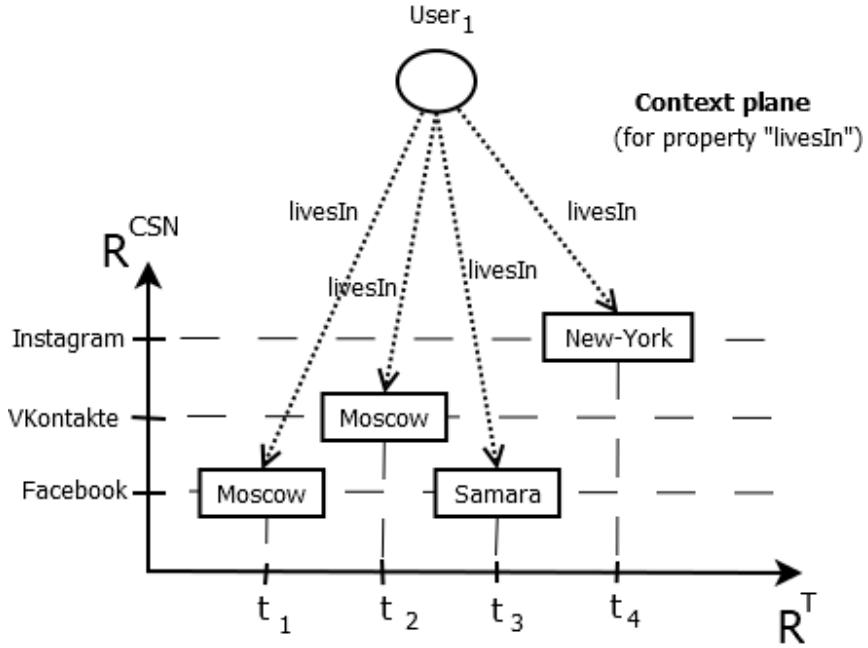


Рисунок. 2.1. Соотношение контекстов времени и источника

2.2. Алгоритм формирования обучающей выборки

Профиль пользователя в социальной сети можно представить следующим образом:

$P = \{Dq, Stat, G\}$, где $\forall i\ p_i \in P$: $p_i = \{< Dq_i, Stat_i, SocG_i >\}$,

где Dq – информация о пользователе, извлеченная из анкеты профиля;

S – это статистические данные, а также динамика активности пользователя;

$SocG$ – это социальный граф пользователя, включающий профили, связанные с анализируемым профилем отношениями: друзья, подписчик, родственник, коллега и пр.

Тексты социальных сетей имеют ряд особенностей, затрудняющих процесс семантического анализа:

- Наличие грамматических ошибок;
- Наличие эмотиконов.

Поэтому этап предобработки постов пользователей социальной сети включает следующие этапы [11]:

- Графематический анализ текста (разбиение текста на простые предложения). В большинстве случаев для обучающей и тестовой выборки использовались тексты не больше 3 предложений.
- Исправление орфографических ошибок. Для решения данной задачи использовалась программная библиотека DeepPavlov [163].
- Удаление стоп-слов.
- Лемматизация слов и словосочетаний.

В качестве метода индексации использовался метод n-грамм [127]. выбор признаков – TF-IDF [69].

В качестве классификаторов использовалось два метода: метод опорных векторов (SVM) и метод случайного леса (RF).

Метод опорных векторов – это метод линейной классификации. Алгоритм классификации выглядит следующим образом:

$$a(x) = sign\left(\sum_{i=1}^{n} \lambda_i c_i \mathbf{x_i} \cdot \mathbf{x} - b\right),$$

где $C_i = \{1;-1\}$, в зависимости от того, какому классу принадлежит точка x_i. Каждое x_i – это p-мерный вещественный вектор, нормализованный значениями [0,1], b-гиперпараметр,$\lambda=(\lambda_i,\dots,\lambda_n)$– вектор двойственных переменных.

При формировании обучающей выборки для сентимент-анализа производится разметка текстов в соответствие с их эмоциональной окраской и возникает необходимость предобработки текстовых данных [101].

Процесс формирования обучающей выборки, состоящий из двух этапов отбора, представлен на рисунке 2.2. Каждый этап отбора выполняется для каждой конкретной эмоции.

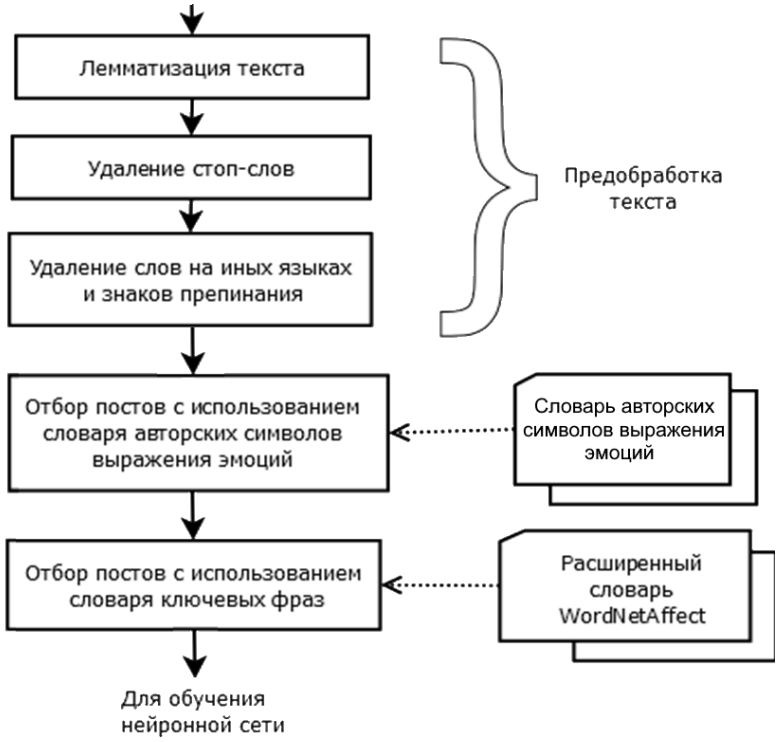


Рисунок 2.2. – Процесс отбора постов

Первым этапом выполняется предобработка постов, состоящая из трех действий – приведение букв к нижнему регистру, удаление всех символов, не являющихся буквами или пробелами и удаление слов на языках, отличных от русского.

Вторым этапом алгоритма происходит лемматизация всех предтсавленных в посте слов, после чего лемматизированные слова проверяются по словарю. Если в посте будет присутствовать слово из конкретной группы, то пост будет принадлежать определенному классу эмоциональной окраски.

На третьем этапе выполняется отбор постов с использованием словаря авторских символов выражения эмоций («смайлы» или «эмодзи»). Каждый пост проверяется на содержание авторского символа, и если символ встретился, то пост добавляется в определенный список [147]. Примеры таких символов представлены на рисунке 2.3.

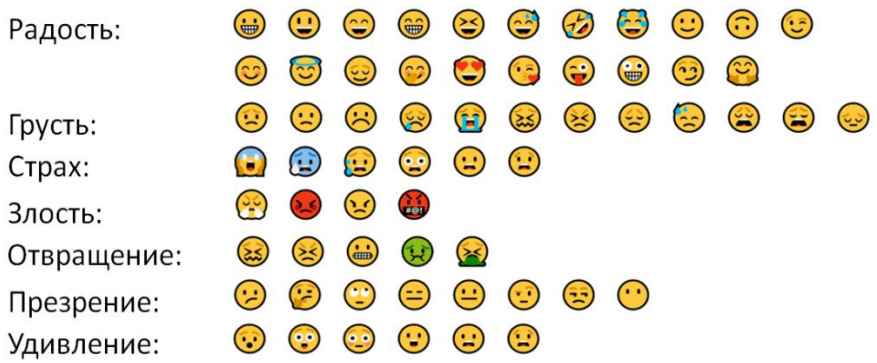


Рисунок 2.3. – Словарь авторских символов выражения эмоций

Формально разработанный словарь описывается следующим образом:

$$D^E = \{D^E_{joy}, D^E_{sad}, D^E_{surp}, D^E_{anger}, D^E_{disg}, D^E_{cont}, D^E_{fear}\},$$

где D^E_{joy} – множество элементов класса «радость»,

D^E_{sad} – множество элементов класса «грусть»,

D^E_{surp} – множество элементов класса «удивление»,

D^E_{anger} – множество элементов класса «злость»,

D^E_{disg} – множество элементов класса «отвращение»,

D^E_{cont} – множество элементов класса «презрение»,

D^E_{fear} – множество элементов класса «страх».

Пример работы отбора представлен в таблице 2.2.

Таблица 2.2. – Процесс отбора постов при помощи словаря авторских символов

Текст	**Эмоция**	**Фильтр пройден**
я очень люблю осень 😀	радость	Да
продукты на полках снова появились	-	Нет
опять я забыла помыть машину 😒	грусть	Да
дорожники ремонтируют асфальт во время дождя ради повышенного коэффицента	-	Нет
картины и фотографии это все-таки разные вещи	-	Нет
гонки были просто великолепны 😍	радость	Да
я тебя жареные гвозди заставлю есть 😫	злость	Да

Последним этапом выполняется отбор постов с использованием словаря ключевых фраз. Если пост содержит ключевое слово, то он добавляется в определенный список, иначе исключается из рассмотрения. За

основу словаря был взят тезаурус WordNetAffect [13]. Пример расширенного словаря WordNetAffect представлен в таблице 2.3.

Таблица 2.3. – Расширенный словарь WordNetAffect

Текст	**Словарь**	**Эмоция**
с неприязнью	WNA-RU	отвращение
вызывать тошноту	WNA-RU	отвращение
яростно	WNA-RU	злость
негодующе	WNA-RU	злость
мерзость	WNA-RU-EXT	отвращение
бесящий	WNA-RU-EXT	злость

Полученный алгоритм можно описать следующими шагами:

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.

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:

4.1. Удаление слов/словосочетаний, содержащих символы латинского алфавита.

4.2 Исправление орфографических ошибок с использованием программной библиотеки DeepPavlov.

4.3 Лемматизация текста, удаление стоп-слов.

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.

2.3. Подход к сопоставлению профилей пользователей с использованием гибридизации различных подходов структурированных и неструктурированных данных

2.3.1. Критерии схожести профилей

Так как нет общепринятой методики идентификации пользователей в различных ресурсах [170], то для решения данной задачи в социальных сетях «ВКонтакте», «Одноклассники» и «Facebook» было решено разработать собственную методику.

Идентификация и поиск страниц производятся на основании данных профиля в одной из названных сетей. Так как в различных сетях пользовательские данные представлены в различных форматах, например, по-разному указывается отчество, форматы представления времени, мест работы и учебы различны, то для сравнения данных необходимо привести все поля к унифицированному виду. Далее представлены поля, которые использовались для сравнения аккаунтов:

- Фамилия, имя, отчество пользователя;
- Дата рождения;
- Место проживания;
- Место рождения;
- Друзья;
- Текстовые заметки (посты);

- Место работы;
- Место учебы;
- Контакты, email, номер телефона;
- Аватар профиля, а также фотографии из профиля.

Данная информация загружается, как для исходного профиля, так и для искомых профилей в других социальных сетях. Загруженные данные профилей сопоставляются с данными исходного профиля. Для унификации имен использовалась транслитерация и приведение к нижнему регистру.

Анкетные данные профиля пользователя в социальной сети отличаются структурированностью, требуют меньше затрат на предобработку.

2.3.2. Критерий схожести лиц

Критерий наличия схожих лиц на фотографиях. Так как одинаковые фотографии в различных сетях могут иметь различные разрешения, то сравнение по пикселям невозможно. Для решения данной проблемы можно воспользоваться методом гистограмм направленных градиентов [124]. Данный метод позволяет получить векторное представление объектов на изображении. В данном случае объектами являются лица пользователей. Идея метода заключается в вычислении гистограммы градиентов внутренних пикселей. Пример такой гистограммы представлен на рисунке 2.4.

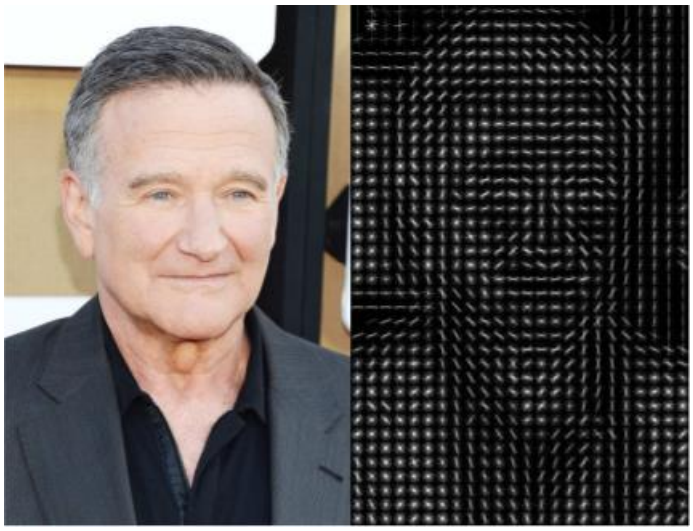


Рисунок 2.4. Пример гистограммы направленных градиентов

На данной диаграмме представлены направления градиентов, которые позволяют выделить некоторые формы в исходном изображении. При вычислении градиентов производится свертка изображения с ядрами [-1, 0, 1] и $[-1, 0, 1]^T$, что дает нам две матрицы D_x и D_y, производные вдоль осей x и y соответственно. На основании этих матриц вычисляются углы и модули градиентов в каждой точке изображения. Полученные с помощью данного метода векторы пропускаются через некоторый классификатор (например, SVM) с целью определения паттернов лица на изображении. Полученные вектора с паттернами лиц попарно сравниваются для изображений из разных социальных сетей. Сравнение происходит путем вычисления евклидова расстояния между векторами. Если данное расстояние меньше или равно 0.85, то векторы, как и изображения, считаются совпадающими.

2.3.3. Критерий схожести контактов, мест работы и учебы

Критерий наличия схожих контактов. Профили, содержащие ссылки друг на друга с большой долей вероятности относятся к одному и тому же человеку.

Критерий наличия схожего места работы и места учебы. Поскольку в разных социальных сетях может быть разный способ отображения места работы и учебы человека, то строки должны пройти предобработку, аналогичную той, что проходят посты социальных сетей. После этого производится сравнение строк по следующей метрике:

$$\frac{n}{max(l1,\ l2)},$$

где n – количество попарно совпадающих лемм; l1, l2 – количество лемм в 1 и 2 строке соответственно.

Если значение этого критерия более 0,85, то строки считаются похожими.

Критерий наличия одинаковых друзей. Данный показатель рассчитывается путем попарного сравнения имен друзей. Имена проходили такую же предобработку, как и имя пользователя на анализируемой

странице, описанную выше. Чем больше совпадений, тем выше данный профиль в списке кандидатов.

2.3.4. Критерий схожести сообщений

Критерий наличия схожих постов. Для сравнения текстовых заметок использовалось две метрики. Первая – это нахождения расстояния Левенштейна, которое рассчитывается согласно модели:

$$D(i,j) = \begin{cases} 0, i = 0, j = 0 \\ i, j = 0, i > 0 \\ j, i = 0, j > 0 \\ min(\begin{matrix} D(i,j-1)+1, D(i-1,j)+1, \\ D(i-1,j-1)+m(S[i],S[j]) \end{matrix}), j > 0, i > 0 \end{cases}$$

Модель описывает минимальное количество операций вставки, удаления и замены одного символа, который необходимых для трансформации двух строк друг в друга.

Вторая метрика – это алгоритм шинглов. В данном методе большой текст разбирвается на маленькие кусочки, называеые шинглами, после чего происходит вычисление хэшей данных шинглов и попарное сравнение хешей. Для метода шинглов использовалась следующая метрика:

$$\frac{2*n}{s1+s2},$$

где n – количество совпадающих хешей шинглов; s1, s2 – количество шинглов в первой и второй строке соответственно.

Визуальное представление алгоритма шинглов представлено на рисунке 2.5.

Рисунок 2.5. Алгоритм шинглов

Построение семантического представления социального портрета

пользователя социальной сети предполагает определение множества интересов пользователя посредством семантического анализа текстовых данных со страниц пользователя.

Данная цель достигается путем решения задачи классификации текстовых данных со страницы профиля пользователя социальной сети.

Классами, в рамках данной задачи, будут категории интересов пользователя, которые напрямую связаны с соответствующими предметными областями (музыка, политика, общественные движения, искусство и пр.). Отсюда набор анализируемых текстовых фрагментов:

$$D^{sn} = \{d_1, d_2, \dots, d_n\}.$$

Формально тематики интересов пользователя можно представить следующим образом:

$$C^{sn} = \{c_r\}, \text{ где } r = 1, \dots, m.$$

При этом тематики/категории могут иметь иерархическую структуру:

$$H^{sn} = \{< c_j, c_p >, c_j, c_p \in C^{sn}\}$$

(т.е. тематика c_p является дочерней по отношению к тематике c_j).

Иерархия тематик в рамках данной модели реализована в виде предметной онтологии [116], в качестве объектов которой представлены соответствующие тематики.

Отсюда базовой задачей является установление соответствия между анализируемым текстовым фрагментом d_i и конкретной тематикой C^{sn}. Предполагается, что текстовый фрагменты, относящиеся к одной тематике, содержат схожие лингвистические единицы (слова или словосочетания).

Формируется словарь, который состоит из наборов терминов, характерных для каждой тематики. Полученный словарь представляет собой лингвистическую основу базы знаний анализа текстовых ресурсов социальных сетей:

$$F^{sn}(C^{sn}) = \cup (c_r),$$

где $F^{sn}(c_r) = < f_1, \dots, f_l, \dots, f_z >$.

Из каждого текстового фрагмента анализируемого текста извлекается

набор признаков (терминов), которые в значительной степени выделяют его среди других текстовых фрагментов:

$$F(d_i) =< f_1^i, \dots, f_l^i, \dots, f_y^i >.$$

Набор признаков интересов пользователей должен совпадать с набором семантических признаков всех анализируемых текстовых фрагментов, отсюда следует задача обеспечения полноты анализа:

$$F^{sn}(C^{sn}) = F(D) = \cup F(d_i).$$

Установление текстовому фрагменту d_i в соответствие определенная тематика c_r осуществляется следующим образом:

$$F(d_i) \cap F^{sn}(c_r).$$

Численный показатель, определяющий степень соответствия анализируемого текстового фрагмента той или иной тематике, вычисляется следующим образом:

$$Val_{ir} = \frac{count(F(d_i) \cap F^{sn}(c_r))}{count(F^{sn}(c_r))}, Val_{ir} = [0..1],$$

где $count(F(d_i) \cap F^{sn}(c_r))$ – это число совпавших терминов словарей $F(d_i)$ и $F^{sn}(c_r)$ соответственно; $count\left(F^{\text{sn}}(c_r)\right)$ – это общее число терминов вв сформированном словаре $F^{sn}(c_r)$.

После этого формируется массив степеней соответствия текстового фрагмента конкретным тематикам:

$$\delta(d_i) =< Val_{i1}, Val_{i2}, \dots, Val_{im} >$$

Итоговое значение показателя степени принадлежности текстовых высказываний пользователя социальной сети к конкретным тематикам, определяющим интересы пользователя, получается по следующей формуле:

$$\mu_r = \frac{\sum_i^n \begin{cases} 1, max(<Val_{i1}, Val_{i2}, \dots, Val_{im}>) = Val_{ir} \\ 0, max(<Val_{i1}, Val_{i2}, \dots, Val_{im}>) \neq Val_{ir} \end{cases}}{n},$$

где n – это число текстовых фрагментов; m – число категорий.

2.3.5. Критерий совпадения социальных графов

Формально набор статистических данных и динамики активности пользователя представляется следующим образом:

$$Stat_i = < \{time_j^{sn}, sel_k^{sn}\}, \{< char_s^{sn}, vl_s^{sn} >\}_{s=\overline{1,|char^{sn}|}} >_{j=\overline{1,n};k=\overline{1,m}}.$$

где $time_j^{sn}$ – это временной промежуток анализа;

sel_k^{sn} – это соответствующий период времени;

n – это общее число временных промежутков;

m – это число периодов времени анализа активности пользователя социальной сети.

Набор показателей ($char_s^{sn}$), по которым собирается статистика и определяется динамика изменения их числа, включает следующие компоненты:

- Число друзей пользователя;
- Число подписок пользователя;
- Число фотографий пользователя (особенно данный показатель актуален для социальной сети Instagram);
- Число сообществ или групп пользователя;
- Число лайков (знаков одобрения/согласия), поставленных пользователем отдельным постам/сообщениям/фотографиям;
- Число репостов текстовых сообщений;
- Число подписок и пр.

vl_s^{sn} – это соответствующие значения перечисленных показателей.

Социальный граф пользователя – это граф, включающий профили в качестве узлов графа, связанные с анализируемым профилем различными видами отношений (дугами):

$$SocG = (N^{sn}, A^{sn}, T^{sn}),$$

где N^{sn} – это множество узлов;

A^{sn} – это множество дуг, состоящих из пар идентификаторов пользователей, связанных между собой определенным отношением;

T^{sn} – это множество типов отношений, связывающих между собой пользователе:

- Друг (в частных случаях, «лучший друг»);
- Подписчик;

- Коллега по работе;
- Родственник и пр.

Таким образом, подход к сопоставлению профилей пользователей в разных социальных сетях можно представить алгоритмом:

Шаг 1. Применение критерия схожести анкет профилей.

Шаг 2. Применение критерия наличия схожих лиц на фотографиях.

Шаг 3. Применение критерия наличия схожих контактов.

Шаг 4. Применение критерия наличия схожего места работы и учебы.

Шаг 5. Применение критерия наличия схожих сообщений со страниц профиля с посредством нахождения расстояния Левенштейна и использования алгоритма шинглов.

Шаг 6. Применение критерия совпадения социальных графов.

2.4. Определение психологических характеристик пользователя социальных сетей

2.4.1. Классификация психологических характеристик пользователя с использованием метода «Большой пятерки»

Схематично общий процесс психолингвистического анализа данных социальных сетей с использованием машинного обучения и модели Большой Пятерки представлен на рисунке 2.6.

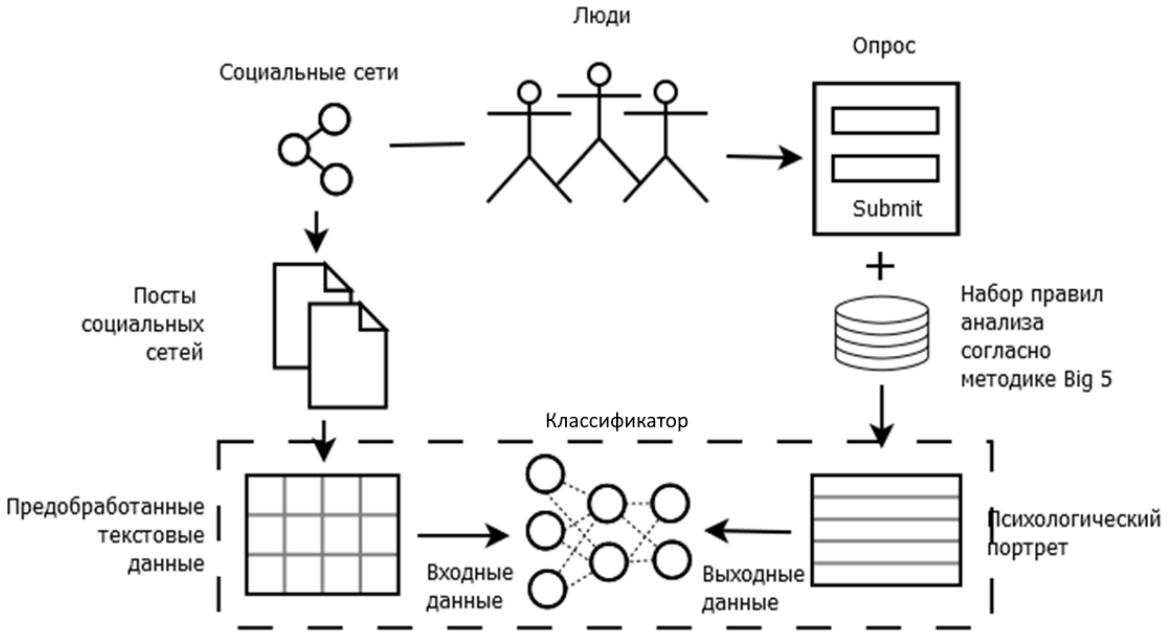


Рисунок 2.6 – Подход к психолингвистическому анализу данных социальных сетей

Основой предложенного подхода является решение задачи классификации текстов пользователей социальных сетей с целью определения психолингвистических характеристик автора.

Классификация текстов состоит из нескольких последовательных этапов (рисунок 2.7).

Рисунок 2.7 – Этапы классификации текстов

В качестве исходных данных в рамках предложенного подхода используется текст из «постов» страниц пользователей в социальных сетях. Использовались только тексты, написанные автором, копии текстов других страниц («репосты») не учитывались. В основном тексты содержали личные мнения, рассуждения и мысли авторов. Основные источники текстов на страницах [113]:

- сообщения (посты);
- текстовые статусы;
- комментарии к собственным и иным сообщениям.

Размеры текстов варьировались от одного-двух предложений до

нескольких десятков предложений.

Одной из моделей личности, применяемой в психологии, является «Большая пятерка» (Big five) – это пятифакторная модель личности. Одной из главных особенностей данной модели является возможность преобразования входящих в нее черт в достаточно полный психологический портрет. Она состоит из пяти основных факторов объединяющих группу черт личности. Для упрощения составления данной модели были разработан пятифакторный опросник личности, представленный в виде теста. Классический вариант модели «Большая пятерка» выделяет следующие основные факторы:

- Нейротизм (neuroticism);
- Экстраверсия (extraversion);
- Открытость опыту (openness to experience);
- Согласие, или сотрудничество (agreebleness),
- Сознательность, или добросовестность (conscientiousness).

Основным тест для автоматизированного тестирования является «тест 5PFQ», разработанный Heijiro Tsuji (Хеджиро Тсуджи). Данный тест был также переведен на русский язык Хромовым А.Б. Опросник содержит парные высказывания – оценка от 0 до 4.

Пример вопроса: «Все новое вызывает у меня интерес» – «Часто новое вызывает у меня раздражение»

2.4.2. Психолингвистический анализ текстовых данных социальных сетей

В рамках данной работы было решено использовать пятифакторный метод оценки личности на основе анализа текстовых данных, извлеченных из профилей пользователя, также называемый «Большая пятерка» [63]. Пятифакторная модель личности является диспозициональной моделью личности и включает в себя пять различных шкал (рисунок 2.8).

Рисунок 2.8. Пятифакторный метод анализа личности

Задача психолингвистического анализа – это задача бинарной классификации по пяти факторам. Задача классификации предполагает разбиение множества объектов на классы посредством определения соответствия характеристик объектов конкретным правилам.

В качестве входных значений на классификатор подаются предобработанные текстовые массивы, извлеченные из постов пользователя в социальной сети.

Выходными значениями для нейронной сети являются:

$$Out = \{N, E, O, A, C\}$$

где N – Нейротизм, |N|=2;

E – Экстраверсия, |E|=2;

O – Открытость опыту, |O|=2;

A – Согласие, или сотрудничество, |A|=2;

C – Сознательность, или добросовестность, |C|=2.

Алгоритм психолингвистического анализа генерирует описание психофизиологического портрета человека в соответствии с вышеописанной

пятифакторной моделью путем анализа текстовых данных со страниц его профилей в разных социальных сетях (рис.2.6).

Полученный метод определения психологических характеристик пользователя социальных сетей можно описать алгоритмом:

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:

1.4.1. Удаление слов/словосочетаний, содержащих символы латинского алфавита.

1.4.2 Исправление орфографических ошибок с использованием программной библиотеки DeepPavlov.

1.4.3 Лемматизация текста, удаление стоп-слов.

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».

2.5. Алгоритм анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей

Процесс сентимент-анализа, включающий алгоритм формирования обучающей выборки, алгоритмы предобработки текста для формирования векторов и нейросетевой подход для определения тональности текстов, представлен на рисунке 2.9.

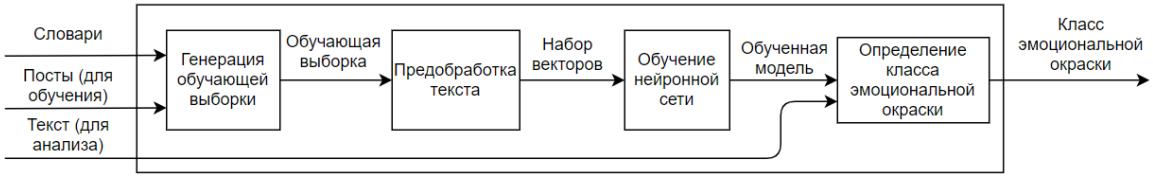


Рисунок 2.9 – Процесс сентимент-анализа

Разработанный алгоритм сентимент-анализа текстовых сообщений (постов) социальных сетей включает следующие этапы:

На вход алгоритму подаются посты, полученные из социальной сети «ВКонтакте» из открытых групп, в количестве 2,5 млн. штук.

Происходит генерация обучающей выборки с использованием разработанных словарей. Формирование выборки происходит в несколько этапов. На выходе получается размеченная обучающая выборка.

Происходит предобработка обучающей выборки, то есть преобразование текстовых данных в набор векторов с помощью моделей «BERT» и «word2vec».

Происходит обучение модели нейронной сети с помощью набора векторов.

После обучения возвращается точность на обучающей и тестовой выборке, и появляется возможность подавать текст на вход нейронной сети и получать класс эмоциональной окраски.

Для того чтобы обучить нейронную сеть классифицировать текстовые данные, необходимо представить текст в векторном виде [51]. Для этого

используются модели «word2vec» и «BERT».

Модель алгоритма «BERT» является функцией, которая преобразует текст в вектор. В данном алгоритме каждому слогу ставится в соответствие число. При инициализации происходит скачивание обученной модели для поределенного языка. Модель разбивает входную последовательность на токены – слоги, а затем каждый слог преобразует в число, тем самым формируя вектор. Подробнее о «BERT» написано в работах [33] и [110].

Математическую модель «BERT» можно представить в виде:

$$\Theta = \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ w_n \end{bmatrix},$$

где Θ – это вектор, состоящий из множества слов w из словаря загруженной модели. Алгоритм преобразует входную последовательность в вектор.

Пусть w_1, w_2 … w_n – набор слов входной последовательности и s_{m1}, s_{m2} … s_{mn} – множество слогов выходной последовательности w_n, тогда функция $(s_m) = f(w_{11},w_{12}…w_{1n})$ преобразует входную последовательность слов в выходную последовательность слогов. Таким образом, появляется возможность преобразовывать входную последовательность в векторное представление.

Для сравнения эффективность модели «BERT» так же была опробована модель «word2vec».

Модель алгоритма «word2vec» является функцией, которая преобразует текст в вектор. В данном алгоритме каждому слову ставится в соответствие число. Подробнее о «word2vec» написано в работе [89].

Математическую модель «word2vec» можно представить в виде набора векторов всех слов, входящих в корпус:

$$\Theta = \begin{bmatrix} V_{w1} \\ V_{w2} \\ . \\ V_{wn} \\ U_{w1} \\ U_{w2} \\ . \\ U_{wn} \end{bmatrix},$$

где Θ это вектор, состоящий из векторов v и u длины d всех слов.

Алгоритм присваивает каждому слову число равное его вероятности встретиться в реальном тексте.

$$P(w_o|w_c) = \frac{e^{s(w_o,w_c)}}{\sum w_i \in V e^{s(w_o,w_c)}}$$

где w_o – вектор целевого слова, w_c – вектор контекста, вычисленный из векторов других слов, вокруг текущего слова. $s(w_1,w_2)$ – функция, сопостовляющая число двум векторам.

Для оптимизации вероятностей встречаемости слов используется дивергенция (изменение разрыва) Кульбака-Лейблера:

$$KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

где p(x) – распределение вероятностей слов, q(x) – распределение вероятностей, генерируемое моделью.

Модель нейронной сети может быть представлена множеством слоев, которые используются в ее архитектуре. Математический нейрон можно представить в виде функции, выход которой рассчитывается с помощью его входов и матрицы весов:

$$y = f(u), \text{ где } u = \sum_{i=1}^{n} w_i x_i + w_0 x_0$$

где x_i – это входы нейрона, а w_i – значения матрицы весов. Функция u называется индуцированным локальным полем, а f(u) – передаточной функцией.

Сигналы на входе могут варьироваться от 0 до 1. Вход x_0 и вес w_0 служат для инициализации нейрона.

С каждым нейроном связано понятие функции активации, которую можно обобщенно представить, как:

$$f(x) = tx$$

где t – некоторый множитель, отвечающий за распределение функции активации.

При проведении исследования определения тональности текстов были использованы различные архитектуры нейронных сетей. Слои, использованные в данной работе, относятся к типам LSTM, Bidirectional LSTM, CNN, MLP, GRU.

Слой Embedding – входной слой нейронной сети, представляет собой словарь:

$$Emb = \{Size(D), Size(S_{vec}), L_{Sec}\},$$

где Size(D) - размер словаря, $Size(S_{vec})$ – размер векторного пространства, $Size(S_{vec})$, L_{Sec} – длина выходного вектора, равная максимальной длине вектора, полученного при обработке всех слов.

- Слой Conv1D – сверточный слой нейронной сети, состоящий из последовательности фильтров.
- Слой MaxPooling1D – слой, отвечающий за уменьшение размерности карт признаков.
- Слой LSTM – рекуррентный слой нейронной сети, состоящий и последовательности блоков математических операций.
- Слой Dropout – данный слой используется, чтобы не допустить переобучения. Данный слой позволяет исключить некоторый процент нейронов из модели.
- Слой Dense – выходной полносвязный слой.
- Слой Flatten – слой, использующийся для преобразования многомерных векторов в одномерный вектор.
- Слой Bidirectional – двунаправленный слой, создающий 2 параллельно-работающих экземпляра слоя, переданного в параметре.

Рекуррентная нейронная сеть, в отличие от персептрона,

состоящего из слоев нейронов, состоит из блоков выполнения математических операций над сигналами. Особенность этих блоков в том, что они сохраняют информацию для дальнейшего использования.

Принцип работы LSTM подробнее рассмотрен в работе [65]. Архитектура рекуррентной нейронной сети представлена на рисунке 2.10.

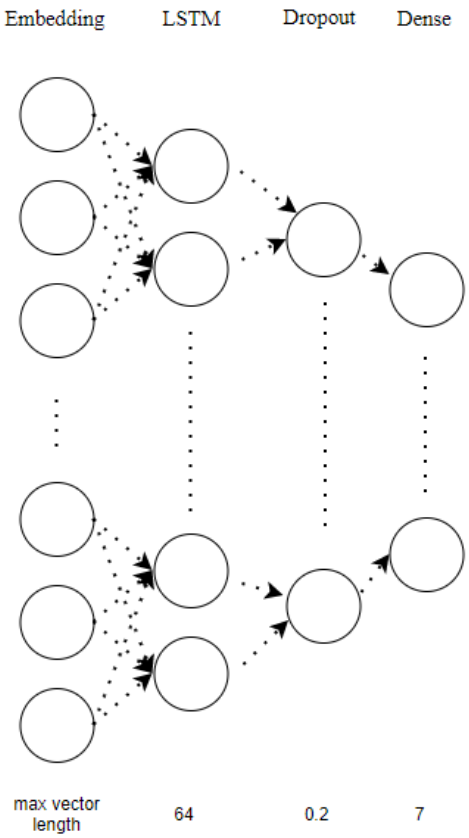


Рисунок 2.10 – Архитектура рекуррентной нейронной сети

Последовательность работы рекуррентной нейронной сети предполагает последовательное прохождение данных через все слои архитектуры нейронной сети и состоит из следующих этапов:

- На слой Embedding подается набор векторов, полученные с помощью моделей «word2vec» или «BERT» на этапе предобработки. Слой Embedding фактически является словарем.
- Передача данных в рекуррентный слой LSTM размером в 64 блока.
- Передача данных в слой Dropout. Данный слой исключает некоторый процент нейронов для того, чтобы исключить возможное переобучение нейронной сети.
- Передача данных в слой Dense. Полносвязный слой преобразует данные в

выходной вектор, состоящий из 7 нейронов.

GRU – рекуррентный слой, основанный на том же принципе, что и LSTM, но представляет собой более простую структуру. Соответственно, GRU менее затратный при вычислениях. Слой GRU отдает предпочтение недавним событиям, что, соответственно, увеличивает весь свежей информации над исторической. Подробнее в работе [74].

Архитектура рекуррентной нейронной сети показана на рисунке 2.11.

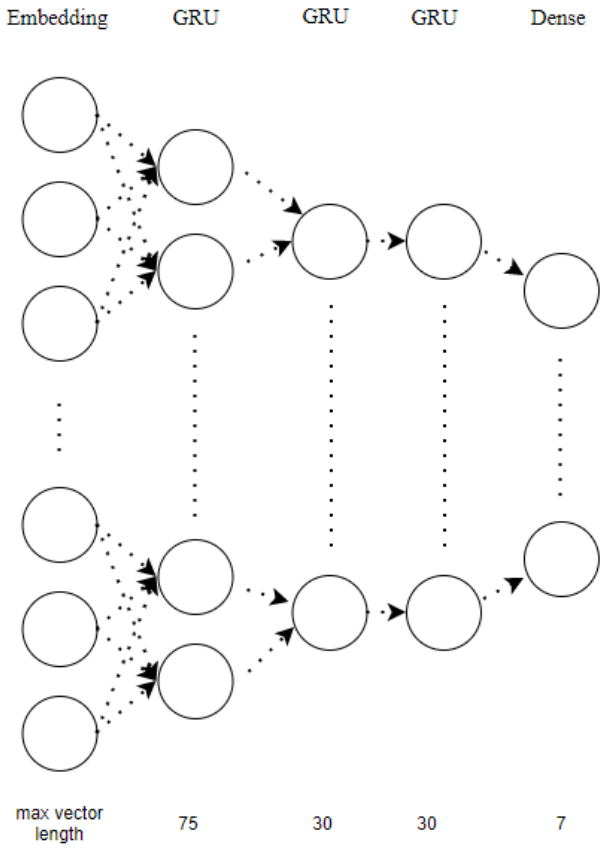


Рисунок 2.11 – Архитектура рекуррентной нейронной сети

Последовательность работы рекуррентной нейронной сети предполагает последовательное прохождение данных через все слои архитектуры нейронной сети и состоит из следующих этапов:

- На слой Embedding подается набор векторов, полученные с помощью моделей «word2vec» или «BERT» на этапе предобработки. Слой Embedding фактически является словарем.
- Передача данных в рекуррентный слой GRU размером в 75 блоков.
- Передача данных в рекуррентный слой GRU размером в 30 блоков.
- Передача данных в рекуррентный слой GRU размером в 30 блоков.

- Передача данных в слой Dense. Полносвязный слой преобразует данные в выходной вектор, состоящий из 7 нейронов.

Двунаправленная рекуррентная нейронная сеть, в отличие от обычной LSTM сети, использует 2 LSTM слоя, которые объединяются слоем Bidirectional, и просматривает входную последовательность в обоих направлениях и получает более насыщенные представления.

Архитектура двунаправленной рекуррентной нейронной сети представлена на рисунке 2.12.

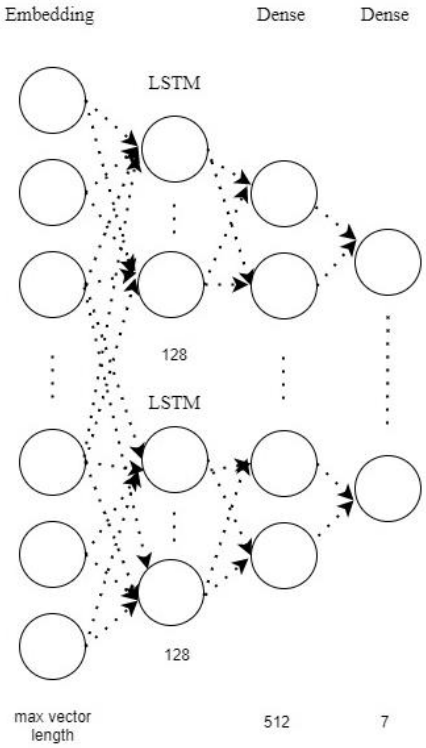


Рисунок 2.12 – Архитектура двунаправленной рекуррентной нейронной сети

Последовательность работы двунаправленной рекуррентной нейронной сети предполагает последовательное прохождение данных через все слои архитектуры нейронной сети и состоит из следующих этапов:

- На слой Embedding подается набор векторов, полученные с помощью моделей «word2vec» или «BERT» на этапе предобработки. Слой Embedding фактически является словарем.
- Передача данных в 2 параллельных рекуррентных слоя LSTM размером в 128 блоков.

- Передача данных в слой Dense, состоящий из 512 нейронов. Данный слой является полносвязным.
- Передача данных в слой Dense. Полносвязный слой преобразует данные в выходной вектор, состоящий из 7 нейронов.

Архитектура двунаправленной нейронной сети с использованием слоев GRU схожа с архитектурой, представленной на рисунке 2.12.

Сверточная нейронная сеть использует операцию свертки для выявления признаков. Операция свертки заключается в перемещении небольшоко окна по всей матрице и нахождения среднего или максимального значения. Архитектура сверточной нейронной сети представлена на рисунке 2.13.

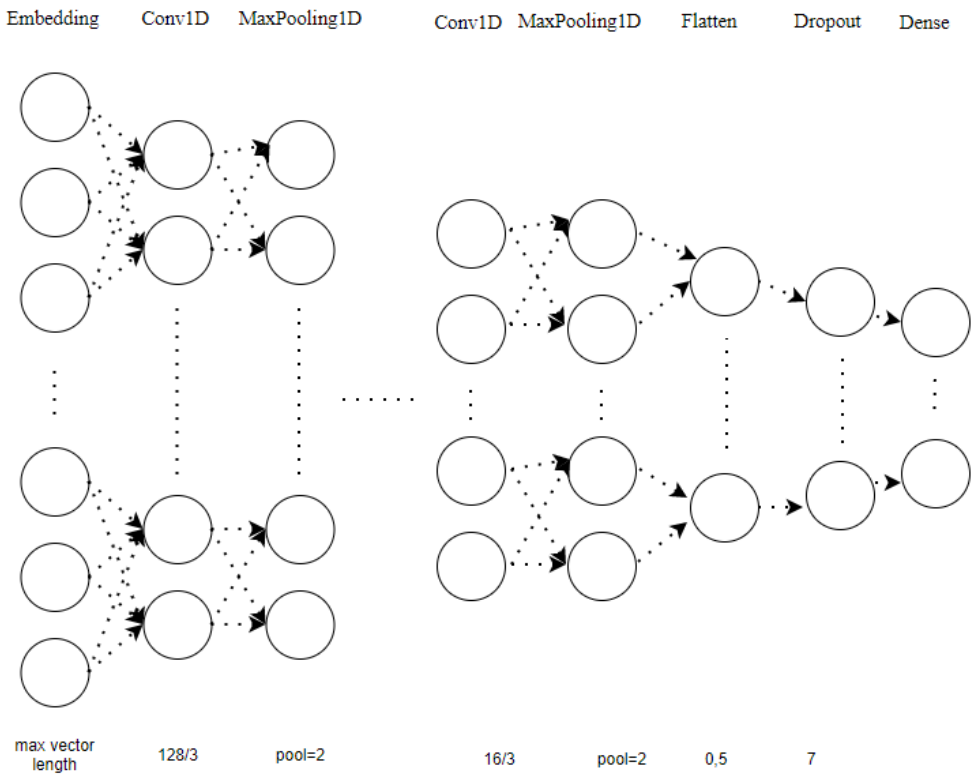


Рисунок 2.13 – Архитектура сверточной нейронной сети

Последовательность работы сверточной нейронной сети предполагает последовательное прохождение данных через все слои архитектуры нейронной сети и состоит из следующих этапов:

- На слой Embedding подается набор векторов, полученные с помощью моделей «word2vec» или «BERT» на этапе предобработки. Слой

Embedding фактически является словарем.

- Передача данных в сверточный слой размером 128 блоков.
- Передача данных в субдискретизирующий слой размерностью равной 2.
- Передача данных в сверточный слой размером 64 блока.
- Передача данных в субдискретизирующий слой размерностью равной 2.
- Передача данных в сверточный слой размером 32 блока.
- Передача данных в субдискретизирующий слой размерностью равной 2.
- Передача данных в сверточный слой размером 16 блока.
- Передача данных в субдискретизирующий слой размерностью равной 2.
- Передача данных в слой Flatten.
- Передача данных в слой Dropout, необходимый для исключения переобучения нейронной сети путем исключения лишний нейронов.
- Передача данных в слой Dense. Полносвязный слой преобразует данные в выходной вектор, состоящий из 7 нейронов.

Многослойный персептрон использует полносвязные слои Dense. Архитектура многослойного персептрона представлена на рисунке 2.14.

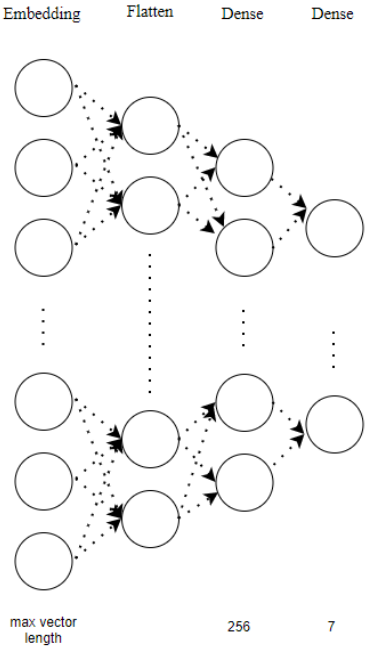


Рисунок 2.14 – Архитектура многослойного персептрона

Последовательность работы многослойного персепторона предполагает последовательное прохождение данных через все слои

архитектуры нейронной сети и состоит из следующих этапов:

- На слой Embedding подается набор векторов, полученные с помощью моделей «word2vec» или «BERT» на этапе предобработки. Слой Embedding фактически является словарем.
- Передача данных в слой Flatten.
- Передача данных в слой Dense размером 256 блоков. Данный слой является полносвязным.
- Передача данных в слой Dense. Полносвязный слой преобразует данные в выходной вектор, состоящий из 7 нейронов.

Архитектура нейронной сети со сверточным и рекуррунтным слоями применяется для распознания текста. В данной модели перед LSTM слоем добавляют сверточный слой и субдискретизирующий слой. Сверточный слой используется для выделения признаков, а LSTM слой уже работает с выделенными признаками. Архитектура нейронной сети со сверточным и рекуррентным слоями показана на рисунке 2.15.

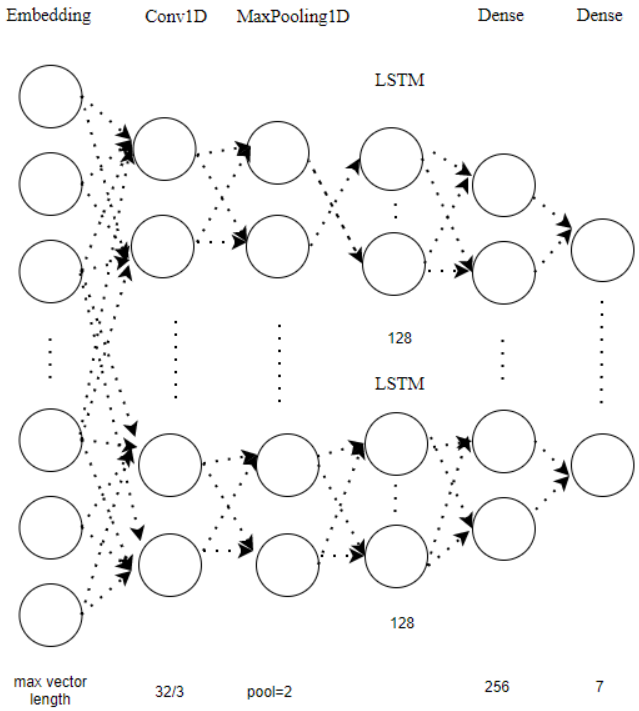


Рисунок 2.15 – Архитектура нейронной сети со сверточным и рекуррентным слоями

Последовательность работы нейронной сети со сверточным и рекуррентным слоями предполагает последовательное прохождение данных

через все слои архитектуры нейронной сети и состоит из следующих этапов:

- На слой Embedding подается набор векторов, полученные с помощью моделей «word2vec» или «BERT» на этапе предобработки. Слой Embedding фактически является словарем.
- Передача данных в сверточный слой размером 32 блока.
- Передача данных в субдискретизирующий слой размерностью равной 2.
- Передача данных в 2 параллельных рекуррентных слоя LSTM размером в 128 блоков.
- Передача данных в слой Dense размером 256 блоков. Данный слой является полносвязным.
- Передача данных в слой Dense. Полносвязный слой преобразует данные в выходной вектор, состоящий из 7 нейронов.

Итоговый алгоритм анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей выглядит следующим образом:

Шаг 1. Формирование обучающей выборки для обучения нейронной сети с использованием разработанного алгоритма совместного применения расширенного тезауруса WordNet-Affect и словаря авторских символов выражения эмоций.

Шаг 2. Векторизация обучающей выборки посредством применения языковой модели BERT.

Шаг 3. Обучение нейронной сети эффективной архитектуры на полученной обучающей выборке.

Шаг 4. Извлечение, предобработка и векторизация текстовых данных социальных сетей с использованием подходов, примененных при формировании обучающей выборки.

Шаг 5. Классификация предобработанных и векторизированных текстовых данных с использованием обученной на Шаге 3 нейронной сети на 7 классов тональности.

2.6. Выводы по главе

В рамках второй главы описаны разработанные алгоритмы, методы и подходы к обработке текстовой информации социальных сетей в задачах формирования социального портрета пользователя.

1. Предложен метод унификации данных различных социальных сетей на основе онтологического подхода, отличающийся выделением в качестве узлов онтологии пользователей социальных сетей, сущностей каждой социальной сети (группа, пост, комментарий, вложение) и объектов реального мира, а в качестве отношений онтологии – Object Properties (hasFriend, hasFollower) и DataType Properties (отношения между объектами онтологии).

2. Предложен оригинальный алгоритм формирования обучающей выборки, применяемый в задачах сентимент-анализа текстовых данных для обучения классификаторов, который отличается одновременным использованием словаря авторских символов выражения эмоций и расширенного словаря WordNetAffect.

3. Предложен оригинальный подход сопоставления профилей пользователей в различных социальных сетях, который отличается набором критериев сопоставления – гибридизацией подходов анализа графической информации, структурированных данных анкет, текстовых данных, а также социальных графов профилей.

4. Разработан метод определения психологических характеристик пользователя социальных сетей, отличающийся гибридизацией алгоритмов обработки естественного языка текстовых данных, машинного обучения и метода «Большой пятерки».

5. Предложен новый алгоритм анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей, который отличается использованием семантических подходов наряду с машинным обучением.

Глава 3. Реализация программного комлекса интеллектуального анализа текстовых данных социальных сетей на основе интеграции семантических подходов и машинного обучения

3.1. Общая концепция программного комплекса

3.1.1. Диаграмма развертывания программного комплекса

Разработанные алгоритмы были реализованы в программном комплексе, обеспечивающем поиск аккаунтов одного человека в разных социальных сетях, объединение и построение психологического портрета данного человека. Диаграмма развертывания программного комплекса представлена на рисунке 3.1.

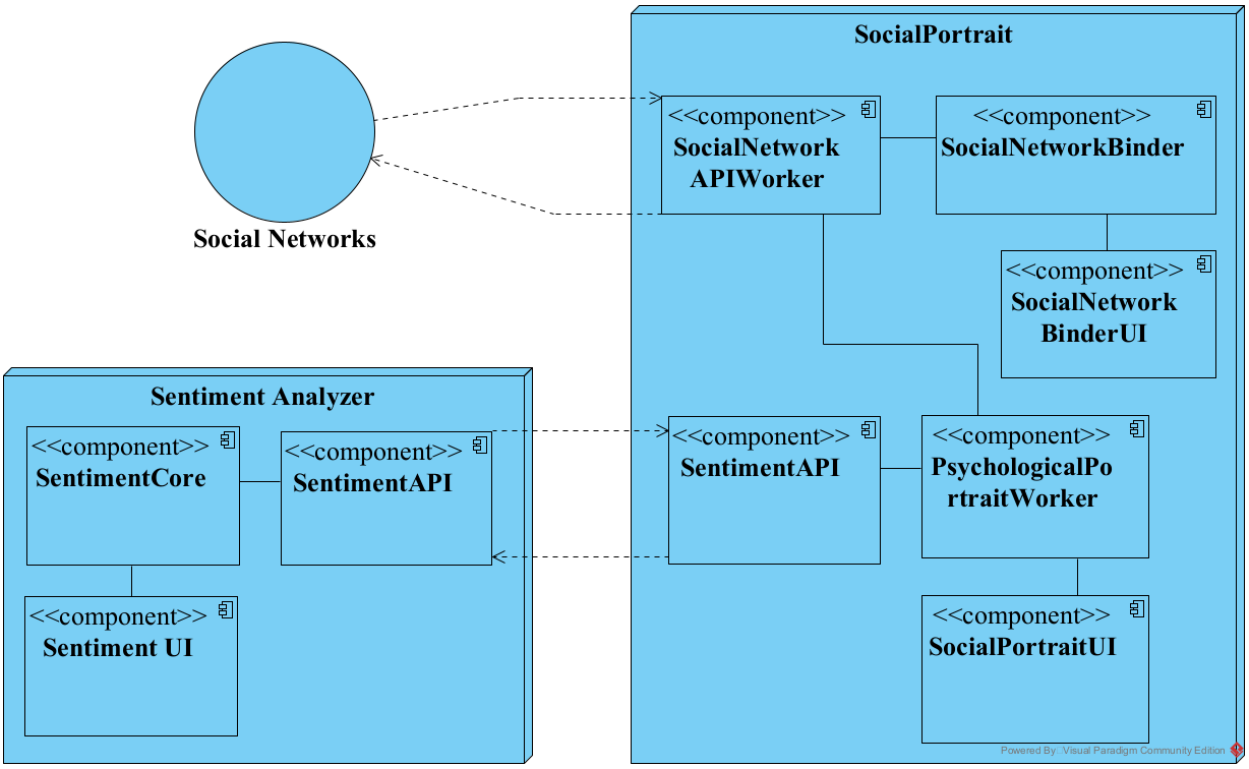


Рисунок 3.1. Диаграмма развертывания программного комплекса.

Отдельные части программного комплекса могут быть использованы независимо – так, сервис определения тональности текста может быть

использован при помощи API или отдельно, при помощи веб-интерфейса.

3.1.2. Подходы к извлечению данных из социальных сетей

Разрабатываемая платформа формирования социального портрета соискателя предполагает прямое обращение к социальным сетям в рамках двух процессов:

- Поиск профилей соискателя в разных социальных сетях на основе запроса, содержащего конкретные значения критериев поиска;
- Извлечение (получение) данных анкет и текстовых данных по ссылкам на необходимые профили разных социальных сетей.

Платформа предполагает автоматическое извлечение слабоструктурированных и неструктурированных данных из следующих социальных сетей:

- ВКонтакте (API – https://vk.com/dev);
- Facebook (API – https://developers.facebook.com);
- Одноклассники (API – https://apiok.ru/dev);
- Instagram (API – https://developers.facebook.com).

Каждая из перечисленных социальных сетей имеет собственный API, позволяющий извлекать определенный набор данных. Наборы полей, которые позволяют извлекать API, разные и определяют уровень функциональности API [52].

В связи с этим для обеспечения полноты и надежности извлечения данных из социальных сетей был разработан комплексный подход к извлечению данных, который предполагает использование как API социальной сети, так и метод прямого парсинга страниц профилей.

Для идентификации в API всех социальных сетей используется специальный ключ доступа, который называется access token. Токен – это строка из цифр и латинских букв, которая передаете на сервер вместе с запросом.

Для автоматизации процесса парсинга страниц социальной сети используется драйвер Selenium, который позволяет управлять поведением

браузера. Последовательность действий при извлечение данных из социальной сети представлена на рисунке 3.2.

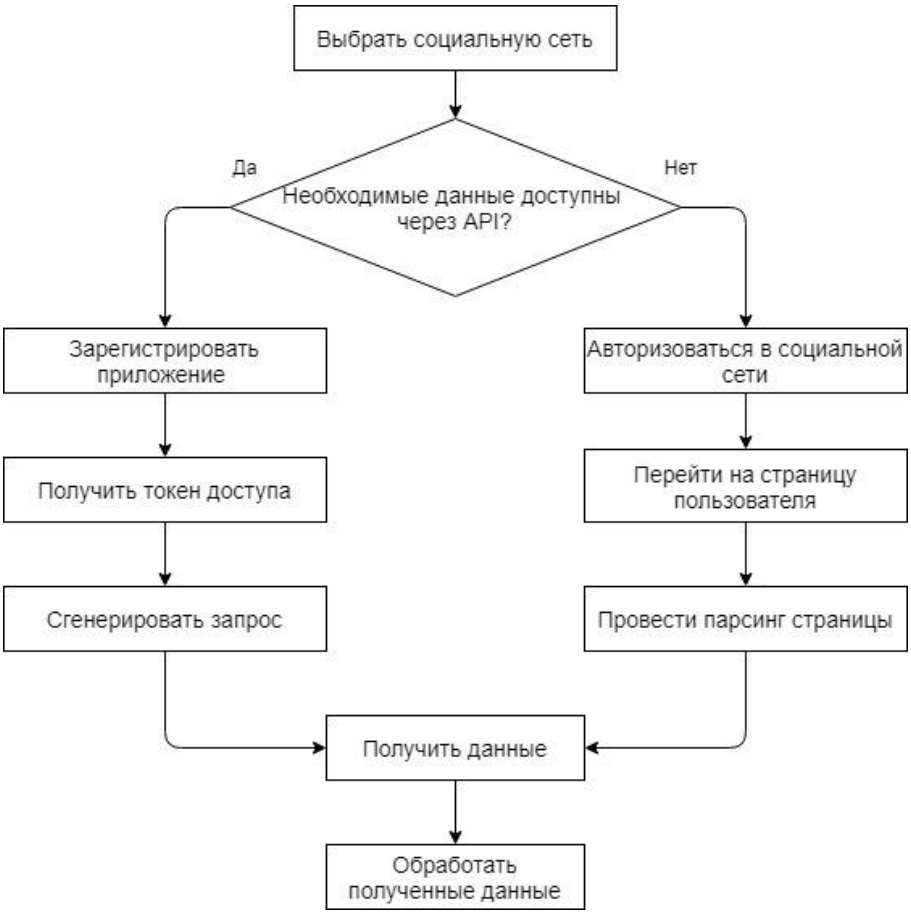


Рисунок 3.2. Извлечение данных из социальной сети

Также парсинг необходим в тех случаях, если профиль в социальной сети, через который обеспечивается подключение к API, будет заблокирован роботом социальной сети. Тем самым обеспечивается надежность получения необходимых данных. Необходимо отметить, что платформа имеет возможность анализировать только публичные данные со страниц пользователей, просмотр которых не ограничен настройками приватности.

3.2. Проектирование и реализация подсистемы анализа тональности текстов

3.2.1. Диаграмма вариантов использования системы анализа тональности текстов

Диаграмма прецедентов (вариантов использования) – диаграмма,

отражающая отношения между актерами и прецедентами и позволяющая описать систему на концептуальном уровне. Диаграмма вариантов использования системы анализа тональности текстов представлена на рисунке 3.4.

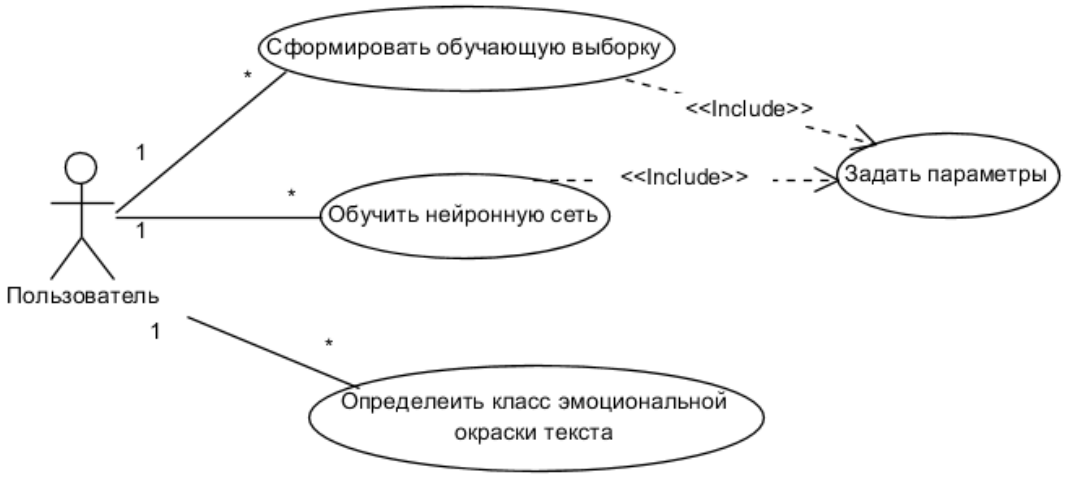


Рисунок 3.4 – Диаграмма вариантов использования системы сентимент-анализа

Диаграмма вариантов использования описывает действия пользователя в системе. Пользователь может сформировать обучающую выборку, обучить нейронную сеть и определить класс эмоциональной окраски текста. Действия сформировать обучающую выборку и обучить нейронную сеть включают в себя задание параметров. Конкретизация вариантов использования:

Использование 1. Сформировать обучающую выборку

1	Пользователь	Сформировать обучающую выборку	Позволяет пользователю сформировать обучающую выборку для обучения нейронной сети.

Основное действующее лицо: Пользователь.

Связи с другими вариантами использования: включает прецедент «Задать параметры»

Данный вариант использования позволяет пользователю сформировать обучающую выборку для обучения нейронной сети. Выборка формируется для семи человеческих эмоций. В результате пользователю

возвращается файл в формате CSV, так же данный файл передается в сервис определения эмоциональной окраски и сохраняется во временную директорию.

Использование 2. Обучить нейронную сеть

2	Пользователь	Обучить нейронную сеть	Позволяет пользователю обучить нейронную сеть сформированной выборкой

Основное действующее лицо: Пользователь.

Связи с другими вариантами использования: включает прецедент «Задать параметры»

Данный вариант использования позволяет пользователю обучить нейронную сеть сформированной выборкой. Файл с данными считывается из директории с временными данными. После этого происходит предобработка данных (лемматизация) и преобразование в вектор с помощью алгоритмов «BERT» или «word2vec». Модель нейронной сети обучается в несколько итераций с помощью алгоритма обратного распространения ошибки.

Использование 3. Определить класс эмоциональной окраски

3	Пользователь	Определить класс эмоциона-льной окраски	Позволяет пользователю определить класс эмоциональной окраски текста

Основное действующее лицо: Пользователь.

Связи с другими вариантами использования: отсутствует

Данный вариант использования позволяет пользователю определить класс эмоциональной окраски текста. Текст подается на вход нейронной сети, предобрабатывается и преобразуется в вектор, затем подается в нейронную сеть. На выходе получается класс эмоциональной окраски.

Использование 4. Задать параметры

2	Пользователь	Задать параметры	Позволяет пользователю задать параметры для формирования обучающей выборки и для обучения нейронной сети

Основное действующее лицо: Пользователь.

Связи с другими вариантами использования: включается в прецеденты «Сформировать обучающую выборку» и «Обучить нейронную сеть».

Данный вариант использования позволяет пользователю задать параметры для формирования обучающей выборки и для обучения нейронной сети. Пользователь может задать следующие параметры: использовать стемминг или лематизацию при формировании обучающей выборки, удалять стоп слова или нет, количество эпох обучения, использовать модель «BERT» или алгоритм «word2vec» для векторизации текста. Так же пользователь выбирает архитектуру нейронной сети.

Для лучшего понимания функциональных зависимостей были разработаны диаграммы последовательности для основных действий.

3.2.2. Диаграмма последовательности системы анализа тональности текстов

Диаграмма последовательности – диаграмма, на которой для набора объектов показан жизненный цикл определенного объекта и взаимодействие актеров ИС в рамках определенного прецедента [119].

Диаграммы последовательности показывают взаимодействие пользователя с приложением, а также показывают взаимодействия классов приложения.

Диаграмма последовательности, показывающая действия приложения при формировании обучающей выборки, показана на рисунке 3.5.

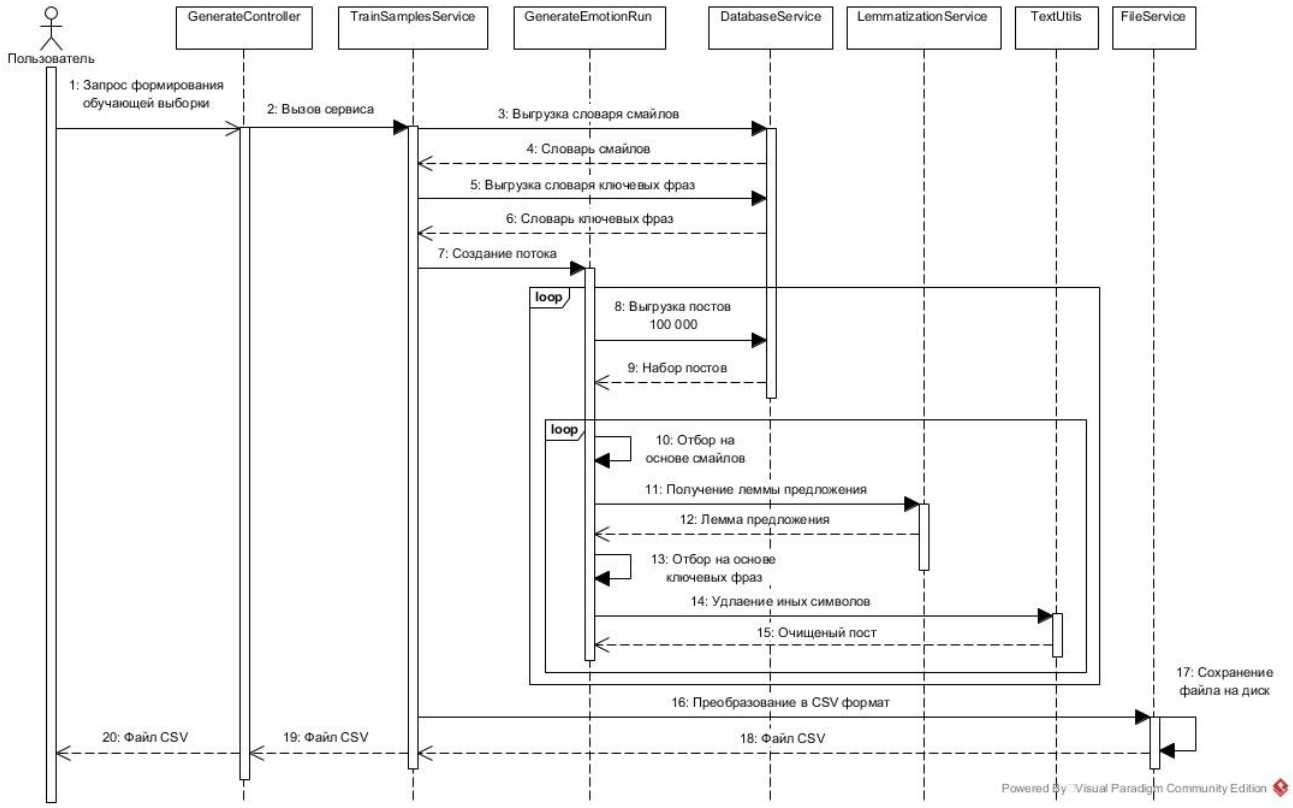


Рисунок 3.5 Формирование обучающей выборки

На диаграмме присутствует 2 цикла: один цикл (внешний) выгружает посты из БД по 100 тыс. штук, а второй (внутренний) обрабатывает текстовые сообщения по 1 посту за раз.

Диаграмма последовательности, показывающая действия приложения при обучении нейронной сети и определении тональности текста, показана на рисунке 3.6.

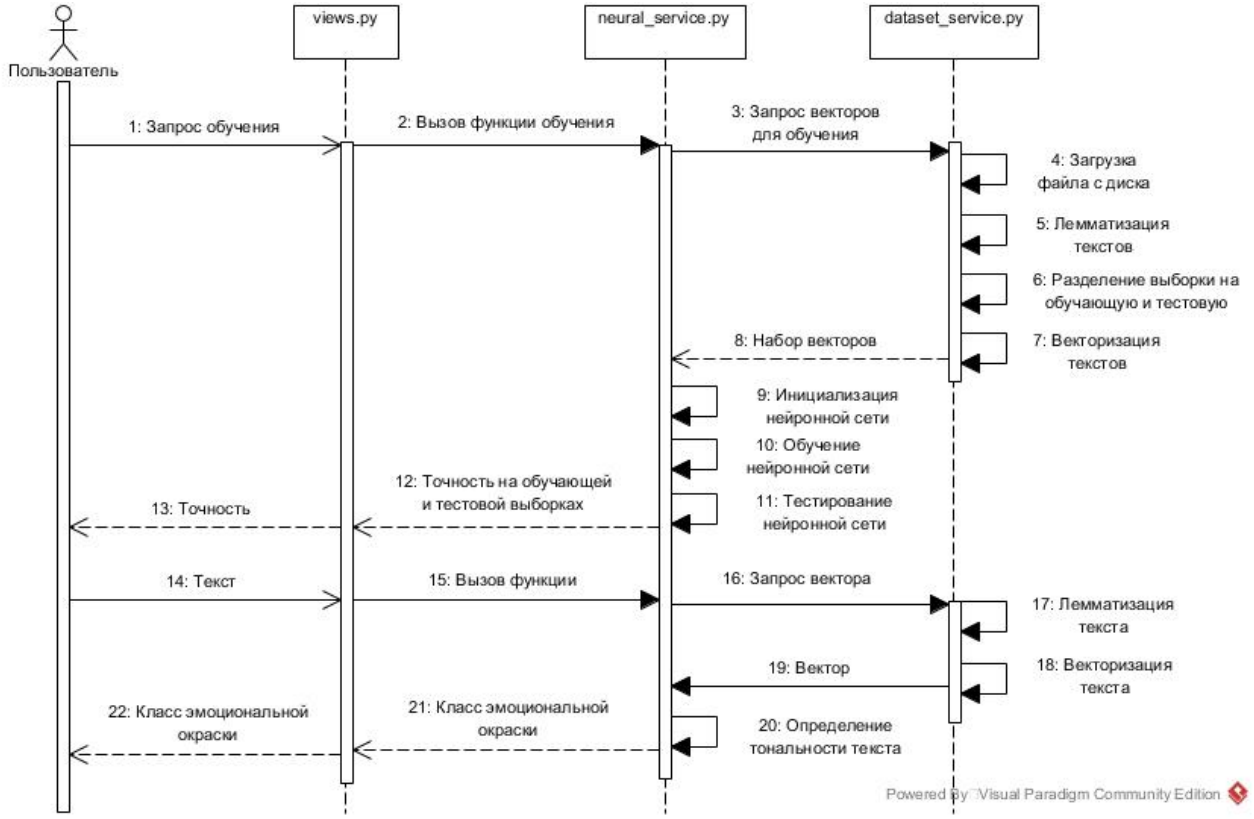


Рисунок 3.6 Обучение нейронной сети и определение тональности

3.2.3. Диаграмма классов системы анализа тональности текстов

Диаграмма классов – диаграмма, демонстрирующая классы системы, атрибуты классов, переменные классов, методы классов и взаимосвязи между ними. На диаграмме классов отображены классы разрабатываемой системы. Диаграмма классов разрабатываемой системы представлена на рисунке 3.7.

Система состоит из двух микросервисов: микросервиса, формиующего обучающую выборку на основе пользовательских данных и микросервиса определения тональности текстов. Рассмотрим микросервис формирования обучающей выборки.

Микросервис формирования обучающей выборки реализован с использованием фреймворка «Spring Framework» на языке программирования Java. Точкой входа приложения является класс SpringBootWebApplication, который запускает приложение и создает все классы-контроллеры и классы-сервисы.

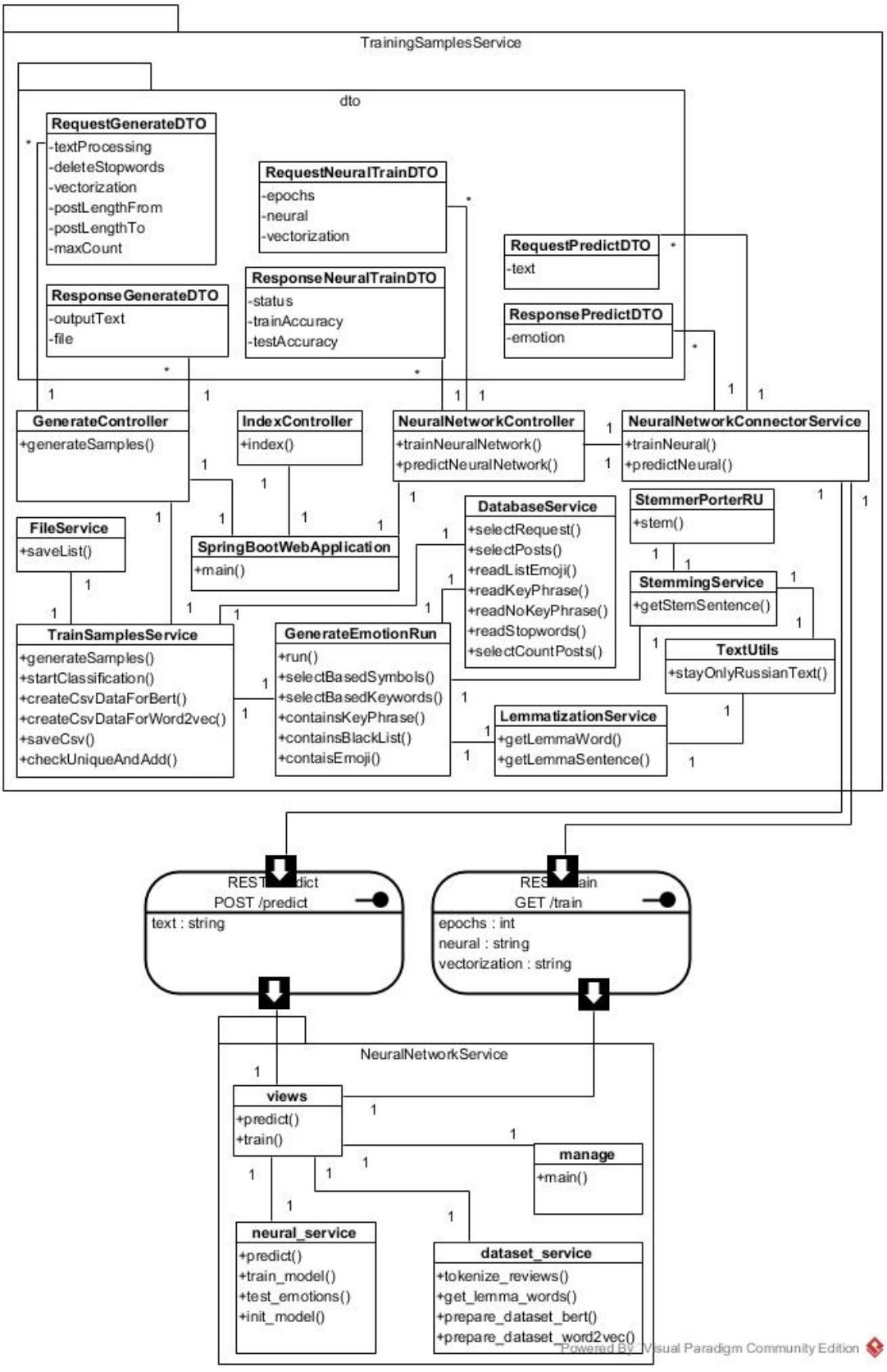


Рисунок 3.7 Диаграмма классов

При открытии приложения в браузере вызывается IndexController и пользователю возвращается web-страница с интерфейсом, через который происходит дальнейшее взаимодействие с системой. После выбора параметров в интерфейсе пользователь нажимает кнопку «Сгенерировать» и

система отправляет GET запрос в класс-контроллер GenerateController, в котором вызывается метод generateSamples(). Данный метод вызывает метод generateSamples() в классе-сервисе TrainSamplesService. Далее происходит выгрузка словаря авторских символов выражения эмоций, словаря ключевых фраз и словаря стоп-слов с помощью методов в классе DatabaseService, и вызов функции startClassification(), которая создает 7 потоков (по 1 потоку для каждой эмоции), инициализируя объекты класса GenerateEmotionRun.

В классе GenerateEmotionRun происходит отбор постов на основе словаря авторских символов выражения эмоций с помощью функции selectBasedSymbols() и отбор на основе словаря ключевых фраз с помощью функции selectBasedKeywords(). С помощью функции contaisEmoji() проверяется наличие авторского символа в тексте поста. С помощью функции containsKeyPhrase() проверяется наличие фразы в тексте поста.

С помощью функции containsBlackList() проверяется наличие фраз, которые не должны встречаться в тексте поста. Если фраза из «черного» списка встречается в тексте, то пост исключается из дальнейшей обработки. Данный список содержит такие слова, как «цена», «купить» и ссылки. Данный список нужен для исключения рекламных постов. Затем производится очистка постов от лишних символов (кроме символов кириллицы и пробелов) с помощью функции stayOnlyRussianText() класса TextUtils.

После того, как все 7 потоков отработали, создается выборка с помощью функций createCsvDataForBert() и createCsvDataForWord2vec(). Отличия данных методов заключается в том, что первый метод возвращает выборку с постами, количество символов в которых находится в указанном диапазоне (параметр в интерфейсе). А второй метод возвращает выборку с постами, количество слов в которых находится в указанном диапазоне. После этого выборка сохраняется на диск с помощью метода saveList() класса FileService и возвращается пользователю.

Классы DTO служат для взаимодействия интерфейса пользователя и сервисов между собой. Взаимодействие происходит с помощью REST

запросов в формате JSON. Параметры, передаваемые в запросах, представлены в классах DTO. Взаимодействие с сервисом определения тональности текстов выполняется с помощью класса NeuralNetworkConnectorService. Рассмотрим микросервис определения тональности текстов.

Микросервис определения тональности текстов реализован с помощью фреймворка «Django» на языке программирования Python. Точкой входа приложения является класс manage, который инициализирует все классы-сервисы и запускает приложение. В классе views описаны точки входа приложения.

При обучении нейронной сети вызывается функция train(), которая вызывает функцию train_model() класса neural_service. После этого происходит вызов метода prepare_dataset_bert() или метода prepare_dataset_word2vec() класса dataset_service (в зависимости от выбранного метода в интерфейсе). На данном этапе производится лемматизация текстов и приведение в векторную форму с помощью модели «BERT» или алгоритма «word2vec».

Для лемматизации предложений используется функция get_lemma_words(). Затем происходит инициализация модели с помощью функции init_model() и обучение нейронной сети. После обучения происходит тестирование точности модели на тестовой выборке с помощью функции test_emotions(). И затем точность модели на обучающей и тестовой выборках возвращается пользователю.

Для определения эмоции с помощью обученной модели нейронной сети используется функция predict(), которая вызывает функцию с таким же именем в классе neural_service. На данном этапе происходит лемматизация текста, преобразование в вектор (с помощью методов описанных выше) и подача в нейронную сеть. На выходе пользователю возвращается класс эмоциональной окраски текста.

3.2.4. Диаграмма «сущность-связь» системы анализа тональности текстов

Диаграмма «сущность-связь» – диаграмма, предназначенная для разработки моделей данных. Диаграмма отражает сущности и связи между ними. На основе ER-диаграммы строится структура БД.

ER-диаграмма системы анализа тональности текстов представлена на рисунке 3.8.

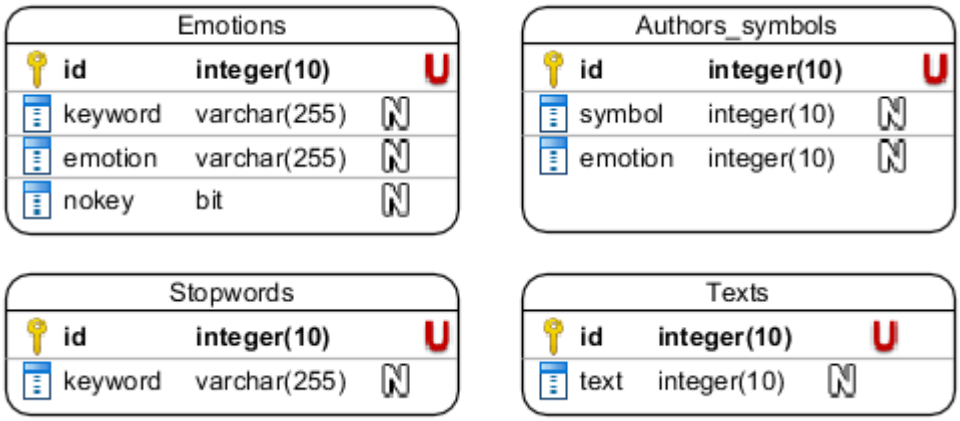


Рисунок 3.8 – Диаграмма «сущность-связь»

Таблица «Emotions» содержит словарь ключевых фраз. Ее атрибуты:

- «id» – идентификатор записи в таблице.
- «keywords» – ключевая фраза.
- «emotion» – название эмоции
- «nokey» – булевая метка для фраз, входящих в «черный список» – фразы, которые не должны учитываться[53].

Таблица «Authors_symbols» содержит словарь авторских символов выражения эмоций. Атрибут «id» – идентификатор записи в таблице. Атрибут «symbol» – символ или группа символов выражения эмоции. Атрибут «emotion» – название эмоции.

Таблица «Stopwords» содержит словарь стоп-слов. Атрибут «id» – идентификатор записи в таблице. Атрибут «keywords» – ключевое слово.

Таблица «Texts» содержит посты для обработки. Атрибут «id» – идентификатор записи в таблице. Атрибут «text» – текстовое сообщение (пост).

3.2.5. Диаграмма компонентов системы анализа тональности текстов

Диаграмма компонентов – структурная диаграмма, отражающая программную систему в виде модулей. На диаграмме, как правило, отражены файлы, библиотеки, модули, исполняемые файлы, пакеты.

Диаграмма компонентов системы анализа тональности текстов представлена на рисунке 3.9.

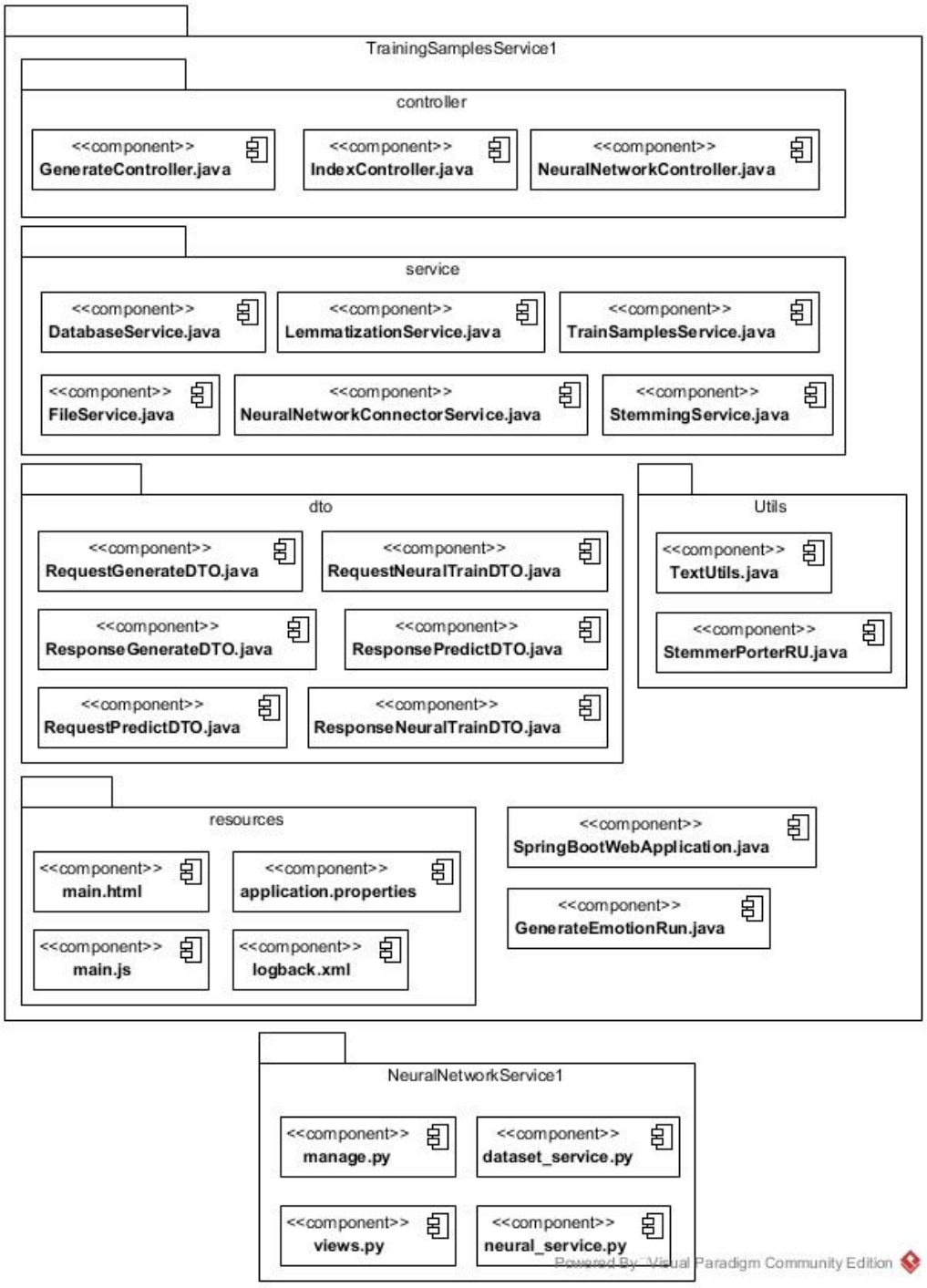


Рисунок 3.9 Диаграмма компонентов

Диаграмма компонентов отражает компоненты разрабатываемого приложения. При запуске приложения инициализируются классы-контроллеры и классы-сервисы. Классы-контроллеры используют экземпляры объектов DTO при работе системы. Файлы main.html и main.js служат для создания пользовательского интерфейса, файл logback.xml используется при работе логгера, а файл0application.properties хранит конфигурации приложения.

3.2.6. Диаграмма развертывания системы анализа тональности текстов

Техническое обеспечение системы представлено диаграммой развертывания. На диаграмме развертывания показаны аппаратные узлы, которые также будут входить в целевую систему наравне с программным обеспечением.

Диаграмма развертывания системы анализа тональности текстов представлена на рисунке 3.10.

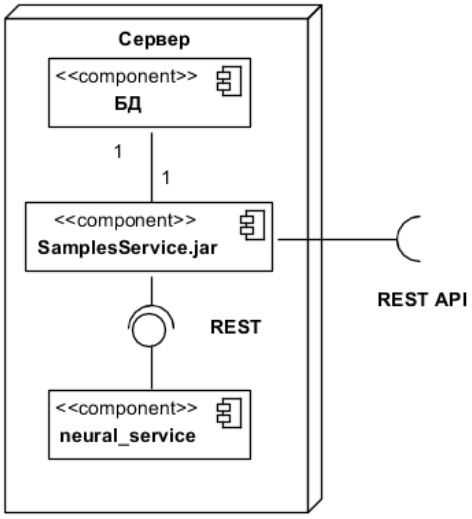


Рисунок 3.10 – Диаграмма развертывания

Микросервисы имеют REST архитектуру, и взаимодействие между ними происходит посредством запросов. Точкой входа системы служит микросервис формирования обучающей выборки, который перенаправляет запросы в микросервис определения тональности текстов. Микросервисы могут быть интегрированы в систему, имеющую REST архитектуру.

3.2.7. Диаграмма потоков данных системы анализа тональности текстов

Схема потоков данных приложения может быть отражена на диаграмме DFD. Диаграмма отражает процессы преобразования входных данных в выходные, а также показывает отношения между этими процессами [56]. Диаграмма потоков данных для разработанного приложения представлена на рисунке 3.11.

Рисунок 3.11 – Диаграмма потоков данных

3.2.8. Описание входных и выходных данных системы анализа тональности текстов

Схема, иллюстрирующая потоки входных и выходных данных каждого компонента системы, представлена на рисунке 3.12.

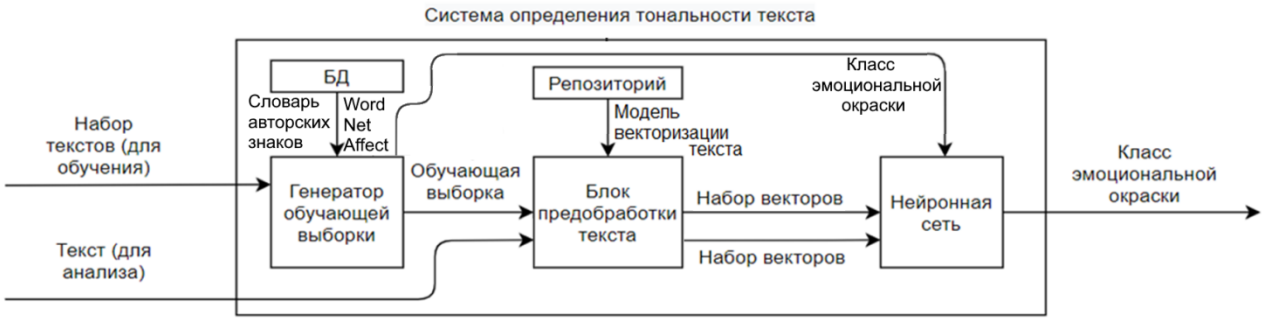


Рисунок 3.12 – Схема потоков входных и выходных данных

Исходными данными служат посты из открытых групп социальной сети «ВКонтакте» в количестве 2,5 млн. штук. Данные представлены в текстовом виде.

Обучающая выборка подается на вход блоку предобработки текста, в котором с помощью языковых моделей «word2vec» или «BERT» текст приводится к векторному виду.

Далее набор векторов подается на вход нейронной сети, где происходит обучение в несколько эпох. После того, как нейронная сеть обучена, на вход сети подается текст, а на выходе получается класс эмоциональной окраски текста. Классификация выполняется по 7-ми человеческим эмоциям.

3.2.9. Описание реализации подсистемы анализа тональности текстов

Для реализации подсистемы анализа тональности текстов была выбрана микросервисная архитектура. Сервис состоит из 2-х микросервисов: сервис формирования обучающей выборки и сервис сентимент-анализа.

Сервис формирования обучающей выборки был реализован на языке Java с использованием фреймворка Spring Framework. Словари, по которым производится отбор на разных этапах формирования обучающей выборки, хранятся в БД PostgreSQL. Так как текстовых сообщений в БД около 2,5 млн., посты выгружаются по 100 тысяч штук за раз. При формировании выборки происходит распараллеливание потоков, то есть для каждой эмоции выборка формируется в отдельном потоке. Всего 7 потоков. После формирования выборка предобрабатывается согласно описанному алгоритму. Результат работы алгоритма записывается в CSV файл и возвращается пользователю.

Сервис определения тональности текстов реализован на языке Python с использованием фреймворка Django. Сервис выполняет предобработку обучающей выборки перед подачей в нейронную сеть. Сначала все слова приводятся к начальной форме с использованием лемматизации, затем выполняется предобработка текстов с помощью алгоритма BERT. После происходит обучение нейронной сети, и затем оценка точности классификации текстов на тестовой выборке. Сервис предоставляет несколько нейронных сетей на выбор. После обучения мы можем подавать в

нейронную сеть тексты и получать класс эмоциональной окраски. Пользовательский интерфейс представлен на рисунке 3.13.

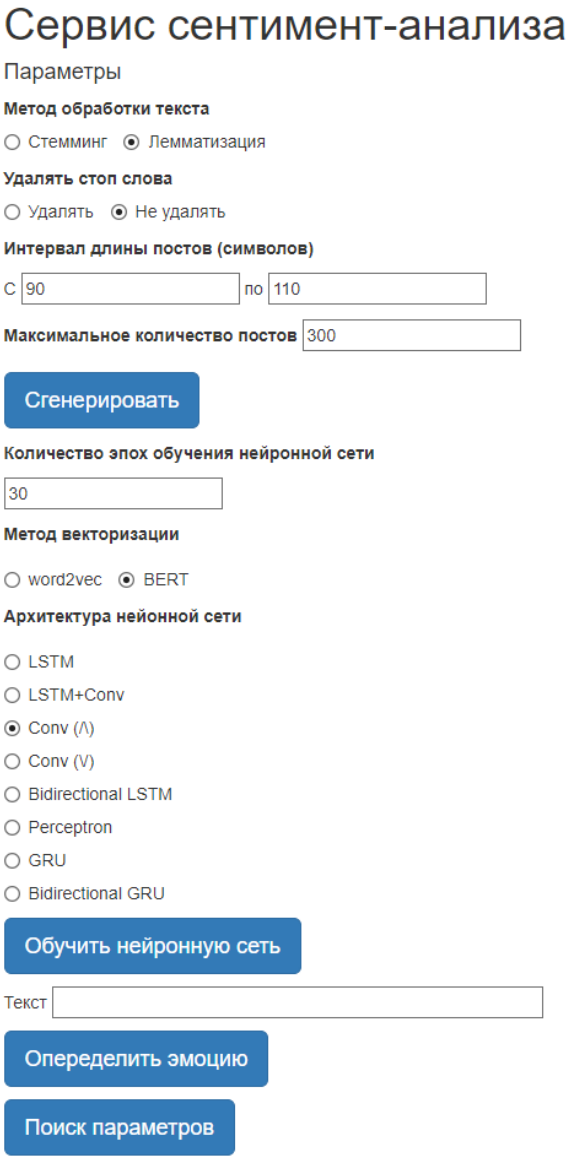


Рисунок 3.13 – Пользовательский интерфейс

Консольный вывод при формировании обучающей выборки представлен на рисунке 3.14.

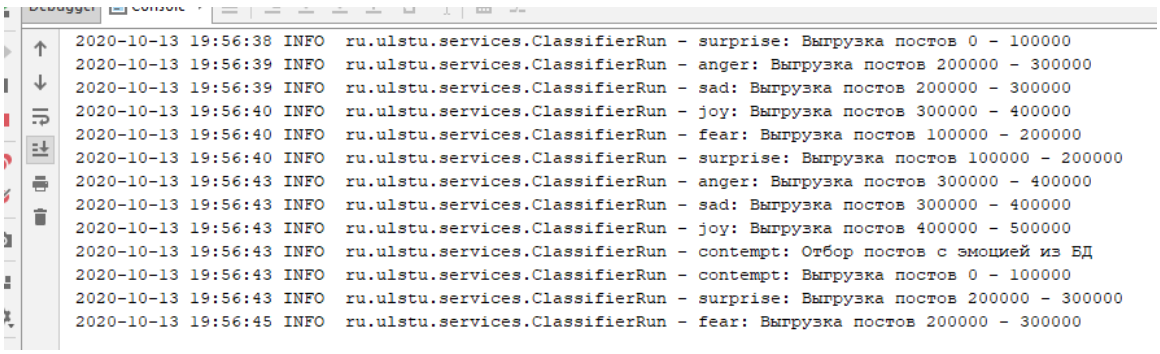


Рисунок 3.14 – Формирование обучающей выборки

Консольный вывод при обучении нейронной сети представлен на

рисунке 3.15.

```
Epoch 13/30
590/590 - 3s - loss: 0.1054 - accuracy: 0.9516
Epoch 14/30
590/590 - 3s - loss: 0.0894 - accuracy: 0.9586
Epoch 15/30
590/590 - 3s - loss: 0.0774 - accuracy: 0.9661
Epoch 16/30
590/590 - 3s - loss: 0.0839 - accuracy: 0.9676
Epoch 17/30
590/590 - 3s - loss: 0.0545 - accuracy: 0.9799
Epoch 18/30
590/590 - 3s - loss: 0.0478 - accuracy: 0.9799
Epoch 19/30
590/590 - 3s - loss: 0.0380 - accuracy: 0.9867
Epoch 20/30
590/590 - 3s - loss: 0.0249 - accuracy: 0.9942
Epoch 21/30
590/590 - 3s - loss: 0.0288 - accuracy: 0.9925
Epoch 22/30
590/590 - 3s - loss: 0.0266 - accuracy: 0.9913
Epoch 23/30
590/590 - 3s - loss: 0.0134 - accuracy: 0.9964
Epoch 24/30
590/590 - 3s - loss: 0.0119 - accuracy: 0.9978
Epoch 25/30
590/590 - 3s - loss: 0.0116 - accuracy: 0.9961
Epoch 26/30
590/590 - 3s - loss: 0.0107 - accuracy: 0.9966
Epoch 27/30
```

Рисунок 3.15 – Обучение нейронной сети

3.3. Проектирование и реализация программного комплекса психолингвистического анализа социальных сетей

3.3.1. Диаграмма вариантов использования программного комплекса психолингвистического анализа социальных сетей

Диаграмма вариантов использования отражает способы взаимодействия пользователя и системы. Диаграмма вариантов использования разрабатываемой системы представлена на рисунке 3.16.

Пользователь может осуществлять «Поиск профилей в социальных сетях», в результате данного действия он получит набор найденных профилей. Также пользователю доступно действие «Добавить профиль по ссылке», составной частью которого является «Выбор профилей для анализа». Также для выбранного профиля пользователь может осуществить «Поиск похожих профилей», который также выведет список найденных страниц. Пользователю доступен вариант использования «Удаление выбранного профиля», если какой-то из профилей он добавил по ошибке. Главным вариантом использования является «Анализ выбранных профилей»,

в результате которого пользователь должен получить характеристики человека.

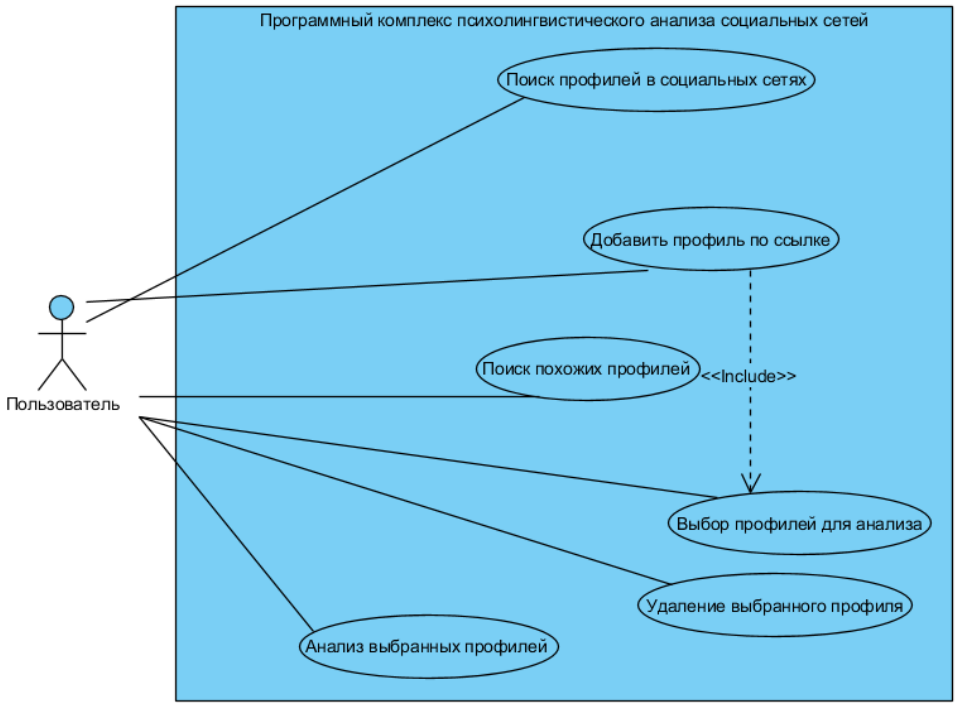


Рисунок 3.16. Use case диаграмма программного комплекса психолингвистического анализа социальных сетей

3.3.2. Диаграмма классов программного комплекса психолингвистического анализа социальных сетей

Диаграмма классов отражает статическое отображение системы при ее проектировании, показывая ее структуру и взаимодействие [88]. Так как разработанный комплекс основывается на клиент-серверном взаимодействии, а каждая из частей имеет достаточно сложную структуру и разный стиль проектирования, то было решено составить отдельные диаграммы классов для клиента и для сервера.

На рисунке 3.17 отображена диаграмма классов клиентской части приложения.

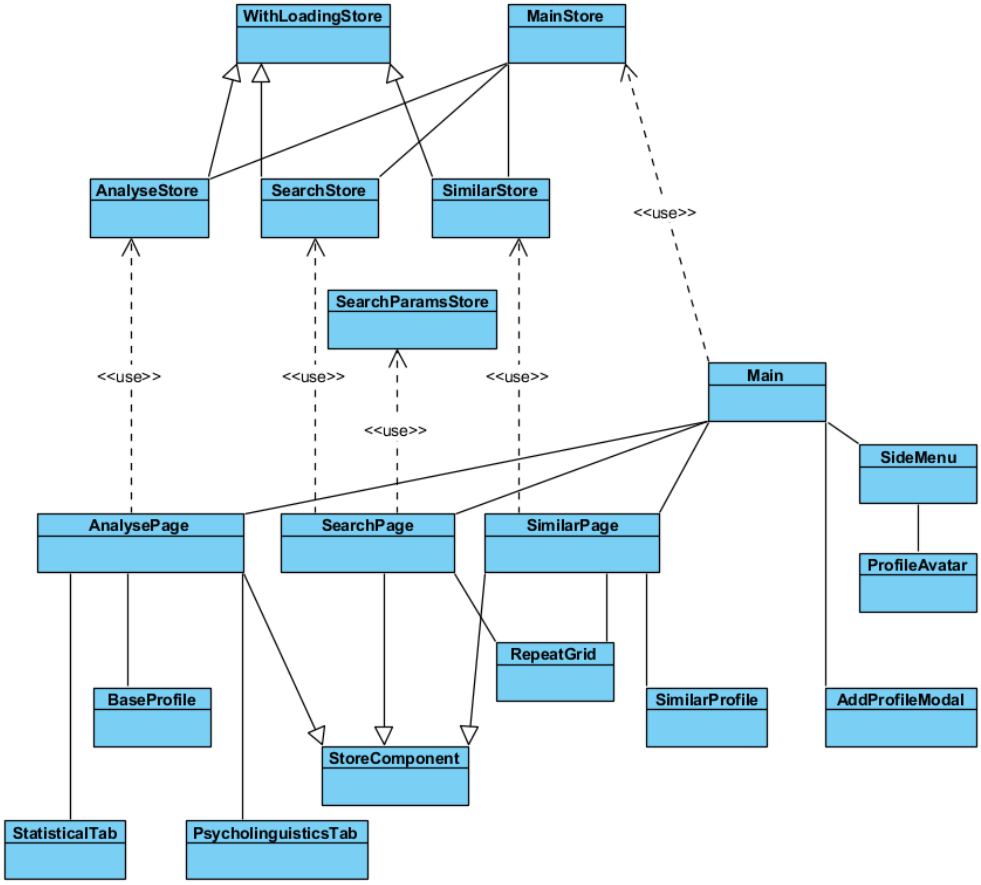


Рисунок 3.17. Диаграмма классов клиентской части приложения

Клиентская часть приложения разделена на классы компонентов и классы потоков данных. Первые отвечают за отображение приложения, вторые за хранение данных. Названия классов потоков данных оканчиваются на «Store», поэтому такие классы принято называть сторами. Оба типа классов имеют схожую архитектура в рамках приложения.

Имеется основной компонент Main, который связан отношением «use» с основным стором MainStore. Также имеются страницы SearchPage, SimilarPage, AnalysePage, которые также связаны со сторами SearchStore, SimilarStore, AnalyseStore соответственно. SearchPage – это класс компонента поиска профилей, SimilarPage – класс компонента поиска похожих профилей, а AnalysePage – класс компонента отображения анализа.

Все эти классы связаны отношением обобщения с классом StoreComponent, так как они расширяют его. Данный класс содержит методы доступа к классам потоков данных. Соответствующие сторы данных классов также связаны отношением обобщения с классом WithLoadingStore, который содержит в себе вспомогательные методы загрузки данных. Также на

диаграмме присутствует класс SideMenu, который отвечает за отображение бокового меню навигации по клиентской части приложения.

На рисунке 3.18 представлена диаграмма классов серверной части приложения.

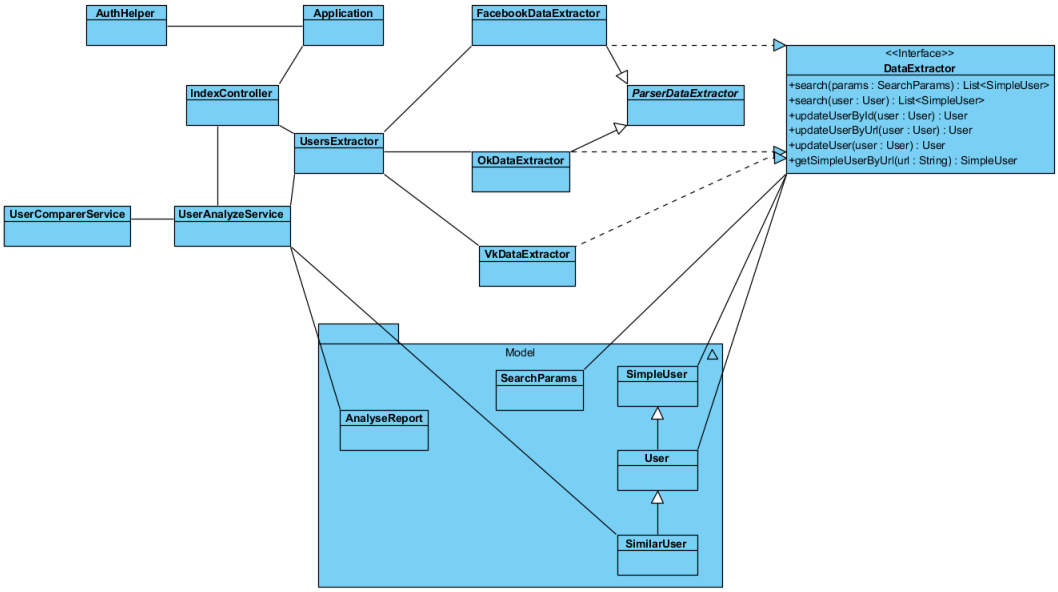


Рисунок 3.18. Диаграмма классов серверной части приложения

Серверная часть также разделена на классы модели данных и классы обработки логики приложения. Модель данных включает в себя классы:

- AnalyzeReport, который описывает результаты анализа пользовательских профилей;
- SearchParams, который описывает параметры запроса поиска;
- SimpleUser, описывающий основные свойства профиля пользователя;
- User, расширяющий SimpleUser, содержащий некоторые более подробные свойства профиля;
- SimilarUser, расширяющий User, содержащий дополнительные свойства, характерные для схожих профилей.

Основной класс приложения – это Application, он запускает приложение. При запуске приложения с помощью класса AuthHelper происходит процедура аутентификации в социальных сетях. В классе IndexController содержатся методы обработки клиентских запросов. За извлечение данных из социальных сетей «ВКонтакте», «Одноклассники» и

«Facebook» отвечают классы VkDataExtrator, OkDataExtractor и FacebookDataExtractor соответственно.

Данные классы реализуют интерфейс DataExtractor, содержащий методы поиска (search) для поиска с использованием параметров и поиска с использованием уже загруженного профиля, обновления профиля (updateUserById, updateUserByUrl, updateUser) для загрузки дополнительных полей профиля, а также извлечения базового профиля (getSimpleUserByUrl) для получения основных свойств профиля, описанных в классе SimpleUser. OkDataExtractor и FacebookDataExtractor также расширяют абстрактный класс ParserDataExtractor, содержащий методы работы с данными, загружаемыми способом парсинга HTML-страниц.

3.3.3. Диаграмма последовательности программного комплекса психолингвистического анализа социальных сетей

Диаграмма последовательности [88] фокусируется на обмене сообщениями между несколькими линиями жизни. На рисунке 3.19 представлена диаграмма последовательности для одной из возможных последовательностей использования системы. Основными этапами данной последовательности являются:

- Запрос поиска профилей;
- Добавление профиля по ссылке;
- Нахождение похожих профилей, также включает получение вектора изображений;
- Анализ профиля, также включает получение результатов классификации.

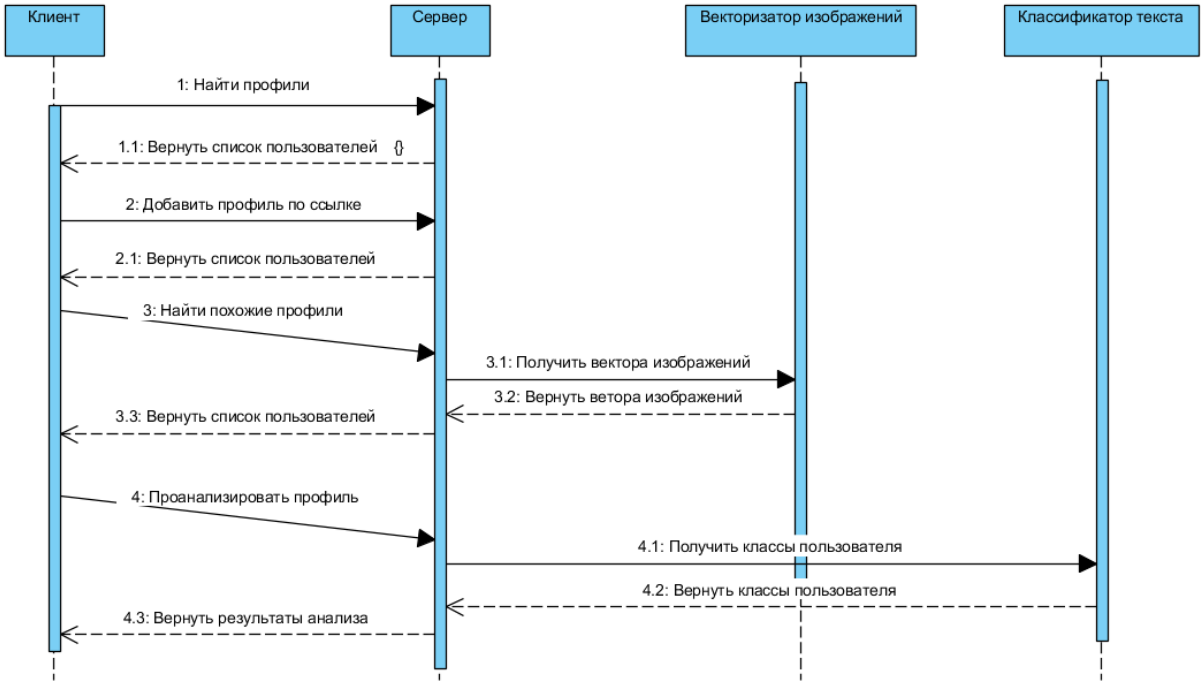


Рисунок 3.19. Диаграмма последовательности программного комплекса психолингвистического анализа социальных сетей

3.3.4. Диаграмма развертывания программного комплекса психолингвистического анализа социальных сетей

Диаграмма развертывания определяет набор конструкций, которые можно использовать для определения архитектуры выполнения систем, представляющих назначение программных артефактов узлам [139]. На рисунке 3.20 представлена диаграмма развертывания разрабатываемой системы.

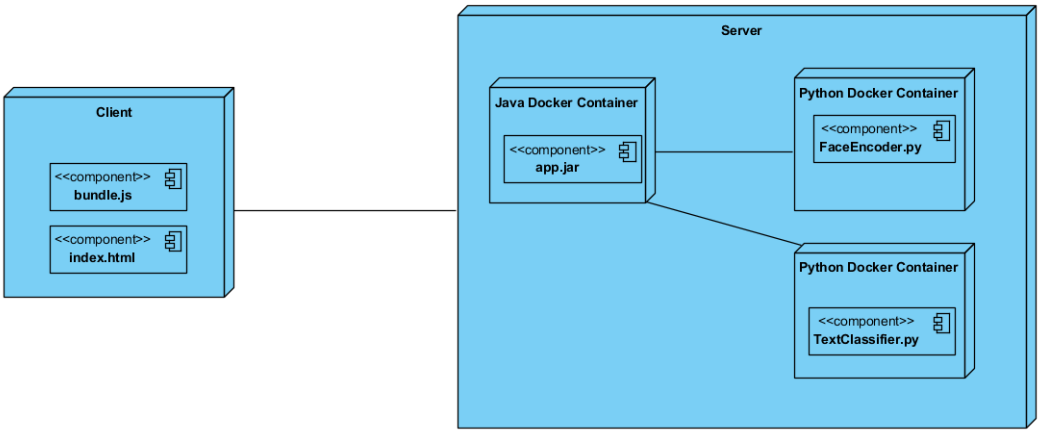


Рисунок 3.20. Диаграмма развертывания

На диаграмме можно увидеть два аппаратных узла: Client и Server.

Клиентский узел содержит в себе компоненты bundle.js и index.html, которые представляют собой собранное JavaScript приложение. Серверный узел включает три аппаратных узла, которые являются Docker контейнерами. Java Docker Container содержит компонент app.jar являющийся результатом сборки Java приложения. Два Python Docker Container содержат компоненты FaceEncoder.py и TextClassifier.py, которые являются Python приложениями для векторизации изображений и классификации текста соответственно.

3.3.5. Программная реализация клиентской части программного комплекса психолингвистического анализа социальных сетей

Клиентская часть системы представляет собой приложение, разработанное на языке JavaScript с использованием библиотеки React JS [2]. React JS – это декларативная, эффективная и гибкая библиотека JavaScript для создания пользовательских интерфейсов. Она позволяет создавать сложные пользовательские интерфейсы из небольших и изолированных частей кода, называемых компонентами. Данная библиотека существенно упрощает написание кода для клиентской части, воспроизводимой браузером, а также хорошо структурирует код за счет компонентного подхода.

Для упрощения взаимодействия с потоком данных была использована библиотека Mobx [48]. MobX – это библиотека, которая делает управление состоянием простым и масштабируемым за счет прозрачного применения функционально-реактивного программирования.

И React, и MobX предоставляют оптимальные и уникальные решения общих проблем при разработке приложений. React предоставляет механизмы для оптимальной визуализации пользовательского интерфейса с помощью виртуального DOM, что уменьшает количество дорогостоящих мутаций DOM. MobX предоставляет механизмы для оптимальной синхронизации состояния приложения с вашими компонентами React, используя реактивный

граф состояний виртуальной зависимости, который обновляется только в случае крайней необходимости и никогда не устареет.

В качестве библиотеки визуального дизайна использовалась библиотека Ant Design. Для отображения диаграмм и графиков использовалась библиотека ECharts, которая позволяет декларативным образом задавать настройки для отображения различной статистической информации, при этом всю сложность отрисовки и отображения библиотека берет на себя.

С разработанным пользовательским интерфейсом можно ознакомиться на рисунке 3.21. На данном рисунке можно увидеть левое боковое меню, через которое можно вызывать различные функции приложения, шапку с названием системы, а также содержимое текущей страницы, в данном случае это страница поиска профилей. На странице поиска профилей отображаются карточки с именами, изображениями найденных профилей, кнопкой добавления, а также параметры поискового запроса.

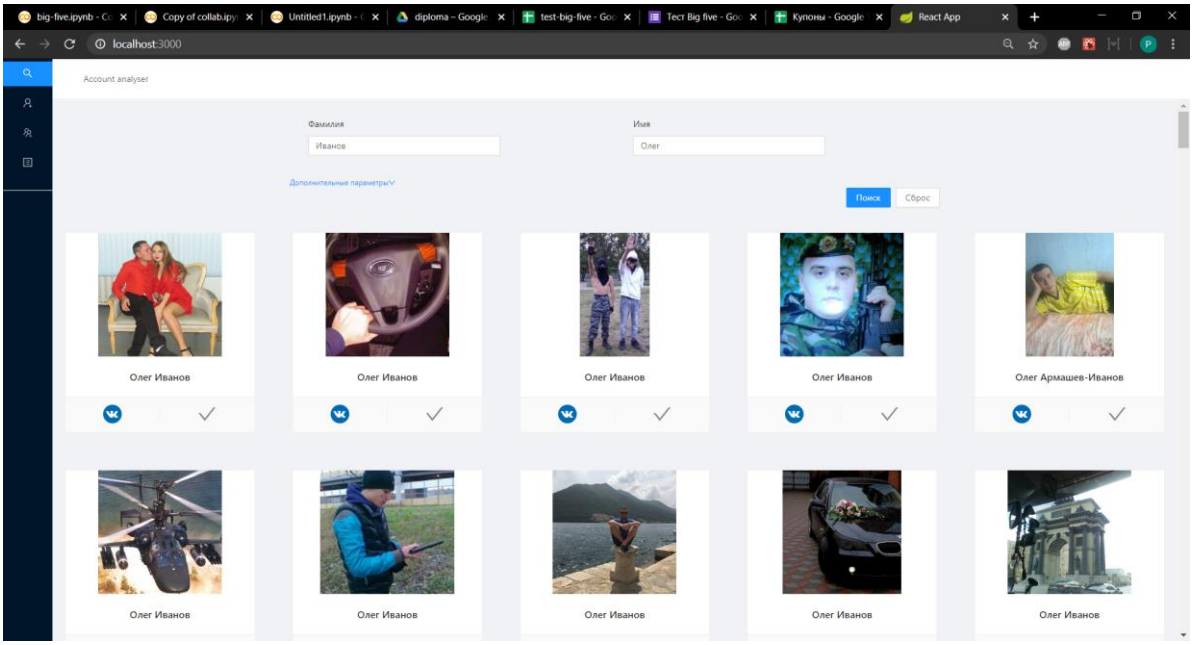


Рисунок 3.21. Страница поиска пользователей клиентского приложения

На рисунке 3.22 представлено окно добавления профиля пользователя по ссылке.

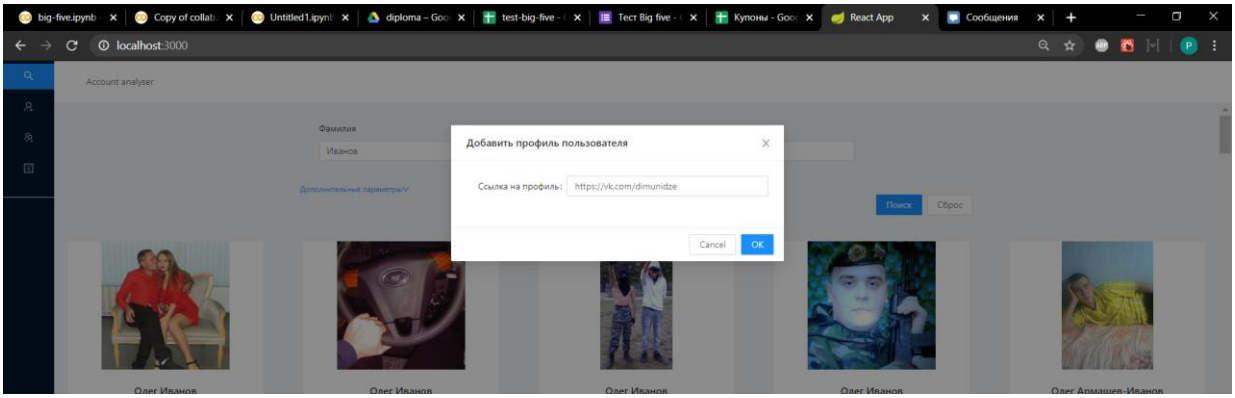


Рисунок 3.22. Окно добавления профиля по ссылке

В результате добавления профиля его иконка отображается в левом боковом меню вместе с иконкой социальной сети и кнопкой удаления выбранного пользователя (рисунок 3.23).

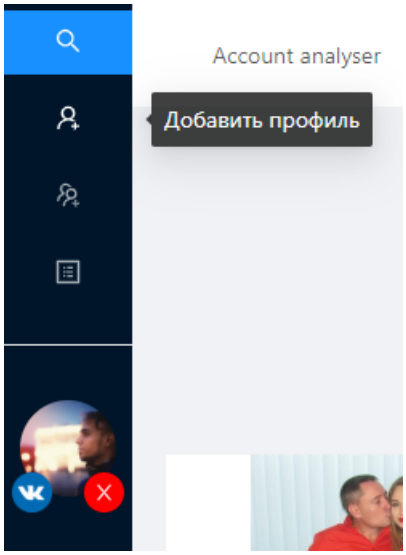


Рисунок 3.23. Иконка добавленного профиля в боковом меню

Страница поиска похожих профилей имеет практически полное совпадение со страницей простого поиска профилей, за исключением отсутствия параметров поиска (рисунок 3.24). Здесь карточки включают информацию о совпадающих данных из профилей: наличие общих контактов, образовательных учреждений, записей, изображений, мест работы, а также друзей.

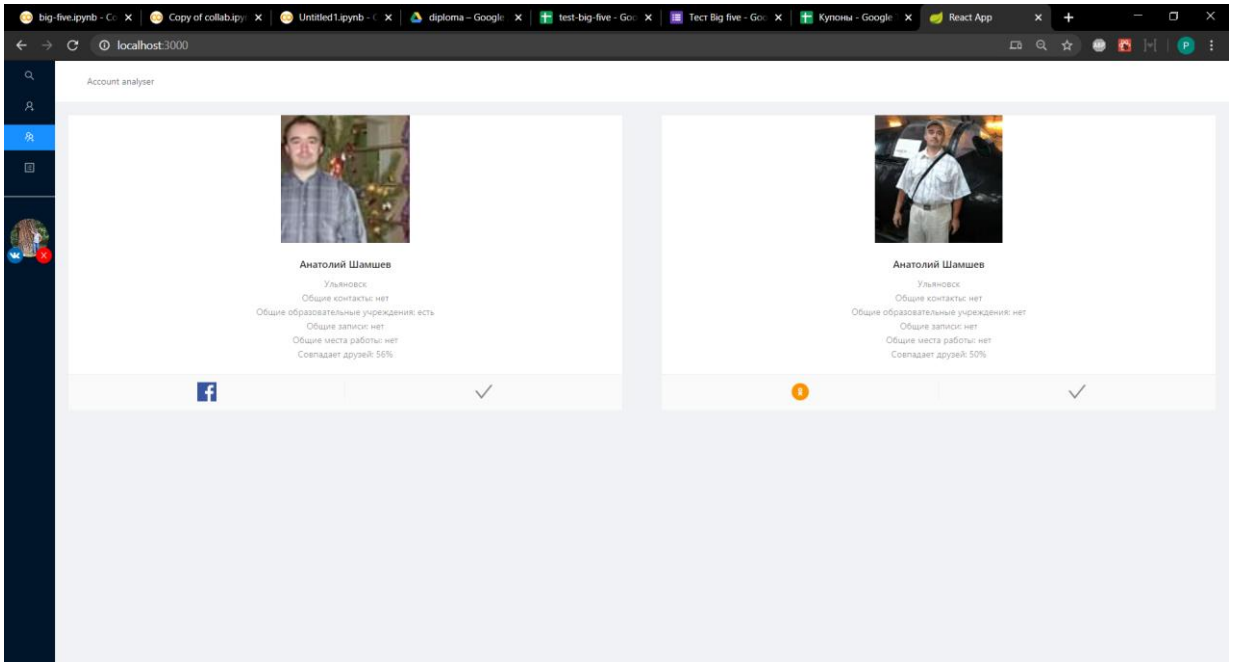


Рисунок 3.24. Страница поиска похожих профилей

Страница анализа выбранных профилей представлена на рисунке 3.25. На странице присутствуют две вкладки: «Статистика» и «Психолингвистика». На данном рисунке отображена первая из этих вкладок. Вкладка содержит графики, отображающие статистику по отметкам «нравится» («Лайки»), копиям («Репосты»), а также комментариям для записей данного пользователя на выбранных страницах на временной шкале.

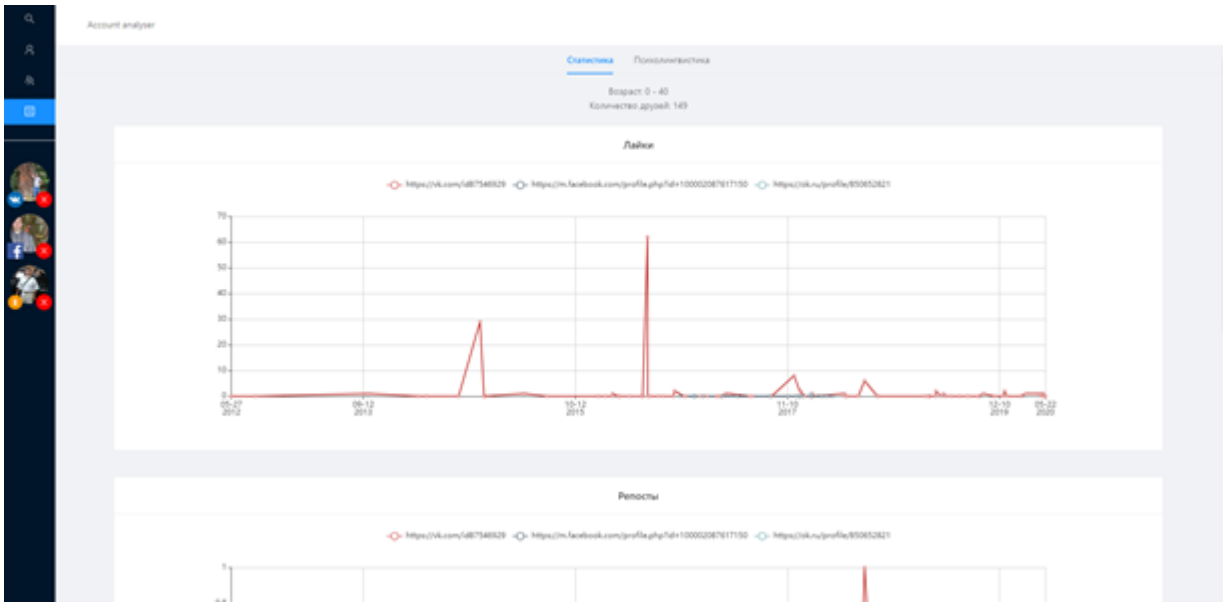


Рисунок 3.25. Страница анализа выбранных профилей

На вкладке «Психолингвистика» (рисунок 3.26) представлена информации по тому, какие психологические характеристики система предоставила пользователю в соответствии с пятифакторной моделью

личности на основании текстов, размещенных на выбранных страницах. Также представлена интерпретация данных характеристик для лучшего понимания.

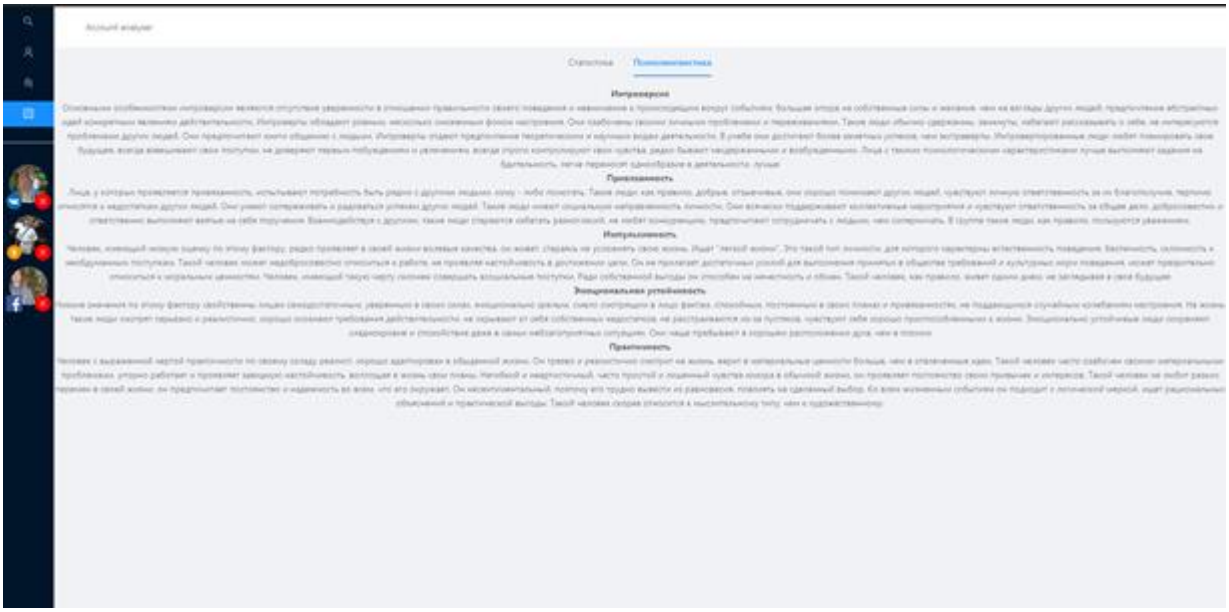


Рисунок 3.26. Вкладка «Психолингвистика» страницы анализа профилей

3.3.6. Программная реализация серверной части программного комплекса психолингвистического анализа социальных сетей

Программная реализация серверной части осуществлялась на языке Java с использованием фреймворка Spring Boot [44]. Spring Boot позволяет легко создавать автономные производственные приложения, которые легко и просто запускать. Большинству приложений Spring Boot требуется очень небольшая конфигурация. Основные преимущества данного фреймворка:

- Обеспечение радикально более быстрого начала разработки;
- Широкие возможности по настройке при значительно более легком старте;
- Ряд нефункциональных возможностей, которые являются общими для большинства проектов (таких как встроенные серверы, безопасность, метрики, проверки работоспособности и внешняя конфигурация);
- Встроенная поддержка сервера приложений позволяет

значительно упростить запуск приложений в продуктивной среде;

- Отсутствие генерируемого кода и настройки XML.

В соответствии с концепцией MVC приложение разделено на слои данных, управления данными и отображения. Так как за слой отображения данных отвечает клиентская часть приложения, то серверная часть приложения выполняет управление данными, а также содержит модели данных. Точками входа в систему являются методы контроллера: searchUser, analyse, getSimilar, getUserByUrl. Первый метод отвечает за поиск пользователей, второй – за анализ, третий – за поиск похожих, четвертый – за получения профиля по ссылке.

Для упрощения работы с моделями данных была использована библиотека Project Lombok. Данная библиотека позволяет избавиться от необходимости написания большого количества дублирующего кода доступа к данным вручную, так как библиотека самостоятельно его генерирует.

Для доступа к информации из социальной сети «ВКонтакте» использовалось официальное VK Java API, позволяющее получить доступ к данным социальной сети без необходимости генерации запросов самостоятельно.

Для получения данных из других социальных сетей использовалась библиотека jsoup [35]. jsoup – это библиотека для работы с реальным HTML. Она предоставляет очень удобный API для извлечения URL-адресов, а также для извлечения и обработки данных с использованием лучших методов HTML5 DOM и селекторов CSS. jsoup реализует спецификацию WHATWG HTML5 и анализирует HTML в том же DOM, что и современные браузеры.

Для морфологического и статистического анализа текстов использовались библиотеки из пакета apache.lucene.morphology [46]. Данные библиотеки использовались для лемматизации данных о работе, образовании пользователя.

3.3.7. Реализация классификатора текстов с целью определения психолингвистических характеристик автора

Программная система классификации текстов реализована на языке программирования Python. При разработке системы использовалась библиотека Scikit-learn [63]. Scikit-learn – это библиотека машинного обучения с открытым исходным кодом, которая поддерживает обучение с учителем и без учителя. Она также предоставляет различные инструменты для подбора моделей, предварительной обработки данных, выбора и оценки моделей, и многие другие утилиты.

Для обучения модели было решено воспользоваться сервисом Google Colab, который предоставляет серьезные вычислительные мощности для работы с методами интеллектуального анализа данных.

Входные данные загружаются в систему из размеченного CSV файла. Алгоритмы для обучения с учителем требуют, чтобы у каждого документа в обучающей выборке была пометка определенной категории. В нашем случае, категория – это одна из пяти характеристик метода «Большая пятерка». Так как в данных имеется пять бинарных независимых классов, то было решено обучить такое же количество классификаторов.

Исходная выборка включала в себя данные психологического опроса пользователей, а также данные профилей социальных сетей этих пользователей.

После фильтрации выборки необходимо было произвести предобработку текста. Для удаления стоп-слов была использована библиотека nltk.corpus.stopwords, которая содержит в том числе русскоязычный словарь. Лемматизация была произведена с помощью библиотеки mystem [24]. Данная библиотека также поддерживает русскоязычные тексты и отличается хорошим качеством лемматизации, используя для этого как словарь для слов, содержащихся в нем, так и методы машинного обучения для лемматизации неизвестных слов.

Для токенизации, очистки пунктуации и подсчета вхождений

используется класс CountVectorizer, который поддерживает подсчет N-грам слов или последовательностей символов и строит словарь индексов признаков. Значение индекса слова в словаре связано с его частотой употребления во всем обучающем корпусе. Для определения TF-IDF [115] меры текста также используется высокоуровневый компонент TfidfTransformer.

Для упрощения работы с цепочкой vectorizer => transformer => classifier в scikit-learn есть класс Pipeline, который функционирует как составной (конвейерный) классификатор, который был применен для цепочки CountVectorizer => TfidfTransformer => классификатор. Scikit-learn включает в себя классификатор методом опорных векторов SGDClassifier, а также классификатор методом случайного леса RandomForestClassifier.

Оптимальные параметры работы алгоритма классификации были выбраны с использованием компонента GridSearchCV. Экземпляр grid search ведет себя как обычная модель scikit-learn и с помощью метода полного перебора подбирает оптимальные параметры алгоритма классификации, в нашем случае.

Для получения метрик точности, полноты и правильности (accuracy) созданной модели использовалась библиотека sklearn.metrics.

3.3.8. Реализация непрерывной интеграции и доставки в программном комплексе психолингвистического анализа социальных сетей

С целью упрощения разработки и разворачивания системы было решено воспользоваться современными принципами непрерывной интеграции и доставки кода. Они включают в себя сервисы непрерывной интеграции, хранения образов, сборки приложения.

На первом шаге производится сборка приложения с помощью Maven в jar файл. На втором шаге производится сборка docker образа приложения и его отправка в хранилище образов. На третьем шаге производится

копирование данного образа в хранилище образов Heroku, а также запуск процедуры разворачивания внутри Heroku. К сожалению, бесплатный аккаунт Heroku позволяет использовать только 512 мегабайт оперативной памяти, чего может быть недостаточно для некоторых функций приложения. Тем не менее, сохранение образа в хранилище Gitlab позволяет использовать данный образ на любой виртуальной машине.

3.4 Выводы по главе

В рамках третьей главы дано описание разработанного программного комплекса интеллектуального анализа текстовых русскоязычных данных социальных сетей, обеспечивающего логический вывод результатов психолингвистического и сентимент-анализа данных профилей пользователей. Программный комплекс состоит из двух модулей.

Первый модуль – подсистема анализа тональности текстов, использующая семантические подходы и методы машинного обучения. В качестве обучающих данных выступает выборка постов социальной сети, основанная на словаре авторских символов выражения эмоций и ключевых фраз.

Второй модуль – подсистема психолингвистического анализа социальных сетей, состоящая из двух компонент – компоненты объединения профилей в различных социальных сетях, принадлежащих одному человеку, позволяющая в автоматизированном режиме найти все профили интересующего человека, и компоненты определения психолингвистических характеристик автора, основанная на модели личности «Большая пятерка» и позволяющая получить психологический портрет пользователя с указанием отношения к определенному объекту (при взаимодействии с первым модулем).

Глава 4. Анализ адекватности разработанных моделей и методов на основе вычислительных экспериментов и практики применения

4.1. План экспериментов

Для оценки эффективности разработанных методов и алгоритмов был проведен следующий ряд экспериментов на свободных русскоязычных данных социальных сетей:

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

В рамках первого набора экспериментов проведено исследование эффективности работы алгоритма поиска схожих профилей в социальных сетях. Для этого составлена выборка профилей, в которой будут находиться страницы профилей различных людей, имеющих аккаунты в некоторых социальных сетях. Далее для этих профилей применен разработанный алгоритм и собраны данные по тому, сколько и в каких сетях данный аккаунт нашел дубликатов искомого профиля. Была оценена корреляция успешной работы алгоритма на различных социальных сетях, а также по различным критериям сопоставления.

2. Эксперименты по оценке алгоритма формирования обучающей выборки.

Для этого для текстовых сообщений, извлеченных из социальной сети «ВКонтакте», был применен предложенный алгоритм и получена обучающая выборка, состоящая из неструктурированных данных, классифицированных по семи человеческим эмоциям, для обучения нейронной сети. Была проведена оценка работы алгоритма формирования обучающей выборки с различными параметрами и определены наилучшие параметры для формирования выборки и обучения классификатора.

3. Эксперименты по сентимент-анализу текстовых данных.

В следующем наборе экспериментов проведено исследование точности классификации текстов с помощью нейронных сетей. Для этого было проведено сравнение несколько архитектур нейронных сетей для классификации текстов по их эмоциональной окраске. Обучающая выборка была сформирована с наилучшими параметрами, полученными в первом эксперименте.

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

Эксперименты проводились на выборке из участников психологического опроса по методу «Большая пятерка», содержащей загруженные данные профилей этих людей. Была проведена оценка результатов классификации не только на текстовой информации, но и на другой метаинформации профилей.

4.2. 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, так как не все из них имели аккаунты во всех сетях сразу. Для каждого из этих аккаунтов мы попытались подобрать похожие с помощью разработанного сервиса. В результате экспериментов была составлена диаграмма, представленная на рисунке 4.1.

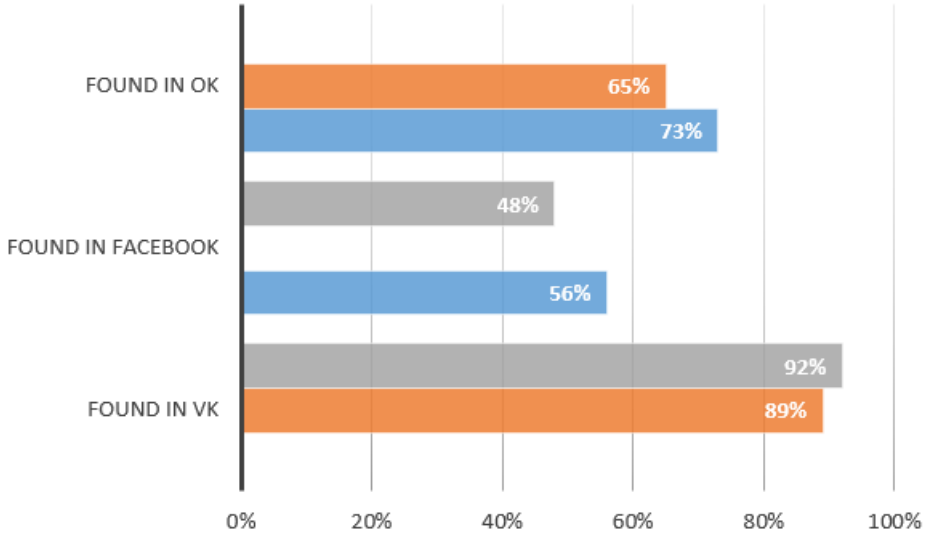


Рисунок 4.1. Диаграмма процентного соотношения найденных профилей

На диаграмме можно увидеть, что лучше всего система справилась с нахождением профилей в социальной сети VK, а хуже всего – Facebook. Это связано с удобством извлечения данных из соответствующих ресурсов. VK API позволяет быстро извлекать большие объемы данных, что увеличивает качество распознавания, в то время как парсинг других сетей отнимает множество ресурсов, что вынуждает ограничивать количество извлекаемых данных. Были подсчитаны результаты для критериев сравнения профилей, результат показан на рисунке 4.2.

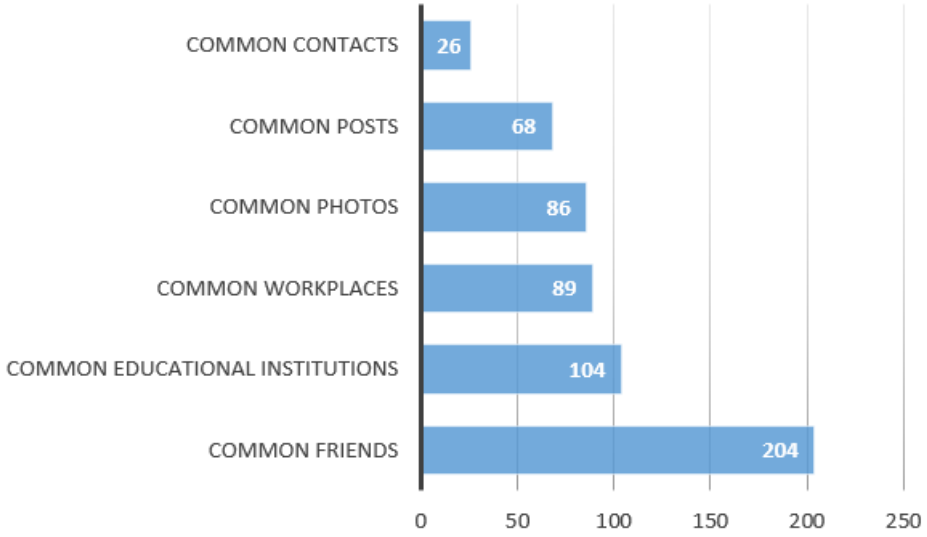


Рисунок 4.2. Диаграмма критериев сравнения профилей

На диаграмме можно увидеть, что во всех случаях системе удалось найти хотя бы одного общего друга. Результаты же совпадения других критериев гораздо меньше. В два раза хуже удавалось находить общие

образовательные учреждения и общие места работы. Это связано с тем, что пользователи реже указывают эти данные на своих страницах. Также зачастую формат указанных данных не позволяет корректно сопоставить их.

Еще меньше система справилась с нахождением общих лиц на фотографиях. Это обусловлено множеством факторов, таких как, например, точность самой модели, качество и количество выгружаемых фотографий. Только в трети экспериментов удалось найти общие записи на страницах. Это связано с тем, что не всегда пользователи заполняют страницы одинаковыми постами. Меньше всего было найдено перекрестных ссылок на профили, так как такую информацию пользователи предоставляют реже всего.

4.3. Эксперименты по сентимент-анализу текстовых данных

4.3.1. Эксперименты по оценке алгоритма формирования обучающей выборки

Для оценки эффективности алгоритма формирования обучающей выборки было обработано около 2,5 миллионов текстовых сообщений социальной сети. Текстовые сообщения были загружены через API «ВКонтакте» из открытых групп.

Из сформированного множества текстов была сформирована обучающая и тестовая выборка. Обучающая выборка содержала 70% постов, тестовая выборка содержала 30% постов.

В ходе проведения экспериментов были отобраны посты, длина которых находилась в заданном интервале (от 90 до 110 символов или от 40 до 50 слов). При формировании векторов короткие вектора дополняются нулями до максимальной длины вектора, из-за этого нейронная сеть не могла обучиться.

Точность классификации на обучающей выборке равна 1.0 для всех

экспериментов.

Так же была проверена следующая гипотеза: «Обучающая выборка, сформированная с использованием словаря авторских символов выражения эмоций и словаря ключевых фраз, имеет более высокое качество, чем выборка, сформированная с использованием только словаря ключевых фраз или выборка, сформированная с использованием словаря только авторских символов выражения эмоций».

4.3.2. Статистика этапов формирования обучающей выборки

При формировании обучающей выборки на каждом этапе отбора выводилось количество сообщений в каждой группе. На каждом этапе отбора выборка уменьшалась в среднем в 2-3 раза. Выборка содержала посты различной длины. Количество сообщений для каждой группы после двух этапов отбора представлено в таблице 4.1.

Таблица 4.1 – Статистика формирования бучающей выборки

Эмоция	Сообщений после 1 этапа	Сообщений после 2 этапа
Отвращение	5011	1206
Страх	4640	2435
Радость	237838	74307
Грусть	7271	2628
Удивление	2738	1534
Злость	1362	511
Презрение	9961	5613

Проведенный эксперимент обучения выборками со стоп словами и без стоп слов показал большую точность выборки со стоп словами.

4.3.3. Оценка разных языковых моделей при формировании обучающей выборки

Для определения более подходящего под поставленные задачи метода векторизации была проведена серия экспериментов с использованием языковых моделей Word2Vec и BERT. Обучающая выборка была составлена

на основе авторских символов выражения эмоций и ключевых фраз. Архитекстуры нейронной сети были использованы различные, но в финальном эксперименте – ранее показавшая лучший результат архитектура нейронной сети гибридного подхода.

По результатам эксперимента точность составила 79% при использовании языковой модели Word2Vec и 87% при использовании языковой модели BERT.

Эксперимент показал, что текст, векторизованный языковой моделью BERT, является более качественными исходными данными для сентимент-анализа текста на русском языке, чем текст, векторизованный языковой моделью Word2Vec.

4.3.4. Оценка использования разных словарей формирования обучающей выборки

Проведенные эксперименты показывают, что сформированная обучающая выборка имеет высокий уровень качества, необходимый для корректного обучения нейронной сети для классификации постов социальной сети. Опробованные архитектуры нейронных сетей показывают, что высокую точность классификации могут иметь не только нейронные сети глубокого обучения, но и полносвязные нейронные сети.

Лучший результат классификации оказался при использовании многослойного персептрона для классификации текстов. Точность – 87%.

Сравнение с классическим методом показывает, что разработанный подход значительно превосходит классический метод классификации – наивный байесовский классификатор.

Так же разработанный подход на 10-15% процентов превосходит методы, опубликованные в научных работах, статьях. В литературном обзоре представлены 2 похожие работы. В работе [7] применяется разметка текстов с использованием смайлов, точность классификации – 0,76. В работе [14] для классификации текстов применяются нейронные сети, точность классификации – 0,71.

4.3.5. Оценка использования разных языковых моделей, словарей формирования обучающей выборки, количества постов и длин сообщений

Для проверки сформулированной гипотезы было сформировано три обучающих набора на основе словаря ключевых фраз и словаря авторских символов выражения эмоций:

1. выборка, сформированная на основе двух словарей;
2. выборка, сформированная на основе второго словаря;
3. выборка, сформированная на основе первого словаря.

Результаты экспериментов приведены в таблице 4.2.

Таблица 4.2 – Результаты экспериментов

Языковая модель	Кол-во постов	Словари при получении обучающей выборки	Выборка сбалансирована	Веса классов	Длина сообщения	Точность на тестовой выборке
word2vec	1042	смайлы и ключевые слова	нет	нет	40-50 слов	0,77
word2vec	1042	смайлы и ключевые слова	нет	да	40-50 слов	0,79
BERT	556	смайлы и ключевые слова	нет	нет	90-110 символов	0,86
BERT	556	смайлы и ключевые слова	нет	да	90-110 символов	0,87
BERT	726	смайлы	нет	да	90-110 символов	0,82
BERT	2100	ключевые слова	да	да	90-110 символов	0,83
BERT	513	смайлы и ключевые слова, без стоп слов	нет	да	90-110 символов	0,82

Проведенные эксперименты показывают, что наивысшая точность классификации текстовых сообщений социальных сетей достигается:

- при использовании модели «BERT» для преобразования текстов в

вектора;

- при заданных весах классах (когда выборка не сбалансирована, задают веса классов, чтобы указать приоритет объектов в обучающей выборке).

Наивысшая точность составила 87%. Эксперименты показывают, что выборка, сформированная с использованием двух словарей является более качественной в рамках задачи обучения нейронной сети и превосходит выборки, использующие данные словари по отдельности.

4.3.6. Оценка разных архитектур нейронных сетей в задаче сентимент-анализа текстовых ресурсов

Эксперимент предполагает сравнение точности классификации постов социальной сети с помощью различных архитектур нейронной сети. Для формирования обучающей выборки используются результаты предыдущих экспериментов: длина поста 90-110 символов, стоп-слова не подлежат удалению, векторизация с использованием модели BERT.

Для проведения экспериментов была сформирована тестовая выборка, состоящая из текстов, которые не присутствовали в обучающей выборке, при этом точность для обучающей выборки составила 1.0. Результаты экспериментов представлены в таблице 4.3.

Таблица 4.3 – Результаты экспериментов

Нейронная сеть	Точность на тестовой выборке
CNN	0,85
LSTM	0,82
Bidirectional LSTM	0,84
MLP	0,87
GRU	0,81
Bidirectional GRU	0,84
LSTM & CNN	0,86

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

В следующем эксперименте будет исследоваться точность классификации постов социальной сети с помощью разработанного подхода и с помощью классического метода.

Для экспериментов была сгенерирована обучающая выборка с наилучшими параметрами: длина поста 90-110 символов, не удалять стоп-слова. На сформированной выборке были обучены 2 модели классификации: нейронная сеть со сверточным и рекуррентным слоями и наивный байесовский классификатор. Обе модели были разработаны на языке Python. Размер тестовой выборки для обеих моделей – 30%.

Из предыдущего эксперимента видно, что точность классификации постов социальной сети по 7 человеческим эмоциям с помощью разработанного подхода (в качестве нейронной сети будем использовать архитектуру со сверточным и рекуррентным слоями) – 0,86.

Для обучения байесовского классификатора была использована та же выборка, что и для разработанного подхода. Тексты были предобработаны с помощью алгоритма «мешок слов» с использованием биграмм. Далее классификатор был обучен на обучающей выборке, содержащей классификацию текстов по 7 человеческим эмоциям. Точность классификации составила – 0,38.

Данный эксперимент показывает, что разработанный подход превосходит классический метод классификации, и что классический метод не применим для классификации постов социальной сети.

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

Исходная выборка состояла из 100 объектов (204 профиля различных социальных сетей). Результаты классификации оценивались по метрике площади под кривой ошибок. Для оценки было произведено пять итераций алгоритма со случайным сбалансированным распределением элементов по

классам в обучающей и тестовой выборках. На каждой итерации оптимальные параметры алгоритма классификации подбирались методом перебора.

Можно видеть, что модели не справились с определением пар классов «Игривость / Практичность» и «Привязанность / Отдаленность». Это связано с большим дисбалансом по классам для данных характеристик в исходной выборке. Метод опорных векторов смог показать чуть лучшие результаты для пары «Контролирование / Естественность», а для пар «Экстраверсия / Интроверсия» и «Эмоциональность / Эмоциональная сдержанность» метод опорных векторов показал наилучшие результаты. Следует отметить, что в исследованиях [100, 36, 58, 31, 83, 84, 159] поведение моделей на последних двух парах характеристик также показало наилучшие результаты.

В рамках проведения экспериментов оценивалась эффективность и точность алгоритмов классификации текстов с целью определения психолингвистических характеристик автора. Задача – сравнить несколько алгоритмов классификации по точности классификации.

Эксперименты проводились на выборке из участников психологического опроса по методу «Большая пятерка», содержащей загруженные данные профилей этих людей. Общая схема проведения экспериментов представлена на рисунке 4.4.

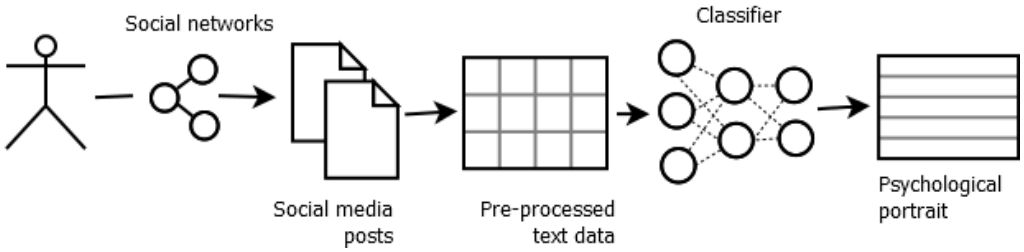


Рисунок 4.4 – Схема проведения экспериментов

Результаты классификации оценивались по метрике AUC ROC. Для оценки было произведено пять итераций алгоритма со случайным сбалансированным распределением элементов по классам в обучающей и тестовой выборках. На каждой итерации оптимальные параметры алгоритма классификации подбирались методом перебора.

Оценка работы алгоритма производится на заранее отобранной тестовой выборке, которая никак не участвуют при обучении классификатора. Также производится экспертная оценка качества. В таблице 4.4 представлены показатели оценки качества системы.

Таблица 4.4 Оценка качества работы классификатора

Класс c_i		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

В таблице приняты следующие условные обозначения: TP – истинно положительное решение; TN – истинно отрицательное решение; FP – ложно положительное решение; FN – ложно отрицательное решение. Согласно определению, точность вычисляется следующим образом:

$$precision = \frac{TP}{TP + FP}.$$

Полнота (TPR) или верная положительная оценка вычисляется по формуле:

$$TPR = \frac{TP}{TP + FN}.$$

Часто можно встретить другую формулу для вычисления точности (accuracy). Эту величину иногда называют правильностью или аккуратностью метода:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

Также необходимо отметить ложная положительная оценка (TPR), вычисляемые по формуле:

$$TPR = \frac{FP}{TN + FP}.$$

В задачах бинарной классификации часто применяют метрику площади под кривой ошибок (AUC ROC) в качестве меры качества классификации [56]. Кривая ошибок показывает зависимость TPR и FPR. Площадь под кривой ошибок является характеристикой качества классификации, не зависящей от соотношения цен ошибок. Чем больше площадь, тем лучше классификация. Численно площадь можно выразить

следующими формулами:

В задачах бинарной классификации часто применяют метрику площади под кривой ошибок (AUC ROC) в качестве меры качества классификации [56]. Кривая ошибок показывает зависимость TPR и FPR. Площадь под кривой ошибок является характеристикой качества классификации, не зависящей от соотношения цен ошибок. Чем больше площадь, тем лучше классификация. Численно площадь можно выразить следующими моделями:

$$\frac{\sum_{i=1}^{q}\sum_{j=1}^{q} I[y_i < y_j] I'[a_i < a_j]}{\sum_{i=1}^{q}\sum_{j=1}^{q} I[y_i < y_j]},$$

$$I'[a_i < a_j] = \begin{cases} 0, a_i < a_j, \\ 0.5, a_i = a_j, \\ 1, a_i > a_j, \end{cases} I'[a_i < a_j] = \begin{cases} 0, y_i \geq y_j \\ 1, y_i < y_j \end{cases},$$

где a_i – ответ алгоритма на i-ом объекте, y_i – его метка (класс), q – число объектов. В идеальном случае площадь под кривой ошибок равна единице, то есть модель всегда верно различает верноположительные и ложноположительные принадлежности классам. Когда площадь под кривой равна 0,5, то это наихудший вариант, так как модель не может различать принадлежность классам. Если площадь менее 0,5, то модель делает обратную классификацию, присваивая объектам противоположный класс.

В качестве входных значений на нейронную сеть подаются предобработанные текстовые массивы, извлеченные из постов пользователя в социальной сети.

Пример выходных значений для фактора эмоциональности указаны в таблице 4.5.

Таблица 4.5 – Выходные значения фактора эмоциональности.

Повышенная эмоциональность:	**Эмоциональная сдержанность:**
Высокие значения по этому фактору свойственны лицам, неуверенным в своих силах, эмоционально незрелым, подверженным колебаниям настроения. Часто такие люди с тревогой ждут неприятностей, а в случае даже мелкой	Низкие значения по этому фактору свойственны лицам самодостаточным, уверенным в своих силах, эмоционально зрелым. Такие люди встречают неприятности без эмоциональных колебаний, практически

неудачи могут впасть в депрессию. У таких людей обычно заниженная самооценка, они обидчивы, но в неудачах обычно обвиняют себя.	всегда следуют планам и могут легко изменить план в случае провала. Такие люди легко приспосабливаются к различным ситуациям, осознают действительность, здраво оценивают внешние факторы.

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 4.5.

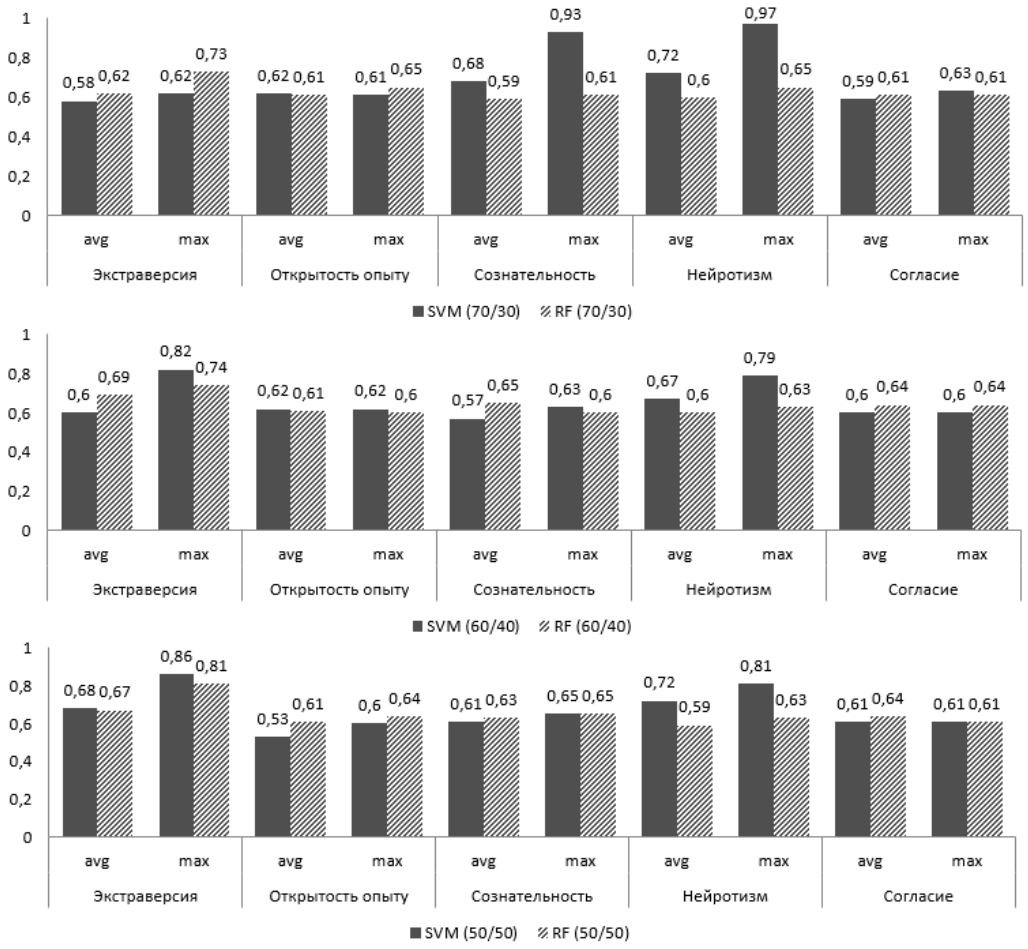


Рисунок 4.5 – Результаты экспериментов

Как видно из результатов экспериментов, наименьшую эффективность предложенный подход показал при определении объектов классов «Agreebleness» и «Openness to experience». Это связано с большим дисбалансом по классам для данных характеристик в исходной выборке. Метод опорных векторов смог показать чуть лучшие результаты для пары «Conscientiousness», а для пар «Neuroticism» метод опорных векторов показал наилучшие результаты. Следует отметить, что в исследованиях [100, 36, 58,

31, 83, 84, 159] поведение моделей на последних двух парах характеристик также показало наилучшие результаты. Таким образом решается задача нахождения психологических особенностей пользователя по особенностям представления его мнения в социальных сетях.

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 экспремента представлены в таблице 4.6.

Таблица 4.6 – Сравнение различных методов классификации

Алгоритм	**Средняя точность метода (%)**
Наивный Байесовский классификатор	57,50
Метод опорных векторов (SVM)	65,12
Линейная регрессия	51,44
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 псхологического показателя. Такой разброс данных связан с недостаточго большой обучающей выборкой, а так же 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».

4.5. Внедрение разработанных алгоритмов и методов

Разработанные подходы были использованы в рамках проекта

«Интеллектуальная платформа формирования социального портрета соискателя на основании семантико-когнитивного анализа профилей в социальных сетях», поддержанного Фонда содействия инновациям по программе «Старт-Цифровые технологии» для ООО «ФаззиЛаб».

Целью проекта была разработка, техническая реализация и проведение тестирования прототипа платформы формирования социального портрета соискателя на основе интеллектуального поиска данных в социальных сетях с применением принципов инженерии знаний.

Эффективность использования научно-технических результатов подтверждена экспериментальными исследованиями, целью которых являлось оценка временных затрат на извлечение, обработку и анализ текстовых данных анализируемых социальных сетей.

В результате проведения экспериментов удалось покрыть тестами функции, реализующие основные алгоритмы системы, что позволило избежать возникновения регрессий в их работе при внесении изменений в программный код.

В результате выполнения нагрузочных тестов были получены следующие показатели:

1. Время отклика прототипа платформы на переход между страницами – не более 0,3 с.
2. Число одновременных запросов к платформе – не менее 100.

По итогам интеграционного тестирования время формирования социального портрета в пределах 3 прогонов колеблется незначительно, требования по затрачиваемому времени на формирование социального портрета выполнены.

Кроме того, разработанные модели и алгоритмы были применены УОСОО «Федерация бадминтона» в рамках проекта «Парабадминтон: все силы – для победы», поддержанного Фондом Президентских грантов для отбора волонтеров, обеспечивающих сопровождение лиц с ПОДА (проект № 18-2-009220).

Целью проекта «Парабадминтон: все силы – для победы» было

создание условий для физической адаптации, социальной интеграции и раскрытия собственного потенциала людей с поражением опорно-двигательного аппарата посредством организации тренировочного процесса и участия спортсменов во Всероссийских соревнованиях по парабадминтону.

В рамках проекта было организовано 4 выезда спортсменов на всероссийские соревнования:

- 1 этап Кубка России по бадминтону-спорт лиц с поражением опорно-двигательного аппарата (г.Новочебоксарск).
- 2 этап Кубка России по бадминтону-спорт лиц с поражением опорно-двигательного аппарата (г.Ставрополь).
- Финальный этап Кубка России по бадминтону-спорт лиц с поражением опорно-двигательного аппарата (г.Казань).
- Открытый Чемпионат России по бадминтону-спорт лиц с поражением опорно-двигательного аппарата (г.Новочебоксарск).

Помимо этого, был организован Открытый чемпионат Ульяновской области по парабадминтону для спортсменов с ПОДА во всех спортивных разрядах.

Эффективность использования научно-технических результатов оценивалась экспериментальными исследованиями, целью которых являлась оценка трудозатрат на поиск и подбор волонтеров для работы со спортсменами с ПОДА по текстовым данным профилей социальных сетей.

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.

По результатам работы разработанного программного комплекса проанализировано 10116 профилей в социальной сети ВКонтакте. Итоги

анализа:

- Удовлетворили условиям поиска – 17 человек.
- После личного собеседования из них было отобрано 9 человек.

По экспертной оценке на обработку одного профиля социальной сети уходит в среднем 30 секунд. Из этого выходит, что на обработку 10116 профилей социальной сети уйдет 82 рабочих часа. Применение разработанного программного комплекса, реализующего алгоритмы интеллектуального анализа текстовых данных социальных сетей, позволило сократить трудозатраты на поиск волонтеров до 68 часов (из них 16 часов – работа эксперта с результатами работы ПО). Разница затраченного времени составило 14 часов.

Кроме того, подходы были использованы в системе интеллектуального поиска и анализа в Интернет-СМИ и социальных сетях в рамках совместного проекта с ФНПЦ АО «НПО «Марс». В рамках данной работы необходимо было осуществить:

1) поиск сообщений в социальной сети "ВКонтакте", имеющей отношение к возникшей ситуативной задаче, формируемой в виде множества ключевых слов;

2) мониторинг необходимых тем в пространстве социальной сети "ВКонтакте" при помощи лингвистических словарей и онтологий;

3) выявление материалов и поисковых запросов в социальной сети "ВКонтакте" на определенную тему;

4) выявлению пользователей в социальной сети "ВКонтакте", распространяющих материалы на определенную тему и анализ их деятельности;

5) мониторингг активности определенных пользователей и групп пользователей, сообществ и т.д. в социальной сети "ВКонтакте";

6) поиску пользователей социальной сети "ВКонтакте" по неполной информации в профиле;

Программная система, разработанная в рамках данного проекта, состояла из четырех подсистем: извлечения данных из общедоступных узлов,

хранения информации, извлеченной из сети Internet, в структурированном виде, подсистемы онтологического поиска информационных ресурсов с учетом особенностей предметной области и подсистемы формирования отчета, содержащего результаты проведенного интеллектуального анализа в разрезах, настраиваемых пользователем.

Для извлечения данных из социальной сети использовался разработанный в рамках данного диисертационного исследования модуль извлечения данных, а структура хранения извлеченных данных в контексте пользователя социальной сети была основана на разработанной в рамках данной диссертационной работы онтологической модели профиля социальной сети.

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

4.6. Выводы по главе

В четвертой главе представлены результаты экспериментов по применению предложенных алгоритмов, реализованных в ввиде программного комплекса.

В рамках исследования были проведены эксперименты по анализу различных методов поиска аккаунтов одного человека в различных социальных. Проведены эксперименты по использованию различных алгоритмов формирования обучающей выборки, используемой в сентимент анализе. По итогам проведенных экспериментов можно сделать вывод о большей эффективности использования одновременно словаря авторских эмоций и ключевых слов. Наибольшую эффективность показала языковая модель BERT.

Помимо этого, представлены результаты экспериментов по использованию различных архитектур нейронных сетей и двух языковых моделей – Word2Vec и BERT. Наилучший результат показал многослойный перцептрон с использванием языковой модели BERT и обучающей выборки,

сформированной при помощи словаря авторских символов и словая ключевых слов.

Преимущество по сравнению с существующими методами достигнуто за счет использования нового алгоритма обучения с учетом особенностей даных социальных сетей и использованием эффективного алгоритма классификации.

Также представлены эксперименты по оценке эффективности алгоритмов классификации текстов с целью определения психолингвистических характеристик пользователя социальной сети. Оцениваемые алгоритмы были основаны на двух моделях классификаторов (SVM и RF), а также включали разделение классифицируемых объектов на обучающую и тестовую выборки в различных соотношениях.

Также в четвертой главе представлены результаты апробации и внедрения разработанных алгоритмов и методов в рамках следующих проектов:

- проект «Интеллектуальная платформа формирования социального портрета соискателя на основании семантико-когнитивного анализа профилей в социальных сетях», поддержанный Фондом содействия инновациям по программе «Старт-Цифровые технологии» для компании ООО «Центр программной инженерии и аналитики «ФаззиЛаб»;
- проект «Парабадминтон: все силы – для победы», поддержанный Фондом Президентских грантов для отбора волонтеров, обеспечивающих сопровождение лиц с ПОДА для УОСОО «Федерация бадминтона»;
- совместный проект с ФНПЦ АО «НПО «Марс» по разработке системы интеллектуального поиска и анализа в Интернет-СМИ и социальных сетях.

Заключение

В диссертационной работе были рассмотрены проблемы анализа слабоструктурированнх данных социальных сетей с целью построения социального портрета пользователя. Были изучены вопросы классификации текстовой информации с целью получения психологических характеристик пользователя, способы идентификации пользователя в различных социальных сетях, а так же изучены вопросы классификации текстовой информации с целью отнесения их к конкретному классу эмоциональной окраски.

В рамках данного исследования была представлена онтологическая модель социального портрета пользователя социальной сети, а также предложен подход к определению психологических характеристик пользователя социальных сетей посредством анализа текстовых сообщений в социальных сетях. Подход заключается в классификации авторских текстов пользователя с использованием машинного обучения. В качестве обучающих и тестовых данных использовались предобработанные текстовые данные со страниц социальных сетей пользователя, а также результаты прохождения пятифакторный опросник личности, определяющие психологические показатели на основании модели «Большой пятерки».

Кроме того, по итогам работы были разработаны:

- алгоритм формирования обучающей выборки, состоящей из постов, классифицированных по 7-ми эмоциям;
- алгоритм классификации текстовых постов социальной сети на основе семантических подходов и машинного обучения;
- алгоритм классификации пользователей социальных сетей по психологическим характеристикам.

Для достижения цели исследования был спроектирован и разработан комплекс психолингвистического анализа профилей, включающий в себя два модуля: подсистема анализа тональности текстов, в качестве обучающих данных которой выступает выборка постов социальной сети, основанная на

словаре авторских символов выражения эмоций и ключевых фраз и подсистема психолингвистического анализа социальных сетей, состоящая из двух компонент. Первая – объединение профилей в различных социальных сетях и вторая – определение психолингвистических характеристик автора.

Для подтверждения эффективности решения были произведены эксперименты. 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 показателях Big5.

Результаты исследований были применены в реальных задачах подбора персонала и внедрены в практику процесса подбора персонала организаций региона.

Дальнейшее развитие исследования может заключаться в расширении базы данных, содержащей посты для анализа, и словарей, используемых при формировании обучающей выборки, определении тональности текстов не в строго определенном интервале длины поста, а текстов любой длины. Может быть рассмотрено применение иных алгоритмов векторизации текста и иных вариантов классификации.

Библиографический список

1. Abdullin Y. B., Ivanov V. V. Deep learning model for bilingual sentiment classification of short texts //Научно-технический вестник информационных технологий, механики и оптики. – 2017. – Vol. 17. – №. 1. – pp. 129-136.

2. Aggarwal S. Modern Web-Development using ReactJS //International Journal of Recent Research Aspects. – 2018. – Vol. 5. – pp. 133-137.

3. Alharbi A. S. M., de Doncker E. Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information //Cognitive Systems Research. – 2019. – Vol. 54. – pp. 50-61.

4. Andrea Esuli , Fabrizio Sebastiani SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining (2006) In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), pp. 417-422

5. Andreev, I. A., Armer, A. I., Krasheninnikova, N. A., Moshkin, V. S Attacking the problem of continuous speech segmentation into basic units //III International conference Information Technology and Nanotechnology. – 2017. – pp. 473-476.

6. Anton Zarubin, Vadim Moshkin, Aleksey Filippov, Albina Koval The approach to the construction of question-answer systems based on the syntagmatic analysis of the text // DS-ITNT 2018// Proceedings of the International conference Information Technology and Nanotechnology. Session Data Science // Samara, Russia, 24-27 April, 2018. pp. 179-185

7. Arevian G., Panchev C. Optimising the hystereses of a two context layer RNN for text classification //2007 International Joint Conference on Neural Networks. – IEEE, 2007. – pp. 2936-2941.

8. Arshad S. Sentiment Analysis / Text Classification Using CNN (Convolutional Neural Network). // Towards Data Science. – 2019.

9. Baccianella S., Esuli A., Sebastiani F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining //Proceedings of the Seventh International Conference on Language Resources and Evaluation

(LREC'10). – 2010.

10. Bartunov S. et al. Joint link-attribute user identity resolution in online social networks //Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM. – 2012.

11. Belov, V., Drozdov, D., Shakurov, R., Moshkin, V., Andreev, I. An integrated approach to mapping user profiles on social networks //CEUR Workshop Proceedings. – 2020. – pp. 225-228.

12. Bhargava R., Arora S., Sharma Y. Neural network-based architecture for sentiment analysis in Indian languages //Journal of Intelligent Systems. – 2019. – Vol. 28. – №. 3. – pp. 361-375.

13. Bobicev V. et al. Emotions in words: Developing a multilingual wordnet-affect //International Conference on Intelligent Text Processing and Computational Linguistics. – Springer, Berlin, Heidelberg, 2010. – pp. 375-384.

14. Bobillo F., Straccia U., Fuzzy ontology representation using OWL 2. International Journal of Approximate Reasoning. Volume 52, 2011, pp. 1073–1094 ().

15. BrandAnalytics [Электронный ресурс] – Режим доступа https://br-analytics.ru – Загл. с экрана (дата обращения: 21.05.2022).

16. Cambria E., Havasi C., Hussain A. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis //Twenty-Fifth international FLAIRS conference. – 2012.

17. Cauwenberghs G., Poggio T. Incremental and decremental support vector machine learning //Advances in neural information processing systems. – 2001. – pp. 409-415.

18. Chen J. et al. Feature selection for text classification with Naïve Bayes //Expert Systems with Applications. – 2009. – Vol. 36. – №. 3. – pp. 5432-5435.

19. Chen, Qufei and Marina Sokolova. “Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical Discharge Summaries.” CoRR abs/1805.00352 (2018).

20. Chetviorkin I.I., Loukachevitch N.V. Sentiment Analysis Track at ROMIP-2012. Компьютерная лингвистика и интеллектуальные технологии. Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2013». Сб. научных статей том 2, pp. 40-50.

21. Cristani, M., Vinciarelli, A., Segalin, C., Perina, A. Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis //Proceedings of the 21st ACM international conference on Multimedia. – 2013. – pp. 213-222.

22. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.

23. Dinu L.P., Iuga I. The Naive Bayes Classifier in Opinion Mining: In Search of the Best Feature Set. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2012. Lecture Notes in Computer Science, vol 7181. Springer, pp 556-567

24. Droganova K. Building a dependency parsing model for Russian with maltparser and Mystem tagset //International Workshop on Treebanks and Linguistic Theories (TLT14). – 2015. – pp. 268.

25. Eureka Engine– NLPub [Электронный ресурс]. – Режим доступа: https://nlpub.mipt.ru/Eureka_Engine (дата обращения: 21.05.2022).

26. Feedot [Электронный ресурс] – Режим доступа http://feedot.com- Загл. с экрана (дата обращения: 21.05.2022).

27. Filippov A., Moshkin V., Yarushkina N. Development of a Software for the Semantic Analysis of Social Media Content. // Recent Research in Control Engineering and Decision Making. ICIT 2019. Studies in Systems, Decision and Control, vol 199. Springer, Cham – 2019 – pp. 421-432

28. ForsMedia [Электронный ресурс] – Режим доступа http://www.fors.ru/business-solutions/forsmedia – Загл. с экрана (дата обращения: 21.05.2022).

29. George A. Miller (1995). WordNet: A Lexical Database for English.Communications of the ACM Vol. 38, No. 11: pp. 39-41.

30. Gjoka, M., Kurant, M., Butts, C. T., Markopoulou, A. Practical

recommendations on crawling online social networks //Selected Areas in Communications, IEEE Journal on. – 2011. – Vol. 29. – №. 9. – pp. 1872-1892.

31. Golbeck J. et al. Predicting personality from twitter //2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. – IEEE, 2011. – pp. 149-156.

32. Hootsuite [Электронный ресурс] – Режим доступа https://hootsuite.com – Загл. с экрана (дата обращения: 21.05.2022).

33. Horev R. BERT Explained: State of the art language model for NLP //Towards Data Science. – 2018. – Vol. 10.

34. Hosmer Jr D. W., Lemeshow S., Sturdivant R. X. Applied logistic regression. – John Wiley Sons, 2013. – Vol. 398.

35. Houston P. Instant jsoup How-to. – Packt Publishing Ltd, 2013.

36. Iacobelli F. Large scale personality classification of bloggers //international conference on affective computing and intelligent interaction. – Springer, Berlin, Heidelberg, 2011. – pp. 568-577.

37. Kain N. Understanding of Multilayer perceptron. // Medium – 2018

38. Ledovaya Y. A., Tikhonov R. V., Bogolyubova O. N. Social networks as a new environment for interdisciplinary studies of human behavior // Vestnik of Saint Petersburg University. Psychology, 7(3), 2017 – pp. 193–210

39. Leskovec J., Faloutsos C. Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – pp. 631-636.

40. Loukachevitch N. V., Lashevich G., Gerasimova A. A., Ivanov V. V., Dobrov B. V. Creating Russian WordNet by Conversion. In Proceedings of Conference on Computatilnal linguistics and Intellectual technologies Dialog-2016, 2016. pp.405-415.

41. Loukachevitch N., Lashevich G. Multiword expressions in Russian thesauri RuThes and RuWordnet //2016 IEEE Artificial Intelligence and Natural Language Conference (AINL). – IEEE, 2016. – pp. 1-6.

42. Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon //Proceedings of the Tenth International Conference on Language

Resources and Evaluation (LREC'16). – 2016. – pp. 1171-1176.

43. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 – 2011 – pp. 142-150.

44. Magnini B., Cavaglia G. Integrating Subject Field Codes into WordNet //LREC. – 2000. – Vol. 1413.

45. Mayers G. Sentiment Analysis from Tweets using Recurrent Neural Networks. // Medium – 2020.

46. McCandless M. et al. Lucene in action. – Greenwich : Manning, 2010. – Vol. 2.

47. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journ. 2014, no. 5, pp. 1093–1113.

48. Mezzalira L. Mobx: Simple state management //Front-End Reactive Architectures. – Apress, Berkeley, CA, 2018. – pp. 129-158.

49. Mikolov T. et al. Distributed Representations of Words and Phrases and their Compositionality. arXiv preprint (2013) //arXiv preprint arXiv:1301.3781. – 2019

50. Mikolov T. et al. Efficient estimation of word representations in vector space. arXiv preprint (2013) //arXiv preprint arXiv:1301.3781. – 2019

51. Moshkin V. Fadeev D., Kurilo D., Andreev I. An Intelligent Search Algorithm for Extremist Texts// Proceedings of ITNT 2021 – 7th IEEE International Conference on Information Technology and Nanotechnology : 7, Samara, 20–24 сентября 2021 года. – Samara, 2021. – DOI 10.1109/ITNT52450.2021.9649291. – EDN JSSMDS.

52. Moshkin V., Andreev I., Yarushkina N. The Extending the Knowledge Base of the Intelligent CAD of a Design Organization Using Semantic Analysis of Wiki-Resources, Advances in Automation // Proceedings of the International Russian Automation Conference, RusAutoConf2020 – 2020

53. Moshkin V., Yarushkina N., Andreev I. Approaches to sentiment analysis of the social network text data //CEUR Workshop Proceedings. – 2020. –

pp. 198-202.

54. Motoyama M., Varghese G. I seek you: searching and matching individuals in social networks //Proceedings of the eleventh international workshop on Web information and data management. – 2009. – pp. 67-75.

55. Najork M., Wiener J. L. Breadth-first crawling yields high-quality pages // Proceedings of the 10th international conference on World Wide Web. – ACM, 2001. – pp. 114-118.

56. Narkhede S. Understanding AUC-ROC Curve // Towards Data Science. 2018. Vol. 26.

57. Narkhede S., Baraskar T. HMR log analyzer: Analyze web application logs over Hadoop MapReduce //International Journal of UbiComp. – 2013. – Vol. 4. – №. 3. – p. 41.

58. Oberlander J., Nowson S. Whose thumb is it anyway? Classifying author personality from weblog text //Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. – 2006. – pp. 627-634.

59. Ouyang X. et al. Sentiment analysis using convolutional neural network //2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing. – IEEE, 2015. – pp. 2359-2364.

60. Ozer D. J., Benet-Martinez V. Personality and the prediction of consequential outcomes //Annu. Rev. Psychol. – 2006. – Vol. 57. – pp. 401-421.

61. Pak A., Paroubek. P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. //LREC (2010).

62. Panicheva P., Bogolyubova O., Ledovaya Y. Revealing interpetable content correlates of the dark triad personality traits //RUSSIR-2016. – Springer Nature, 2016.

63. Pedregosa F. et al. Scikit-learn: Machine learning in Python //the Journal of machine Learning research. – 2011. – Vol. 12. – pp. 2825-2830.

64. Peter Turney Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // Proceedings of the Association for Computational Linguistics. – 2002. – pp. 417–424.

65. Phi M. Illustrated Guide to LSTM’s and GRU’s: A step by step explanation // Towards Data Science. – 2018

66. Piedmont, R. L., Sherman, M. F., Sherman, N. C., Dy-Liacco, G. S., Williams, J. E. Using the five-factor model to identify a new personality disorder domain: the case for experiential permeability //Journal of Personality and Social Psychology. – 2009. – Vol. 96. – №. 6. – pp. 1245.

67. Polyakov I.V., Sokolova T.V., Chepovsky A.A., Chepovsky A.M. Text classification problem and features set. Vestn. NGU. Ser.: Informatsionnye tekhnologii [Novosibirsk State Univ. Journ. of Information Technologies]. 2015, vol. 13, iss. 2, pp. 55–63 (in Russ.).

68. Raad E., Chbeir R., Dipanda A. User profile matching in social networks //2010 13th International Conference on Network-Based Information Systems. – IEEE, 2010. – pp. 297-304.

69. Ramos J. Using tf-idf to determine word relevance in document queries //Proceedings of the first instructional conference on machine learning. 2003.vol. 242. pp. 133-142.

70. RCO Fact Extractor SDK [Электронный ресурс]: RCO. – Режим доступа: http://www.rco.ru/product.asp?ob_no=5047 (дата обращения: 21.05.2022).

71. Rong X. word2vec parameter learning explained //arXiv preprint arXiv:1411.2738. – 2014.

72. Sabuj M. S., Afrin Z., Hasan K. M. Opinion mining using support vector machine with web based diverse data //International Conference on Pattern Recognition and Machine Intelligence. – Springer, Cham, 2017. – pp. 673-678.

73. Safavian S. R., Landgrebe D. A survey of decision tree classifier methodology //IEEE transactions on systems, man, and cybernetics. – 1991. – Vol. 21. – №. 3. – pp. 660-674.

74. Saha S. A Comprehensive Guide to Convolutional Neural Networks – the ELI5 way // Towards Data Science. – 2018

75. Saif H. et al. Contextual semantics for sentiment analysis of Twitter //Information Processing Management. – 2016. – Vol. 52. – №. 1. – pp. 5-19.

76. Segalin C., Cheng D. S., Cristani M. Social profiling through image understanding: Personality inference using convolutional neural networks //Computer Vision and Image Understanding. – 2017. – Vol. 156. – pp. 34-50.

77. Segalin, C., Perina, A., Cristani, M., Vinciarelli, A. The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits //IEEE Transactions on Affective Computing. – 2016. – Vol. 8. – №. 2. – pp. 268-285.

78. SenticNet – concept-level sentiment analysis [Электронный ресурс]. – Режим доступа: https://sentic.net (дата обращения: 21.05.2022).

79. SentiStrength [Электронный ресурс]: SentiStrength – sentiment strength detection in short texts. – Режим доступа: http://sentistrength.wlv.ac.uk/#About (дата обращения: 21.05.2022).

80. SentiWordNet – lexical resource for opinion mining. [Электронный ресурс]. – Режим доступа: https://github.com/aesuli/sentiwordnet (дата обращения: 21.05.2022).

81. simpletransformers [Электронный ресурс]. – Режим доступа: https://github.com/ThilinaRajapakse/simpletransformers (дата обращения: 21.05.2022).

82. Soucy P., Mineau G. W. A simple KNN algorithm for text categorization //Proceedings 2001 IEEE International Conference on Data Mining. – IEEE, 2001. – pp. 647-648.

83. Souri A., Hosseinpour S., Rahmani A. M. Personality classification based on profiles of social networks' users and the five-factor model of personality //Human-centric Computing and Information Sciences. – 2018. – Vol. 8. – №. 1. – pp. 24.

84. Steele Jr F., Evans D., Green R. Is your profile picture worth 1000 words? Photo characteristics associated with personality impression agreement //Third International AAAI Conference on Weblogs and Social Media. – 2009.

85. Tarasov D.S. Deep Recurrent Neural Networks for Multiple Language Aspect-Based Sentiment Analysis. Computational Linguistics and Intellectual Technologies: Proc. of Annual Int. Conf. "Dialogue-2015". Moscow, Russia, 2015,

vol. 2, iss. 14 (21), pp. 65–74.

86. Ting K.M. Precision and Recall. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA, 2011,

87. Transformer – новая архитектура нейросетей для работы с последовательностями [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/341240/(дата обращения: 21.05.2022).

88. Uml O. M. G. 2.0 Superstructure Specification //OMG, Needham. – 2004.

89. Understanding LSTM Networks [Электронный ресурс]. – Режим доступа: http://colah.github.io/posts/2015-08-Understanding-LSTMs (дата обращения: 21.05.2022).

90. Vosecky J., Hong D., Shen V. Y. User identification across multiple social networks //2009 first international conference on networked digital technologies. – IEEE, 2009. – pp. 360-365.

91. Wallach H. M. Topic modeling: beyond bag-of-words //Proceedings of the 23rd international conference on Machine learning. – 2006. – pp. 977-984.

92. Wang S., Huang M., Deng Z. Densely Connected CNN with Multi-scale Feature Attention for Text Classification //IJCAI. – 2018. – pp. 4468-4474.

93. Wangperawong A. Attending to mathematical language with transformers //arXiv preprint arXiv:1812.02825. – 2018.

94. Webb P. et al. Spring boot reference guide //Part IV. Spring Boot features. – 2013. – Vol. 24.

95. What is a Transformer? [Электронный ресурс]. – Режим доступа: https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04 (дата обращения: 21.05.2022).

96. Widiger T. A., Costa P. T., McCrae Jr., R. R. / In P. T. Costa, Jr., T. A. Widiger (Eds.) A proposal for Axis II: Diagnosing personality disorders using the five-factor model. //Personality disorders and the five-factor model of personality – Washington – 2002 – pp. 431-456.

97. Widiger T. A., Mullins-Sweatt S. N. Clinical utility of a dimensional model of personality disorder //Professional Psychology: Research and Practice. –

2010. – Vol. 41. – №. 6. – pp. 488-494.

98. Wiggins J. S., Pincus A. L. Conceptions of personality disorders and dimensions of personality //Psychological assessment: A journal of consulting and clinical psychology. – 1989. – Vol. 1. – №. 4. – pp. 305.

99. WordNet-Affect, FBK-irst © 2009. All Rights Reserved. [Электронный ресурс]. – Режим доступа: http://wndomains.fbk.eu/wnaffect.html (дата обращения: 21.05.2022).

100. Yarkoni T. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers //Journal of research in personality. – 2010. – Vol. 44. – №. 3. – pp. 363-373.

101. 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.

102. Yarushkina, N., Filippov, A., Moshkin, V., Guskov, G., Romanov, A. Intelligent instrumentation for opinion mining in social media //Proceedings of the II International Scientific and Practical Conference Fuzzy Technologies in the Industry, Ulyanovsk, Russia. – 2018. – pp. 50-55.

103. Yarushkina, N., Filippov, A., Moshkin, V., Namestnikov, A., Guskov, G., The social portrait building of a social network user based on semi-structured data analysis // CEUR Workshop Proceedings/ 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction, IS 2019 / Volume 2475, 2019, pp. 119-129.

104. Yekhlakov Y. P., Gribkov E. I. User opinion extraction model concerning consumer properties of products based on a recurrent neural network //Бизнес-информатика. – 2018. – №. 4 (46) eng.

105. You G., Hwang S., Nie Z. Socialsearch: enhancing entity search with social network matching //Proceedings of the 14th International Conference on Extending Database Technology. – 2011. – pp. 515-519.

106. YouScan [Электронный ресурс] – Режим доступа https://youscan.io – Загл. с экрана (дата обращения: 21.05.2022).

107. Zarubin A., Koval A., Filippov A., Moshkin V. Application of Syntagmatic Patterns to Evaluate Answers to Open-Ended Questions // Creativity in Intelligent Technologies and Data Science //Springer, pp.150-162

108. Zhang L., Wang S., Liu B. Deep learning for sentiment analysis: A survey //Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2018. – Vol. 8. – №. 4. – p. 1253.

109. Zou, H., Tang, X., Xie, B., Liu, B. Sentiment classification using machine learning techniques with syntax features //2015 International Conference on Computational Science and Computational Intelligence (CSCI). – IEEE, 2015. – pp. 175-179.

110. Алгоритм Word2Vec [Электронный ресурс]. – Режим доступа: https://neurohive.io/ru/ (дата обращения: 21.05.2022)

111. Алексеев А. А., Лазарева И. М. Морфологический анализ учебных текстов //Актуальные направления научных исследований XXI века: теория и практика. – 2015. – Т. 3. – №. 7-3. – С. 289-292.

112. Анализ точности однофакторного уравнения регрессии [Электронный ресурс]. – Режим доступа: https://studme.org/140829/matematika_himiya_fizik/analiz_tochnosti_odnofaktornogo_uravneniya_regressii (дата обращения: 21.05.2022).

113. Андреев, И. А., Армер, А. И., Крашенинникова, Н. А., Мошкин, В. С. Подход к решению задачи членения слитной речи на речевые единицы //Информационные технологии и нанотехнологии (ИТНТ-2017). – 2017. – С. 473-476.

114. Андреев, И. А., Башаев, В. А., Клейн, В. В., Мошкин, В. С. Определение вероятности терминологичности словоупотреблений в текстах конкретной предметной области //Интегрированные модели и мягкие вычисления в искусственном интеллекте. – 2015. – С. 764-773.

115. Андреев, И. А., Башаев, В. А., Клейн, В. В., Мошкин, В. С., Ярушкина, Н. Г.. Семантическая метрика терминологичности на основе онтологии предметной области //Автоматизация процессов управления. – 2014. – №. 4. – С. 76-84.

116. Андреев, И. А., Бексаева, Е. А., Клейн, В. В., Мошкин, В. С., Серков, И. П. Лингвистический подход к автоматизированному построению предметной онтологии //Прикладные информационные системы. – 2016. – С. 256-263.

117. Антонова А., Соловьев А. Использование метода условных случайных полей для обработки текстов на русском языке. Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2013». Сб. научных статей / Вып. 12 (19).- М.: Изд-во РГГУ, 2013.– С.27-44.

118. Базенков Н. И., Губанов Д. А. Обзор информационных систем анализа социальных сетей / Управление большими системами. Выпуск 41. М.: ИПУ РАН, 2013. С.357-394.

119. Базовые эмоции [Электронный ресурс]. – Режим доступа: https://www.psychologos.ru/articles/view/bazovye-emocii (дата обращения: 21.05.2022).

120. Белов, В. А., Дроздов, Д. С., Шакуров, Р. А., Мошкин, В. С., Андреев, И. А. Комплексный подход к маппингу профилей пользователей в социальных сетях //Информационные технологии и нанотехнологии (ИТНТ-2020). – 2020. – С. 220-225.

121. Богданов А. Л., Дуля И. С. Сентимент-анализ коротких русскоязычных текстов в социальных медиа //Вестник Томского государственного университета. Экономика. – 2019. – №. 47. – С. 220-241.

122. Бодрунова С. С. Кросс-культурный тональный анализ пользовательских текстов в Твиттере //Вестник Московского университета. Серия 10. Журналистика. – 2018. – №. 6. – С. 191-212.

123. Власов, Д. А., Шишков, В. В., Алымов, А. С., Ишин, И. А., Колесников, И. Е., Петров, А. И Описание информационного образа пользователя социальной сети с учетом его психологической характеристики //International Journal of Open Information Technologies. – 2018. – Т. 6. – №. 4. – С. 24-37.

124. Вохминцев, А. В., Соченков, И. В., Кузнецов, В. В., Тихоньких, Д. В. Распознавание лиц на основе алгоритма сопоставления изображений с

рекурсивным вычислением гистограмм направленных градиентов //Доклады Академии наук. – Федеральное государственное бюджетное учреждение" Российская академия наук", 2016. – Т. 466. – №. 3. – С. 261-261.

125. Гречачин В. А. К вопросу о токенизации текста //Международный научно-исследовательский журнал. – 2016. – №. 6 (48) Часть 4. – С. 25-27.

126. Гришеленок Д. А., Ковель А. А. Использование результатов математического планирования эксперимента при формировании обучающей выборки нейросети //Известия высших учебных заведений. Приборостроение. – 2011. – Т. 54. – №. 4. – С. 51-54.

127. Гудков В. Ю., Гудкова Е. Ф. N-граммы в лингвистике //Вестник Челябинского государственного университета. – 2011. – №. 24.

128. Демина Р. Ю., Ажмухамедов И. М. Методика формирования обучающего множества при использовании статических антивирусных методов эвристического анализа //Инженерный вестник Дона. – 2015. – Т. 37. – №. 3. – С. 74.

129. Дли М. И., Булыгина О. В. Особенности применения нейро-сетевых моделей для классификации коротких текстовых сообщений //Программные продукты и системы. – 2019. – Т. 32. – №. 4. – С. 650-654.

130. Ермаков А.Е., Киселев С.Л. Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2015. – Москва, Наука, 2015

131. Ермолаева, П. О., Нагимова, А. М., Носкова, Е. П., Зайнуллина, М. Р., Купцова, А. И. Социальный портрет населения: методология, основные характеристики. // Монография / сост.:– Казань: Казанский (Приволжский) федеральный университет, Артифакт, 2014. – 92 с.

132. Ионова С. В. Эмотивность текста как лингвистическая проблема //Автореф. дисс.. канд. филол. наук. – 1998.

133. Кафтанников И. Л., Парасич А. В. Проблемы формирования обучающей выборки в задачах машинного обучения //Вестник Южно-Уральского государственного университета. Серия: Компьютерные

технологии, управление, радиоэлектроника. – 2016. – Т. 16. – №. 3. – С. 15-24.

134. Клековкина М. В., Котельников Е. В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики //Труды. – 2012. – С. 118-123.

135. Колмогорова А. В., Вдовина Л. А. Лексико-грамматические маркеры эмоций в качестве параметров для сентимент-анализа русскоязычных интернет-текстов //Вестник Пермского университета. Российская и зарубежная филология. – 2019. – Т. 11. – №. 3. – С. 38-46.

136. Корепанова А. А., Абрамов М. В., Тулупьева Т. В. Идентификация аккаунтов пользователей в социальных сетях «вконтакте» и «одноклассники» // Семнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2019, сборник научных трудов – 2019. – С. 153.

137. Коршунов А., Белобородов И., Бузун Н., Аванесов В., Пастухов Р., Чихрадзе К., Козлов И., Гомзин А., Андрианов И., Сысоев А., Ипатов С., Филоненко И., Чуприна К., Турдаков Д., Кузнецов С. Анализ социальных сетей: методы и приложения //Труды Института системного программирования РАН. – 2014. – Т. 26. – №. 1.

138. Котельников Е. В., Клековкина М. В. Определение весов оценочных слов на основе генетического алгоритма в задаче анализа тональности текстов //Программные продукты и системы. – 2013. – №. 4. – С. 296-300.

139. Леоненков А. В. Самоучитель UML 2. – БХВ-Петербург, 2007.

140. Леонтьева Н.Н. О статусе знаний в системах автоматического понимания текста // В сборнике: Компьютерная лингвистика и вычислительные онтологии / Труды XVIII объединенной конференции «Интернет и современное общество» (IMS-2015). 2015. С. 104-115.

141. Лингвистическая онтология "Тезаурус РуТез" [Электронный ресурс]. – Режим доступа: https://www.labinform.ru/pub/ruthes/ (дата обращения: 21.05.2022).

142. Лукашевич Н.В., Левчик А.В. Создание лексикона оценочных

слов русского языка РуСентилекс // Труды конференции OSTIS-2016, С.377-382.

143. Меньшиков И. Л. Анализ тональности текста на русском языке при помощи графовых моделей //УРФУ, Екатеринбург, Россия: конференция.–2012. – 2013.

144. Меньшиков И. Л., Кудрявцев А. Г. Обзор систем анализа тональности текста на русском языке // Молодой ученый. – 2015. – №12. – С. 140-143.

145. Метод полного факторного эксперимента [Электронный ресурс]. – Режим доступа: https://studfile.net/preview/1938844/page:2/ (дата обращения: 21.05.2022).

146. Методология планирования эксперимента: методические указания к лабораторным работам / сост. Т. П. Абомелик. – Ульяновск : УлГТУ, 2011 – 38 с.

147. Мошкин В. С., Андреев И. А. Сравнение эффективности применения алгоритмов сентимент-анализа неструктурированных ресурсов социальных сетей //Системный анализ и информационные технологии САИТ-2019. – 2019. – С. 534-540.

148. Мошкин, В. С., Башаев, В. А., Клейн, В. В., Андреев, И. А. Использование семантической метрики для решения задачи извлечения терминологии из текста проблемной области // Информатика и вычислительная техника. – 2014. – С. 72-78.

149. Найханова Л.В. Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования / Л.В.Найханова. – Улан-Удэ: Издательство БНЦ СО РАН, 2008. – 244 с

150. Определение точности модели [Электронный ресурс]. – Режим доступа: https://studfile.net/preview/5369109/page:8/ (дата обращения: 25.02.2021).

151. Пазельская А., Соловьев А. Метод определения эмоций в текстах на русском языке. Компьютерная лингвистика и интеллектуальные

технологии. Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011». Сб. научных статей / Вып. 11 (18).- М.: Изд-во РГГУ, 2011.– С. 510-523.

152. Посевкин Р. В., Бессмертный И. А. Применение сентимент-анализа текстов для оценки общественного мнения //Научно-технический вестник информационных технологий, механики и оптики. – 2015. – Т. 15. – №. 1. – С. 169-171.

153. Проект ВААЛ [Электронный ресурс]: ВААЛ. – Режим доступа: http://www.vaal.ru/ (дата обращения: 21.05.2022).

154. Раковская Е. Е. Векторная модель представления текстовой информации //Научные исследования: от теории к практике. – 2016. – №. 2-1. – С. 270-272.

155. РуТез – NLPub [Электронный ресурс]. – Режим доступа: https://nlpub.mipt.ru/РуТез (дата обращения: 21.05.2022).

156. Система извлечения знаний из текстов «Аналитический курьер» [Электронный ресурс]: АйТеко. – Режим доступа: http://www.i-teco.ru/ac.html (дата обращения: 21.05.2022).

157. Словарь оценочных слов и выражений русского языка РуСентиЛекс [Электронный ресурс]. – Режим доступа: https://www.labinform.ru/pub/rusentilex/index.htm (дата обращения: 21.05.2022).

158. Смирнова О. С., Шишков В. В. Выбор топологии нейронных сетей и их применение для классификации коротких текстов //International Journal of Open Information Technologies. – 2016. – Т. 4. – №. 8. – С. 50-54.

159. Станкевич, М. А., Игнатьев, Н. А., Смирнов, И. В., Кисельникова, Н. В. Выявление личностных черт у пользователей социальной сети ВКонтакте //Вопросы кибербезопасности. – 2019. – №. 4.

160. Субботин С. А. Быстрый метод выделения обучающих выборок для построения нейросетевых моделей принятия решений по прецедентам // Радиоэлектроника, информатика, управления. – 2015. – №. 1 (32).

161. Тарасова А. Н. Синергия вопросительного и восклицательного

знаков в сетевых текстах (на материале татарского, русского и английского языков) //Вестник Вятского государственного университета. – 2015. – №. 4. – С. 78-84..

162. Татарникова Т. М., Богданов П. Ю. Построение психологического портрета человека с применением технологий обработки естественного языка //Научно-технический вестник информационных технологий, механики и оптики. – 2021. – Т. 21. – №. 1.

163. Тезаурус русского языка RuWordNet [Электронный ресурс]. – Режим доступа: http://www.ruwordnet.ru/ru (дата обращения: 21.05.2022).

164. Трансформеры как графовые нейронные сети [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/491576/ (дата обращения: 21.05.2022).

165. Усталов Д. А. Извлечение терминов из русскоязычных текстов при помощи графовых моделей //Теория графов и приложения= Graphs theory and applications: материалы конференции. – 2012. – С. 62-69.

166. Франсуа Ш. Глубокое обучение на Python. – " Издательский дом"" Питер""", 2018.

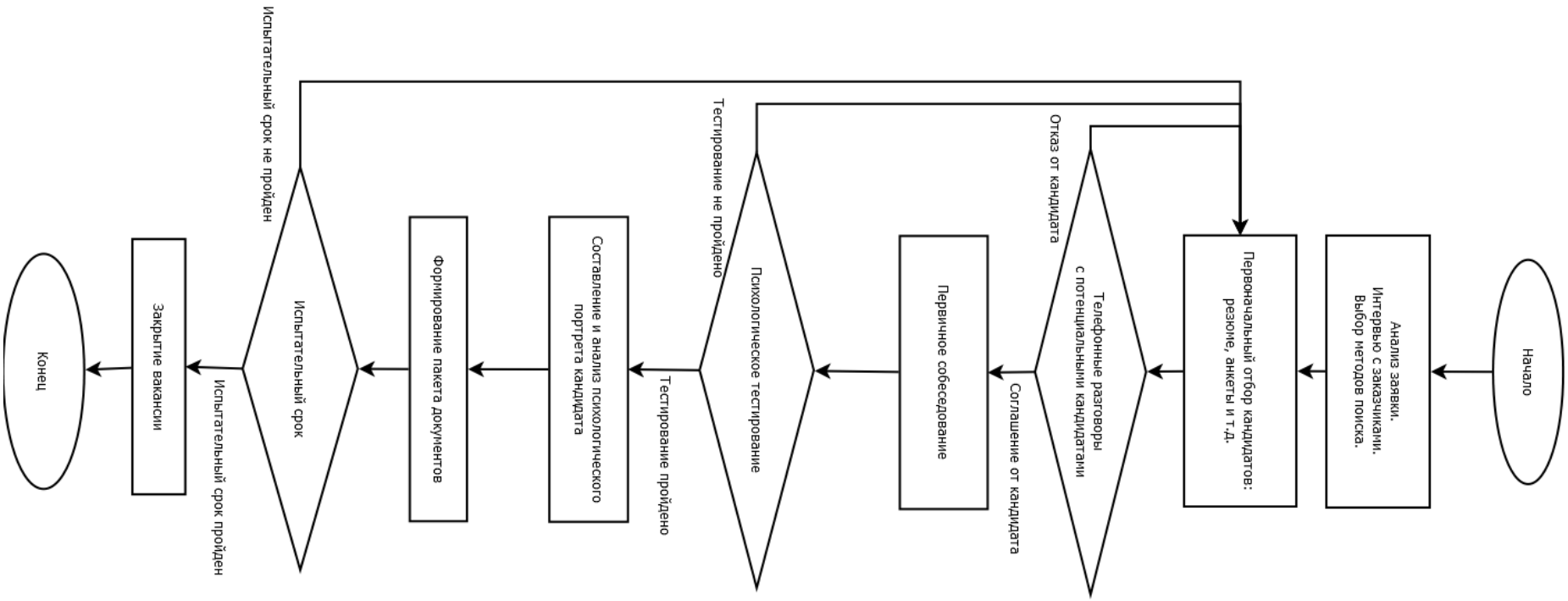
167. Хромов А. Б. Пятифакторный опросник личности: Учебно-методическое пособие //Курган: Изд-во Курганского гос. университета. – 2000. – С. 23.

168. Шипилов О. Ю., Беляев А. С. Анализ эмоционального окраса сообщений в социальной сети твиттер //Вопросы науки. – 2016. – Т. 3. – С. 91-98.

169. Юрганов А. А. Сентимент-анализ как инструмент исследования текстов //Проблемы современной науки и образования. – 2017. – №. 29 (111). – С. 39-41.

170. Ярушкина Н.Г., Андреев И.А. Гуськов Г.Ю., Дударин П.В., Желепов А.С., Мошкин В.С., Наместников А.М., Романов А.А., Филиппов А.А., Эгов Е.Н Интеллектуальный предиктивный мультимодальный анализ слабоструктурированных больших данных // Монография / сост.:– Ульяновск: УлГТУ, 2020. – 220 с.

Приложение А. Блок-схема процесса подбора персонала.



Приложение Б. Таблица сравнения характеристик систем анализа социальных сетей

Характеристика	Google rends	Social Studio	ПРИЗМА	Разрабатываемая система
Пользователи	Интернет-пользователи	Коммерческие организации	Коммерческие организации, государственные структуры	Коммерческие организации, государственные структуры
Уровень анализа данных	Мониторинг с элементами первичного анализа	Мониторинг и анализ	Мониторинг, анализ, прогнозирование, управление	Анализ
Методы анализа	Методы анализа текстов	Базовые методы анализа и поиска текстов на уровне ключевых слов, анализ тональности текстов, визуализация	Методы анализа текста, поиска, визуализация	Методы анализа текста, изображений, визуализация
Объекты анализа социальных сетей	Сеть в целом, информационные сообщения	Сеть в целом, сообщения, мнения, оценки, пользовательская аналитика для ранжирования тематик	Сеть в целом, упоминания, оценки, тнформационная повестка	Профили конкретных пользователей
Режим анализа	Ретроспективный анализ	Анализ в режиме реального времени, ретроспективный анализ с	Ретроспективный анализ	Анализ в режиме реального времени

		ограничением в 30 дней		
Объемы обрабатываемых данных	BigData	Отсутствует информация	Большие	Небольшие
Охват источников данных	Поисковые запросы в Google Поиске, Картинках, Новостях, Покупках и YouTube	Различные медиаресурсы, блоги, сайты, СМИ, социальные сети (Facebook, Twitter, LinkedIn, YouTube, Flickr, Metacafe)	Мониторинг более 900 млн источников	Социальные сети Vkontakte, Facebook, Одноклассники

Приложение В. Акты внедрения

УТВЕРЖДАЮ

Генеральный директор,

председатель НТС ФНПЦ

АО «НПО «Марс», к.т.н.

_____ В.А.Маклаев

"08" 06 2022 г.

А К Т

об использовании результатов кандидатской диссертации И.А. Андреева «Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя»

Научно-техническая комиссия в составе:

председателя комиссии: главный специалист, к.т.н. Э.Д. Павлыгин,

членов комиссии: главный научный сотрудник, д.т.н. Г.П. Токмаков,

начальник отдела развития и поддержания интегрированной автоматизированной системы управления предприятием, к.т.н. А.А. Перцев,

заместитель начальника отдела развития и поддержания интегрированной автоматизированной системы управления предприятием, к.т.н. А.Н. Подобрий.

Настоящим актом подтверждается использование в проектных работах предприятия, следующих научных и практических результатов диссертационной работы И.А. Андреева «Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах

формирования социального портрета пользователя» в рамках проекта «Система интеллектуального поиска и анализа в Интернет-СМИ и социальных сетях»:

- подход к извлечению структурированных и неструктурированных данных из различных социальных сетей;
- унифицированная онтологическая модель профиля социальной сети (в рамках структуры хранения извлеченных данных);
- подход к объединению профилей пользователя в различных социальных сетях.

Данные результаты диссертационной работы использованы на этапе выбора и сравнения методологии автоматизации анализа текстовых данных, извлеченных из социальных сетей и электронных СМИ.

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». 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 в 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%.

Председатель комиссии:
главный специалист, к.т.н. Э.Д. Павлыгин

Члены комиссии:
главный научный сотрудник, д.т.н. Г.П. Токмаков

начальник отдела развития и поддержания интегрированной автоматизированной системы управления предприятием, к.т.н. А.А. Перцев

заместитель начальника отдела развития и поддержания интегрированной автоматизированной системы управления предприятием, к.т.н. А.Н. Подобрий

Ульяновская областная
спортивная общественная организация
«Федерация бадминтона»

ОГРН 115730000549
ИНН 7325138396
КПП 732501001
432071, г. Ульяновск,
пер. Робеспьера, д. 114

А К Т

об использовании результатов кандидатской диссертации И.А. Андреева
“Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя”

Настоящим актом подтверждается использование следующих научных и практических результатов диссертационной работы И.А. Андреева “Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя” для снижения трудозатрат на подбор волонтеров по данным социальных сетей:

- Подход к сопоставлению профилей пользователей в разных социальных сетях посредством гибридизации различных подходов анализа структурированных и неструктурированных данных страниц социальных сетей.
- Метод определения психологических характеристик пользователя социальных сетей с применением методов машинного обучения и метода «Большой пятерки».
- Алгоритм анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей на основании интеграции семантических подходов и методов машинного обучения.
- Программная система психолингвистического и сентимент-анализа открытых текстовых русскоязычных данных профилей пользователей социальных сетей.

Эффективность использования научно-технических результатов подтверждена экспериментальными исследованиями, целью которых являлась оценка трудозатрат на подбор волонтеров для работы со спортсменами с повреждениями опорно-двигательного аппарата (ПОДА) в рамках проекта «Парабадминтон: все силы - для победы», поддержанного Фондом Президентских грантов (проект № № 18-2-009220).

Целью проекта «Парабадминтон: все силы – для победы» было создание условий для физической адаптации, социальной интеграции и раскрытия собственного потенциала людей с поражением опорно-двигательного аппарата посредством организации тренировочного процесса и участия спортсменов во Всероссийских соревнованиях по парабадминтону.

В рамках проекта было организовано 4 выезда спортсменов на всероссийские соревнования:

- 1 этап Кубка России по бадминтону-спорт лиц с поражением опорно-двигательного аппарата (г.Новочебоксарск).
- 2 этап Кубка России по бадминтону-спорт лиц с поражением опорно-двигательного аппарата (г.Ставрополь).
- Финальный этап Кубка России по бадминтону-спорт лиц с поражением опорно-двигательного аппарата (г.Казань).
- Открытый Чемпионат России по бадминтону-спорт лиц с поражением опорно-двигательного аппарата (г.Новочебоксарск).

Помимо этого, был организован Открытый чемпионат Ульяновской области по парабадминтону для спортсменов с ПОДА во всех спортивных разрядах.

Эффективность использования научно-технических результатов оценивалась экспериментальными исследованиями, целью которых являлась оценка трудозатрат на поиск и подбор волонтеров для работы со спортсменами с ПОДА по текстовым данным профилей социальных сетей.

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.

По результатам работы разработанного программного комплекса проанализировано 10116 профилей в социальной сети ВКонтакте. Итоги анализа:

- Удовлетворили условиям поиска – 17 человек.
- После личного собеседования из них было отобрано 9 человек.

По экспертной оценке применение разработанного программного комплекса, реализующего алгоритмы интеллектуального анализа текстовых данных социальных сетей, позволило сократить трудозатраты на поиск волонтеров на 14 часов.

Зам. председателя УОСОО
«Федерация бадминтона» А. В. Костенко

Общество с ограниченной ответственностью
«Центр программной инженерии и аналитики «ФаззиЛаб»

ул.Северный Венец, д.32, г.Ульяновск, 432027
тел. 8 (953) 98-38-627
ОКПО 42929129
ОГРН 1197325020232
ИНН/КПП 7325168471/732501001

«12» мая 2022 года г. Ульяновск

А К Т

об использовании результатов кандидатской диссертации И.А. Андреева "Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя"

Настоящим актом подтверждается использование следующих научных и практических результатов диссертационной работы И.А. Андреева "Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя" для реализации научно-исследовательских и опытно-конструкторских работ (НИОКР) по теме: "Разработка, техническая реализация и тестирование прототипа платформы формирования социального портрета соискателя на основе интеллектуального поиска данных в социальных сетях с применением принципов инженерии знаний" (Проект № 56043, заявка С1ЦТ-66008 в рамках реализации инновационного проекта "Интеллектуальная платформа формирования социального портрета соискателя на основании семантико- когнитивного анализа профилей в социальных сетях"), поддержанного ФГБУ «Фонд содействия развитию малых форм предприятий в научно-технической сфере»:

- Оригинальный подход к сопоставлению профилей пользователей в разных социальных сетях посредством гибридизации различных подходов анализа структурированных и неструктурированных данных страниц социальных сетей.
- Новый метод определения психологических характеристик пользователя социальных сетей с применением методов машинного обучения и метода «Большой пятерки».
- Оригинальный алгоритм анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей на основании интеграции семантических подходов и методов машинного обучения.

– Программная система психолингвистического анализа открытых текстовых русскоязычных данных профилей пользователей социальных сетей.

Целью проекта была разработка, техническая реализация и проведение тестирования прототипа платформы формирования социального портрета соискателя на основе интеллектуального поиска данных в социальных сетях с применением принципов инженерии знаний.

Представленные результаты кандидатской диссертации И.А. Андреева способствовали построению необходимой интеллектуальной программной платформы, а эффективность использования научно-технических результатов подтверждена экспериментальными исследованиями, целью которых являлось оценка временных затрат на извлечение, обработку и анализ текстовых данных анализируемых социальных сетей.

В результате проведения экспериментов удалось покрыть тестами функции, реализующие основные алгоритмы системы, что позволило избежать возникновения регрессий в их работе при внесении изменений в программный код.

В результате выполнения нагрузочных тестов были получены следующие показатели:

1. Время отклика прототипа платформы на переход между страницами – не более 0,3 с.
2. Число одновременных запросов к платформе – не менее 100.

По итогам интеграционного тестирования время формирования социального портрета в пределах 3 прогонов колеблется незначительно, требования по затрачиваемому времени на формирование социального портрета выполнены.

Директор ООО « ФаззиЛаб » **/ Мошкин В.С. /**

Приложение Г. Свидетельства о государственной регистрации программ для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2014612325

Программа для автоматизированного извлечения терминологии из текста на основе лингвистических и статистических методов

Правообладатель: ***федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Ульяновский государственный технический университет» (RU)***

Авторы: ***Ярушкина Надежда Глебовна (RU), Башаев Виталий Александрович (RU), Клейн Виктор Викторович (RU), Андреев Илья Алексеевич (RU)***

Заявка № **2013662129**

Дата поступления **25 декабря 2013 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***25 февраля 2014 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Б.П. Симонов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2014616248

Онтологически-ориентированная система извлечения терминологии

Правообладатель: ***федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Ульяновский государственный технический университет» (RU)***

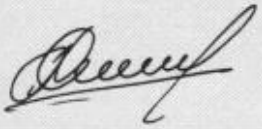
Авторы: ***Ярушкина Надежда Глебовна (RU), Башаев Виталий Александрович (RU), Мошкин Вадим Сергеевич (RU), Клейн Виктор Викторович (RU), Андреев Илья Алексеевич (RU)***

Заявка № **2014613574**

Дата поступления **21 апреля 2014 г.**

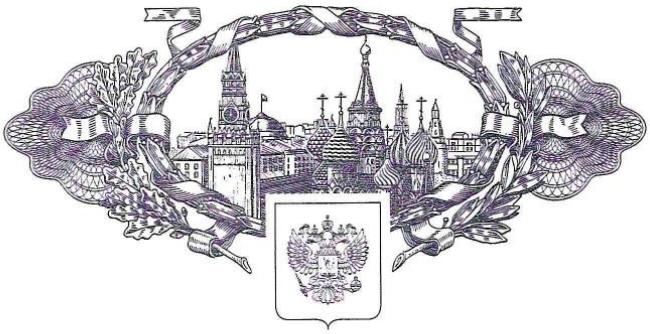
Дата государственной регистрации в Реестре программ для ЭВМ ***18 июня 2014 г.***

Руководитель Федеральной службы по интеллектуальной собственности



Б.П. Симонов

РОССИЙСКАЯ ФЕДЕРАЦИЯ



СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021610716

Интеллектуальная система для Opinion Mining социальных медиа

Правообладатель: ***федеральное государственное бюджетное образовательное учреждение высшего образования «Ульяновский государственный технический университет» (RU)***

Авторы: ***Ярушкина Надежда Глебовна (RU), Мошкин Вадим Сергеевич (RU), Андреев Илья Алексеевич (RU), Константинов Андрей Алексеевич (RU)***

Заявка № **2021610025**
Дата поступления **11 января 2021 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***19 января 2021 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев