Ульяновский государственный технический университет

Стенографический отчет

ЗАСЕДАНИЕ ДИССЕРТАЦИОННОГО СОВЕТА Д212.277.04

Повестка дня:

Защита диссертации **Андреевым Ильей Алексеевичем**

на соискание ученой степени *кандидата технических наук*:

"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).

Официальные оппоненты:

Куприянов Александр Викторович, доктор технических наук, доцент, заведующий кафедрой технической кибернетики, исполнительный директор института информатики и кибернетики, ФГАОУ ВО «Самарский национальный исследовательский университет имени академика С.П. Королева

Абрамов Максим Викторович, кандидат технических наук, руководитель лабораторией теоретических и междисциплинарных проблем информатики, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр РАН» (СПб ФИЦ РАН), г. Санкт-Петербург

Ведущая организация - **ФГБОУ ВО «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича» (СПбГУТ).**

Ульяновск - 2022

ЗАСЕДАНИЕ ДИССЕРТАЦИОННОГО СОВЕТА Д 212.277.04

от 14 сентября 2022 года

на заседании присутствовали члены Совета:

1.	Ярушкина Н.Г., председатель Совета	д.т.н., профессор	05.13.12	технические науки	Очно
2.	Киселев С.К. зам. председателя Совета	д.т.н., доцент	05.13.05	технические науки	Очно
3.	Наместников А.М., ученый секретарь Совета	д.т.н., доцент	05.13.12	технические науки	Очно
4.	Андреев А.С.	д.ф.-м.н., профессор	05.13.01	технические науки	Дистанционно
5.	Афанасьева Т.В.	д.т.н., доцент	05.13.12	технические науки	Дистанционно
6.	Браже Р.А.	д.ф.-м.н., профессор	05.13.05	технические науки	Очно
7.	Васильев К.К.	д.т.н., профессор	05.13.01	технические науки	Очно
8.	Гладких А.А.	д.т.н., профессор	05.13.01	технические науки	Очно
9.	Иванов О.В.	д.ф-м.н., доцент	05.13.05	технические науки	Дистанционно
10.	Клячкин В.Н.	д.т.н., профессор	05.13.01	технические науки	Очно
11.	Крашенинников В.Р.	д.т.н., профессор	05.13.01	технические науки	Очно
12.	Курганов С.А.	д.т.н., доцент	05.13.05	технические науки	Очно
13.	Негода В.Н.	д.т.н., доцент	05.13.12	технические науки	Дистанционно
14.	Перегудова О.А.	д.ф-м.н., доцент	05.13.01	технические науки	Дистанционно
15.	Пиганов М.Н.	д.т.н., профессор	05.13.05	технические науки	Дистанционно
16.	Самохвалов М.К.	д.ф-м.н., профессор	05.13.05	технические науки	Очно
17.	Смирнов В.И.	д.т.н., профессор	05.13.05	технические науки	Очно
18.	Ташлинский А.Г.	д.т.н., профессор	05.13.01	технические науки	Очно

Председатель Совета
д.т.н., профессор Н.Г. Ярушкина

Ученый секретарь Совета
д.т.н., доцент А.М. Наместников

Председатель

Уважаемые коллеги!

На заседании диссертационного Совета Д212.277.04 из **23** члена Совета присутствуют 18 человек, в том числе в удаленном режиме 6 человек. Заседание проходит в смешанном очно-интерактивном режиме. Необходимый кворум имеем.

Членам Совета повестка дня известна. Какие будут суждения по повестке дня? Утвердить? (принято единогласно).

По специальности защищаемой диссертации **05.13.01 «Системный анализ, управление и обработка информации» (информационные технологии и промышленность)** (технические науки) на заседании присутствуют 7 докторов наук.

Наше заседание правомочно.

Председатель

Объявляется защита диссертации на соискание ученой степени *кандидата технических наук* **Андреевым Ильей Алексеевичем** по теме: "*Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя*".

Работа выполнена в Ульяновском государственном техническом университете

Научный руководитель - **к.т.н. Мошкин Вадим Сергеевич**

Официальные оппоненты:

Куприянов Александр Викторович, доктор технических наук, заведующий кафедрой технической кибернетики, исполнительный директор института информатики и кибернетики, ФГАОУ ВО «Самарский национальный исследовательский университет имени академика С.П. Королева

Абрамов Максим Викторович, кандидат технических наук, руководитель лабораторией теоретических и междисциплинарных проблем информатики, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр РАН» (СПб ФИЦ РАН), г. Санкт-Петербург

Присутствуют оба оппонента, оппонент Абрамов М.В. подключен удаленно через средства телеконференцсвязи.

Письменные согласия на оппонирование данной работы от них были своевременно получены.

Ведущая организация - **ФГБОУ ВО «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича» (СПбГУТ).**

Слово предоставляется **Ученому секретарю** диссертационного Совета **Д212.277.04** д.т.н. **Наместникову А.М.** для оглашения документов из личного дела соискателя.

Ученый секретарь

Соискателем **Андреевым Ильей Алексеевичем** представлены в Совет все необходимые документы для защиты кандидатской диссертации (зачитывает):

- заявление соискателя;
- копия диплома о высшем образовании (заверенная);
- справка об обучении в аспирантуре;
- заключение по диссертации от организации, где выполнялась работа;
- отзыв научного руководителя;
- диссертация и автореферат в требуемом количестве экземпляров.

Все документы личного дела оформлены в соответствии с требованиями Положений ВАК.

Основные положения диссертации отражены **Андреевым И.А.** в **36** научных работах, в т.ч. в **4 статьях в изданиях из перечня ВАК, 11 публикациях индексируемых Scopus, 1 коллективная монография, 3 свидетельства на регистрацию программы для ЭВМ.** Соискатель представлен к защите **22.06.2022 г.** (протокол **№8**). Объявление о защите размещено на сайте ВАК РФ **27.06.2022 г.**

Вся необходимая информация по соискателю внесена в ФИС ГНА.

Председатель

Есть ли вопросы по личному делу соискателя к ученому секретарю Совета? (Нет).

Есть ли вопросы к **Андрееву И.А.** по личному делу? (Нет).

Илья Алексеевич, Вам предоставляется слово для изложения основных положений Вашей диссертационной работы.

(Соискатель излагает основные положения работы).

Соискатель

Добрый день, уважаемый председатель, уважаемые члены диссертационного совета, уважаемые присутствующие. Разрешите представить мою диссертационную работу на тему "Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя". Защита проходит по специальности 05.13.01 «Системный анализ, управление и обработка информации» (информационные технологии и промышлен-

ность), научный руководитель - кандидат технических наук Мошкин Вадим Сергеевич.

Социальные сети являются неотъемлемой частью жизни современного человека. Концепция Web 2.0, которая в данный момент превалирует, предполагает формирование электронных ресурсов пользователями, которым предоставляется шаблон. Многие бизнес задачи, которые ранее было невозможно решить из-за недостатка данных, сейчас решаются путем извлечения данных из социальных сетей. На слайде представлен неисчерпывающий список областей применения анализа текста социальных сетей, одной из проблем, которые мы можем решить является подбор персонала для HR.

В сфере подбора персонала изучают психологические особенности человека, его тип характера и темперамент. Эту задачу можно выполнить на основе социальных сетей. При этом текст в социальных сетях абсолютно не структурирован и обладает специфичными особенностями. Методы системного анализа отличаются междисциплинарным подходом к решению сложных проблем, и они применимы для анализа текстовых данных социальных сетей. Таким образом, применение системного подхода для создания новых методов и алгоритмов психолингвистического и сентимент-анализа является актуальной задачей.

Сентимент-анализ — класс методов контент-анализа в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов (мнений) по отношению к объектам, речь о которых идет в тексте. Психолингвистический анализ - это класс методов контент-анализа в компьютерной лингвистике, предполагающий выявление стоящих за вербальной формой текста психологических состояний и отношений, которые заложены в текст автором или группой авторов.

Целью диссертационной работы является снижение трудозатрат на построение социального портрета пользователей социальных сетей посредством автоматизации и учета дополнительных факторов в процессе анализа 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. В качестве примера использованы русскоязычные текстовые данные страниц следующих социальных сетей: ВКонтакте, Одноклассники, Facebook, Instagram. Необходимо упомянуть, что Организация Meta, а также ее продукты Instagram и Facebook, 21 марта 2022г. Тверским судом города Москвы признаны экстремистскими и запрещены на территории РФ, однако на момент проведения исследования данного решения суда еще не было вынесено.

Существует ряд задач, которые были поставлены во время работы над диссертационным исследованием. Некоторые из них - разработка алгоритма формирования обучающей выборки, состоящей из открытых русскоязычных текстовых ресурсов социальных сетей, классифицированных по 7-ми эмоциям; разработка методов классификации текстовых постов социальной сети по классам тональности на основе семантических подходов и машинного обучения; разработка подхода к сопоставлению профилей пользователей в разных социальных сетях посредством анализа структурированных и неструктурированных данных анкет, а также социальных графов профилей; разработка подхода к

определению психологических характеристик пользователя социальных сетей посредством анализа текстовых сообщений в социальных сетях. Кроме того, необходимо разработать программную систему, реализующую описанные алгоритмы и подходы, а также провести эксперименты и внедрение разработанного ПО.

Научная новизна представлена на слайде. Стоит упомянуть, что в диссертационной работе применяются методы онтологического инжиниринга, методы обработки естественного языка, нейросетевые методы, методы машинного обучения, методы теории анализа социальных сетей (SNA), а также объектно-ориентированного программирования при построении программного комплекса. Область исследования соответствует паспорту специальности 05.13.01. - «Системный анализ, управление и обработка информации (технические науки)».

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

Конечно, существуют работы по всем задействованным а рамках диссертационного исследования направлениях. Это сентимент-анализ, поиск и объединение профилей и психолингвистический анализ текста. Иследователей сентимент-анализа можно разделить на 3 группы: использование словарей, классификация с помощью машинного обучения и смешанный подход.

В рамках диссертационного исследования была разработана следующая схема сентимент-анализа: формирование обучающей выборки для обучения нейронной сети с использованием разработанного алгоритма совместного применения расширенного тезауруса WordNet-Affect и словаря авторских символов выражения эмоций, векторизация обучающей выборки посредством применения языковой модели BERT, обучение нейронной сети эффективной архитектуры на полученной обучающей выборке, извлечение, предобработка и векторизация текстовых данных социальных сетей с использованием подходов, примененных при формировании обучающей выборки и классификация предобработанных и векторизированных текстовых данных с использованием обученной нейронной сети на 7 классов тональности.

Обучающая выборка формируется следующим образом: лемматизация текста, удаление стоп-слов, удаление слов на других языка. После чего используются словарь авторских символов выражения эмоций и расширенный словарь WordNetAffect. Пример расширенного словаря WordNetAffect представлен на слайде.

В рамках диссертационной работы был разработан словарь авторских символов выражения эмоций, который содержит 1427 символов, из них к символам выражения авторских эмоций относится 372 знака. Они

были сгруппированы в 62 агрегирующих символа, распределенных на 7 классов.

На следующем слайде представлен алгоритм формирования обучающей выборки и пример выходных данных. В рамках диссертационной работы было использованы предобученные языковые модели Word2Vec и BERT, на экране представлены их математические модели.

Обучение нейронной сети происходит при помощи указанной ранее обучающей выборки и векторизованного текста. В рамках диссертационной работы были использованы следующие нейронные сети: рекуррентные, сверточные, двунаправленные рекуррентные, многослойный персептрон и гибридный подход.

Следующий этап исследования – это унифицированная онтологическая модель профиля социальной сети. Разработанная в рамках диссертационного исследования модель представляется следующим образом: N – множество узлов онтологии; R – множество отношений онтологии; F – множество функций интерпретации онтологии. Если подразделять модель, то у нас есть: узловые объекты, объекты материального мира, отношения между объектами онтологии, отношения между объектами онтологии и значениями встроенного типа и свойств аннотации. Разработанная модель предполагает учет временного контекста и контекста источника данных. Т.е. у нас есть один и тот же тип данных из разных социальных сетей, и мы можем учитывать где человек пребывает в конкретный момент времени с его собственных слов.

После разработки унифицированной модели профиля социальной сети мы можем перейти к сопоставлению профилей различных социальных сетей. Для этого было выделено несколько критериев: критерий схожести анкет, критерий наличия схожих лиц на фотографиях, критерий наличия схожих контактов, критерий наличия схожего места работы и места учебы, критерий наличия схожих постов, включающий 2 метрики: нахождение расстояния Левенштейна и алгоритм шинглов и критерий совпадения социальных графов.

После объединения нескольких профилей различных социальных сетей одного человека мы можем перейти к психологической оценке пользователя. В психологии «Большая пятерка» – это пятифакторная модель личности, разработанная психологами Р. МакКрае и П. Коста таким образом, чтобы из набора входящих в нее черт можно было составить структурированный и довольно полный портрет личности. Классическая «Большая пятерка» выделяет следующие основные факторы: нейротизм; экстраверсия; открытость опыту (openness to experience); согласие и сознательность. Его адаптацией стал тест 5PFQ японского психолога Хедзиро Тсудзи. На русский язык этот опросник переведен А.Б. Хромовым. В данном тесте 75 вопросов, и он определяет выраженность 30 черт личности (5 основных факторов и 25 первичных).

Алгоритм определения психологических характеристик выглядит следующим образом: у нас есть добровольцы, которые предоставили доступ к своим социальным сетям и прошли упомянутый ранее опросник. Их посты в социальных сетях векторизуются и образуют входные данные для обучения классификатора. Результаты опросника – это выходные данные для обучения. После обучения часть с опросником отбрасывается, остается только система поддержки принятия решений. Примеры вопросов из опросника представлены на слайде. По результатам опросника получается следующая информационная модель, в которой ответы на вопросы распределяются по классам.

В рамках диссертационного исследования был создан комплекс программ, состоящий из двух частей: сентимент-анализ и социальный портрет. Взаимодействие двух частей происходит по API. Программа сентимент-анализа, в свою очередь, состоит из двух микросервисов: микросервис формирования обучающей выборки и микросервис определения тональности текста (сентимент-анализ). В ходе работы было обработано более 2,5 миллионов сообщений (коротких текстов), которые для оптимизации были распараллелены на 7 потоков. Обработка происходила этапами по 100 тысяч сообщений за раз.

На слайде представлены скриншоты системы сентимент-анализа текста: интерфейс пользователя и лог взаимодействия с системой по API. На следующем слайде представлены скриншоты системы построения психологических характеристик пользователя. По профилю пользователя в социальной сети Вконтакте мы нашли его аккаунты в социальных сетях Facebook и Одноклассники, после чего модулем дается описание психологического портрета пользователя с указанием каждой черты.

В рамках диссертационной работы был проведен ряд экспериментов, которые можно разделить на 3 части: эксперименты по объединению профилей пользователей в различных социальных сетях, эксперименты по сентимент-анализу текстовых данных и эксперименты по оценке алгоритма психолингвистического анализа текста профилей социальных сетей. Во всех экспериментах использовались результаты анкетирования по методу «Большой пятерки» и публичные открытые данные из аккаунтов социальных сетей, принадлежащих добровольцам из числа сотрудников и студентов УлГТУ.

Первый набор экспериментов был по объединению профилей. В качестве экспериментальной базы было отобрано 100 человек, которые имели 204 аккаунта в 3 социальных сетях. Такое количество обусловлено тем, что не каждый человек имел аккаунт во всех трех социальных сетях. По результатам работы после применения всех критериев мы смогли однозначно установить кому какой аккаунт принадлежит.

Вторым набором экспериментов были эксперименты по сентимент-анализу текстовых данных. На слайде представлен результат работы алгоритма формирования обучающей выборки после фильтрации 2,5 миллионов сообщений (текстовых данных). После формирования обучающей выборки она была разделена на ту, что пойдет на обучение нейронной сети и ту, что будет использована для проверки в соотношении 90 на 10. После чего был проведен набор экспериментов, результаты одного из них представлены на экране - в данном случае это результаты сравнения двух предобученных языковых моделей Word2Vec и BERT. BERT показал лучшую точность при любой архитектуре нейронной сети. Следующий набор экспериментов касался различных параметров системы. Так, например, были проведены эксперименты только с формированием обучающей выборки по словарю авторских символов выражения эмоций и только по словарю WordNetAffect. Лучшую точность показал BERT, при длине сообщения 90-110 символов и использованием весов классов. Различные нейронные сети показали следующие результаты (представлены на слайде). Лучшим себя показал многослойный персептрон.

Последний набор экспериментов был проведен длов метода определения психологического портрета пользователя. По результатам можно увидеть психологический портрет человека. Поскольку по сути построение психологического портрета является задачей бинарной классификации, в работе была применена метрика площади под кривой оши-

бок (AUC ROC) в качестве меры качества классификации. Кривая ошибок показывает зависимость истинно положительных к ложно положительным. Для классификации использовались метод опорных векторов и метод случайного леса. Эксперименты проводились с делением выборки на обучающую и тестовую в соотношениях 50/50, 60/40 и 70/30. Точность классификации варьируется от 0,58 до 0,93 в зависимости от психологической характеристики и размера обучающей выборки.

Результаты диссертационного исследования были внедрены в следующих организациях: 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», УОСОО »Федерация бадминтона» в рамках проекта «Парабадминтон: все силы - для победы», поддержанного Фондом Президентских грантов для отбора волонтеров, обеспечивающих сопровождение лиц с ПОДА, в рамках проекта «Система интеллектуального поиска и анализа в Интернет-СМИ и социальных сетях», реализуемого совместно с Федеральным научно-производственным центром АО «Научно-производственное объединение «Марс» (ФНПЦ АО НПО «Марс»).

Отдельно остановимся на практической значимости для УОСОО »Федерация бадминтона». Задача - поиск волонтеров в социальных сетях для работами с лицами с ПОДА. 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 итогам работы было проанализировано 10115 профилей в социальной сети Вконтакте; удовлетворили условиям поиска – 17 человек. После личного собеседования из них было отобрано 9 человек. Автоматизированная обработка заняла 25 минут. Экспертная оценка экономии времени - 14 человеко-часов на 10115 профилей.

Основные итоги исследования: разработан алгоритм формирования обучающей выборки, состоящей из постов, классифицированных по 7-ми эмоциям, разработан алгоритм классификации текстовых постов социальной сети на основе семантических подходов и машинного обучения, разработан алгоритм классификации пользователей социальных сетей по психологическим характеристикам, разработан программный комплекс, реализующий описанные алгоритмы и методы классификации пользователей. 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 Большой Пятерки. Использование подхода к построению психологического портрета пользователя социальных сетей, включающего разработанные алгоритмы психолингвистического и сентимент-анализа русскоязычных структурированных и неструктурированных ресурсов социальных сетей, позволило сократить трудозатраты на поиск волонтеров, обеспечивающих со-

провождение лиц с ПОДА, в рамках проекта «Парабадминтон: все силы - для победы» для УОСОО «Федерация бадминтона» на 14 часов на 10115 пользователей. Результаты исследований внедрены в практику процесса подбора персонала организаций региона.

В рамках диссертационного исследования опубликованы 32 статьи, 4 из них в журналах списка ВАК, 11 - в изданиях, индексируемых в Scopus и Web of Science, было получено 3 свидетельства о государственной регистрации программ для ЭВМ и результаты вошли в 1 монографию.

На этом мой доклад окончен. Спасибо за внимание.

Председатель

У кого есть вопросы к соискателю?

(Соискатель отвечает на вопросы)

д.т.н., профессор Клячкин В.Н.

У меня есть несколько вопросов. У Вас в работе есть этап - предобработка данных. Поясните, что такое предобработка текстовых данных, о которых идет речь в вашей работе.

Соискатель

Под предобработкой текста подразумевается лемматизация текста для упрощения векторизации, удаление стоп-слов по списку. Например, в этот список входят слова-паразиты, которые не влияют на сам текст. После чего происходит удаление слов на иностранных языках, т.к. разработанный алгоритм предназначен для обработки русскоязычного текста. Кроме того, если мы говорим о работе классификатора, а не о работе модуля формирования обучающей выборки, то векторизация текста также входит в предобработку.

д.т.н., профессор Клячкин В.Н.

Откройте, пожалуйста, выводы. У Вас указана точность 87%. Как она определяется? И какое место здесь имеет AUC ROC?

Соискатель

Обучающая выборка была разделена на 90% и 10%. 10% не участвовали в обучении, но участвовали в экспериментах. Этот набор данных был проверен, и точность определения составила 87%. AUC ROC применяется только для проверки метода определения психологических характеристик.

д.т.н., профессор Клячкин В.Н.

В экспериментах определения психологических характеристик Вы указали, что лучшим оказался метод опорных векторов. А какие-то другие методы применялись?

Соискатель

Использовался метод случайного леса, и в дополнительных экспериментах было проведено дополнительное сравнение, в котором, наряду с методом опорных векторов и методом случайного леса были использованы наивный байесовский классификатор и логистическая регрессия.

д.т.н., профессор Клячкин В.Н.

Тестовая выборка формируется случайным образом, я правильно понимаю? Если доверительный интервал построить за счет случайности и построить доверительный интервал за счет метода разбиения, результаты могут быть соизмеримы. Проводились ли подобные эксперименты?

Соискатель

Нет, подобных экспериментов не проводилось.

д.т.н., доцент Наместников А.М.

Формула онтологии, применяемой в Вашей работе, довольно известная. Какие особенности присутствуют в Вашей онтологической модели, исходя из решаемой задачи, и чем она отличается от других моделей, решающих похожие задачи?

Соискатель

В рамках модели идет учет контекста полученных данных, они хранятся в онтологии. Сама модель построена на основе данных, которые хранятся в социальных сетях, и учитывает их специфику. Введены специфичные для данной онтологии понятия и отношения.

д.т.н., профессор Крашенинников В.Р.

Полученная Вами точность оценки сентимент-анализа составила 87%. Следовательно – 87% правильных. А были ли такие, которые нельзя было однозначно отнести к неправильным?

Соискатель

Такое исследование не проводилось. Результат оценивался только как "правильный" и "неправильный".

д.т.н., профессор Крашенинников В.Р.

13% - это ошибочно определенные тексты. Проводилось ли исследование, почему именно текст был отнесен к ошибочным?

Соискатель

Да, проводилось. Но нужно смотреть каждый конкретный случай, почему именно данный текст был ошибочно определен. Среди этих 13%

частыми являлись случаи, когда определение было ошибочно из-за многозначных словосочетаний. Одним из примеров является словосочетание "ужасно красивый". Соответственно, если улучшить обучающую выборку, то и качество оценки тоже увеличится.

д.т.н., профессор Крашенинников В.Р.

К вопросу об обучающей выборке. Как я понимаю – эти тексты уже кем-то были разбиты на эти группы?

Соискатель

Нет, обучающую выборку мы формируем на основе разработанного в рамках диссертационного исследования алгоритма. Т.е. у нас есть 2,5 миллиона текстов социальной сети, и они формируются в обучающую выборку при помощи двух словарей: WordNetAffect и словаря авторских символов выражения эмоций. Т.е. выборка саморазмеченная пользователями социальных сетей.

д.т.н., профессор Крашенинников В.Р.

Было ли проведено сравнение, в каких текстах ошибки встречаются чаще, в каких реже? Допустим, что ошибки встречаются чаще в длинных текстах или наоборот?

Соискатель

Да, такие эксперименты проводились. По результатам эксперимента было определено, что модель подходит для коротких текстов 90-110 символов. Хуже алгоритм себя показывает, если текст состоит из 40-50 слов. Но тексты социальных сетей отличаются как раз тем, что они коротки.

д.т.н., доцент Киселев С.К.

В работе у Вас отсутствуют ссылки на какие-либо монографии по психолингвистическому анализу. Ссылки есть только на статьи. Неужели нет в психолингвистическом анализе каких-то монументальных трудов, которые можно было бы использовать в вашей работе?

Соискатель

Подобные монографии есть, но в них дается описание психолингвистических методов без автоматизации, поэтому я не счел необходимым включать их в список литературы, однако сослался на существующие методы автоматизации психолингвистического анализа.

д.т.н., доцент Киселев С.К.

В Вашей работе психолингвистические методы изобретены Вами? Оценку Вы производили тоже самостоятельно?

Соискатель

Нет, это общеизвестные методы. Оценка корректности проводилась в соответствие с методикой оценки результатов опросника Хедзиро Тсудзи, переведенного А.Б. Хромовым.

Д.т.н., доцент Киселев С.К.

Замечу, что обычно в списке литературы приведено большое количество различных монографий и учебников, однако в Вашей работе нет ссылок на подобные психолингвистические монографии.

Д.т.н., доцент Наместников А.М.

В продолжение того вопроса, который я задавал ранее. Онтология создается не просто так, не для того, чтобы она была, а для какой-то задачи. Для чего используется онтология в данной работе?

Соискатель

Информация, полученная из профилей социальных сетей, преобразовывается в данную модель онтологии согласно функциям интерпретации в узлы и отношения онтологии, после чего применяется подход к сопоставлению профилей социальных сетей. Т.е., например, у нас есть профиль человека в Twitter и профиль ВКонтакте, и при помощи разработанного подхода и SWRL-правил (используемые, в первую очередь, для учета контекста), он объединяется.

Председатель, д.т.н., профессор Ярушкина Н.Г.

У меня тоже есть вопрос. Вы упомянули три приложения Вашей работы. Было применение для отбора кандидатов в волонтеры, было применение в программном комплексе, который был передан в предприятие, которое реализует инструментарий для кадровых служб. Третье приложение - разработанный алгоритм извлечения, предобработки данных и формирование социального портрета - были использованы в рамках проекта «Система интеллектуального поиска и анализа в Интернет-СМИ и социальных сетях», реализуемого совместно с Федеральным научно-производственным центром АО «Научно-производственное объединение «Марс» (ФНПЦ АО НПО «Марс»). Охарактеризуйте, пожалуйста, подробнее это приложение.

Соискатель

По результатам внедрения были получен акт, содержащие следующие результаты: система использовалась как модуль для программы ситуационного центра «Системы управления регионом» для поиска пользователей с учетом параметров, содержащихся в онтологической модели пользователя: возраст, место проживания, общая эмоциональная окраска комментариев. Поиск был направлен на выявление пользователей, распространяющих информацию на определенную тему, мониторинг необходимых тем, поиск текстовых сообщений, имеющих отношение к возникшей ситуации, а так же поиск пользователей по неполным данным. В результате внедрения методов и алгоритмов, разработанных

в рамках данной диссертационной работы, достигнуто среднее сокращение времени поиска необходимых профилей на 40%.

Председатель, д.т.н., профессор Ярушкина Н.Г.

То есть те модели и подход, что Вы предложили, были использованы при разработке "системы управления регионом" в рамках проекта "ситуационный центр"?

Соискатель

Да, правильно.

Председатель

Есть еще вопросы? (Нет).

Согласны ли члены Совета сделать технический перерыв? (Нет). Тогда продолжаем работу.

Слово предоставляется научному руководителю работы **к.т.н. Мошкину В.С.**

Добрый день, уважаемые коллеги!

Сразу хочется обратить внимание на актуальность работы Ильи Алексеевича. В настоящее время тот огромный объем неструктурированной информации, который содержится в социальных сетях, имеет большую значимость и ценность, поэтому он должен использоваться для решения критически важных задач. Но для решения этих задач необходимо учитывать особенности представления этой информации, в первую очередь неструктурированность и жанровую особенность.

В диссертационном исследовании соискателю удалось с помощью применения и интеграции эффективных методов системного анализа, онтологического инжиниринга, машинного обучения получить хорошие научные результаты в области обработки неструктурированных знаний социальных сетей и извлечения знаний, в том числе и для практических задач, одной из которых является задача поддержки принятия решений при подборе персонала.

Нужно сказать, что помимо научной значимости данной работы, диссертационное исследование имеет также и практическую ценность, которая была оценена на федеральном уровне. Научные изыскания в рамках диссертационной работы были поддержаны "Фондом содействия инноваций" и "Российским фондом фундаментальных исследований".

Илья Алексеевич, помимо научной работы, работает со студентами, преподает на кафедре "Информационные системы" УлГТУ – является старшим преподавателем. Помимо этого он занимается цифровизацией бизнес-процессов УлГТУ, возглавляя лабораторию автоматизации образовательного процесса УлГТУ. Считаю, что Илья Алексеевич является профессионалом высокого уровня и состоявшимся исследователем, и достоин присуждения ему ученой степени кандидата технических наук.

(Отзыв *прилагается*).

Председатель

Ученому секретарю Совета предоставляется слово для оглашения заключения организации, где выполнялась работа и отзыва ведущей организации.

Ученый секретарь оглашает заключение организации, где выполнялась работа. Затем зачитывает отзыв ведущей организации.

(Заключение и отзыв *прилагаются*).

Председатель

На автореферат диссертации поступило 4 отзыва, все они положительные. Согласны ли члены Совета заслушать обзор отзывов или зачитать их полный текст?

Слово для обзора отзывов, поступивших на диссертацию, предоставляется **Ученому секретарю Совета.**

Ученый секретарь зачитывает обзор отзывов.

1. ФГБОУ ВО «Ульяновский институт гражданской авиации имени главного маршала авиации Б.П. Бугаева»

Отзыв подписан доцентом кафедры «Организации аэропортовой деятельности и информационных технологий» Чоракаевым О.Э. (к.т.н., специальность 05.13.12)

Замечания:

•В автореферате можно заметить следующий недостаток: приведены эксперименты по оценке эффективности алгоритмов классификации тональности с использованием только двух языковых моделей – Word2Vec и BERT, при этом не встречается обоснования их применения. В настоящее время существует большой набор модификаций языковых моделей (ELMo, GloVe, TinyBERT), которые могли бы показать высокую эффективность в данной задаче

2. Филиал ФГБОУ ВО «Национальный исследовательский университет «МЭИ» в г. Смоленске.

Отзыв подписан ведущим программистом лаборатории Луферовым В.С. (к.т.н., специальность 05.13.01)

Замечания:

- Одним из пунктов, вынесенных автором в качестве научной новизны, является методика объединения профилей пользователей из различных социальных сетей. В экспериментах показано, что текущего набора алгоритмов сравнения данных профилей, примененных последовательно, достаточно для объединения профилей 100 человек (всего 204 профиля) разных социальных сетей. Однако, данная методика может показать иной уровень эффективности при обработке данных более крупной выборки.
- В списке публикаций встречаются опечатки и ошибки форматирования.

3. ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации»

Отзыв подписан доцентом департамента анализа данных и машинного обучения Факультета информационных технологий и анализа больших данных Андрияновым Н.А. (к.т.н. 05.13.18)

Замечания:

- В описании алгоритма формирования обучающей выборки Шаг №4 включает в себя предобработку данных. Однако предобработка - это достаточно важный шаг, детальное описание которого в автореферате опущено.
- В таблице 1 и далее приводятся метрики точности разных моделей. Однако отсутствует информация о распределении классов в обучающей и тестовой выборках, без которой сложно судить о качестве работы модели, даже если точность составляет 87%.
- В таблице 2 в качестве алгоритма классификации указана «Линейная регрессия». Вероятно, автор имел в виду логистическую регрессию?

4. ФГАОУ ВО «Казанский (Приволжский) федеральный университет»

Отзыв подписан доцентом кафедры информационных систем Невзоровой О.А. (к.т.н., специальность 05.13.17)

Замечания:

Недостатком автореферата является тот факт, что автором не указан объем применяемых им словарей: ключевых фраз на базе тезауруса WordNet-Affect и авторских символов выражения эмоций, которые используются при генерации обучающей выборки

(*Отзывы прилагаются*).

Председатель

Слово для ответа на замечания по заключению и отзывам предоставляется соискателю.

Соискатель

Дам небольшой комментарий на замечения. Да, в автореферате опущен размер словарей из-за необходимости соблюдать размер автореферата, это указано только в тексте диссертации. Второе, что я хотел бы уточнить - значение критерия схожести 0,85 в задаче объединение профилей был получен экспериментальным путем. С остальными озвученными замечаниями ведущей организации о диссертации и озвученными замечаниями отзывов на автореферат согласен.

Председатель

Слово для отзыва предоставляется официальному оппоненту - **д.т.н. Куприянову Александру Викторовичу.**

Добрый день, уважаемый председатель, уважаемые члены диссертационного совета. Прежде всего, разрешите поблагодарить за оказанную честь, потому что я впервые в вашем университете, и мне очень приятно выступить в роли официального оппонента.

Я кратко озвучу основные тезисы своего отзыва и пройдусь по замечаниям. Сразу хотел бы сказать, что работа мне очень понравилась, она импонирует тем, что это действительно очень актуальная тема, и, как это уже обсуждалось, мы ожидаем, что есть много работ, которые анализируют психолингвистические особенности текстов, но на самом деле это не так. Вся психологическая литература нацелена на свои аспекты, а вот что касается лингвистики - работы сейчас идут очень "пионерские", и только последние 3 года начался

рост публикаций, и много диссертаций стало защищаться по данной тематике.

Основная проблема, которая стоит в данной проблемной области, и она стояла, в том числе, и перед соискателем – проблема структурирования информации, которая извлекается из социальных сетей, поскольку там авторы пишут неструктурированно, очень сложно понять, что за чем следует, сложно понять логику высказывания. Всё это сильно усложняет работу. Т.е. если у нас бы был законченный семантический анализ, то было бы, конечно, гораздо проще строить все эти модели. Именно поэтому, я считаю, было очень правильно выбрано направление системного анализа, совмещающее в себе все характеристики междисциплинарного подхода, а соискатель эти методы очень эффективно применил - устранил негативные факторы: борьба с многозначностью, с динамикой в тексте, когда одно и то же слово начинает менять смысл в зависимости от контекста.

Диссертация очень большая, но читается легко, поскольку автор излагает собственным языком, с одной стороны – достаточно научно, но с другой – очень понятно всё выглядит, и читать вполне приятно.

Что касается структуры – она типичная. В первой части описывается состояние методов, вторая часть – это группы методов, которые автор исследует и которые связаны с оценкой тональности, а третья часть описывает авторский подход.

Что еще я хотел бы подчеркнуть – это теоретическая значимость. Прежде всего, методы системного анализа нашли здесь свое крайне понятное применение, поскольку необходимо было построить логическую модель того, что мы исследуем, т.е. что мы имеем в виду, когда анализируем текст. И здесь автор предлагает свою, важную и значимую модель.

Автореферат соответствует диссертации. Те ошибки, которые были в диссертации, также перекочевали и в автореферат.

Теперь позвольте чуть более детально остановиться на основных замечаниях. Некоторые замечание отпадают, да и в вопросах они звучали, но, тем не менее, я вынужден их озвучить, потому что тоже их написал.

1) В диссертации не указано, как формировалась тестовая выборка профилей социальных сетей в задаче психолингвистического анализа этих профилей и насколько эта выборка достоверна. При этом сам автор признаёт, что выборка с большой вероятностью является нерепрезентативной, поскольку добровольцы имеют особый психологический портрет.

2) В диссертации описывается метод определения эмоциональной окраски предложения, предлагающий некую среднюю оценку эмоциональной окраски. Нигде не учитывается тот факт, что внутри одного предложения, в частности, сложносочиненного, могут встречаться различные и даже противоположные эмоции по отношению к разным объектам.

3) Критерий схожести профилей, описанный в пункте 2.3.1 и критерий схожести лиц, описанный в пункте 2.3.2, приведены в работе слишком кратко и поверхностно, без математической формулировки. В представленном объёме работы учёт этих критерием не представляется существенным и целесообразным.

4) В пунктах 2.2 и 2.4 написано, в частности, в описании алгоритма, что для исправления грамматических ошибок использована

библиотека DeepPavlov. Однако в тексте диссертации не встречается описания ее работы и функциональности.

5) В работе был применен словарь ключевых фраз, построенный на базе тезауруса WordNet-Affect, который, наряду со словарем авторских символов выражения эмоций, был использован при генерации обучающей выборки, однако размер данного словаря нигде не указан.

6) В пункте 3.2.1 встречаются участки текста, структурно являющиеся таблицами, которые состоят из одной строки. Эти таблицы не указаны как таблицы, не пронумерованы и не учитываются в общее количество таблиц диссертации.

7) Оценка качества разработанных алгоритмов проводится на основе сравнения точности классификации, при этом в тексте диссертации отсутствует формальное математическое определение используемого понятия. При этом в большей части диссертации точность приводится в абсолютных значениях, а в таблице 4.6 и в заключении точность приводится в процентах.

8) Основные положения, выносимые на защиту, содержат абстрактные формулировки полученных результатом («позволяет классифицировать», «повышает точность» и т.п.) и только в выводах к главам автор приводит конкретные значения улучшаемых показателей.

9) Создаётся ощущение, что задача объединения профилей пользователей в различных социальных сетях, задача сентимент-анализа текстовых данных и задача психолингвистического анализа текста профилей социальных сетей являются тремя различными задачами с различным математическим аппаратом и различными исходными данными.

10) В диссертации на отдельных страницах встречаются некоторые ошибки в оформлении:

- частично нумерация разделов, таблиц и рисунков оформлена не по ГОСТу.
- небрежность форматирования приводит к наличию малозаполненных страниц (например, на пустом месте страницы 117 следовало разместить рисунок 4.1, расположенный на следующей странице).
- либо таблицы начинаются внизу страницы (например, таблица 1.1. на стр. 42), вследствие чего появляются разрывы и переносы таблицы на следующую страницу, что снижает удобочитаемость работы.

Содержание диссертационной работы соответствует паспорту специальности 05.13.01 – «Системный анализ, управление и обработка информации (информационные технологии и промышленность)», а именно пунктам п. 4 – разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации и п. 10 – методы и алгоритмы интеллектуальной поддержки при принятии управленческих решений в технических, экономических, биологических, медицинских и социальных системах.

Диссертация Андреева Ильи Алексеевича на тему «Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя» является самостоятельно выполненной, завершенной научно-квалификационной работой. Цель диссертация была достигнута, новые научные результаты имеют существенное научное и практическое значение.

По содержанию и полученным результатам данная диссертация удовлетворяет критериям Положения о присуждении ученых степеней, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, а её автор, Андреев Илья Алексеевич, заслуживает

присуждения степени кандидата технических наук по специальности 05.13.01 – "Системный анализ, управление и обработка информации (информационные технологии и промышленность) ".

(*Отзыв прилагается*).

Председатель

Соискателю предоставляется слово для ответа на замечания оппонента.

Соискатель

С замечаниями оппонента, доктора технических наук, Куприянова Александра Викторовича, согласен.

Председатель

Слово для отзыва предоставляется официальному оппоненту – **к.т.н. Абрамову Максиму Викторовичу.**

Уважаемые коллеги, добрый день. Я тоже хочу выразить признательность за то, что мне удостоилась честь выступить вторым оппонентом данной диссертации. Уже было задано много вопросов, большая дискуссия сложилась, другой оппонент подробно высказался, поэтому, с вашего позволения, я буду краток и отмечу основные тезисы отзыва, который я представил.

Хочется отметить актуальность тематики, в которой работает соискатель. Несмотря на то, что социальные сети популярны уже почти 20 лет и ведется большое количество исследований, в том числе, по оценке личностных особенностей пользователей социальных сетей, тем не менее, устоявшегося инструментария общеизвестного, общедоступного до сих пор не создано. Существует большое количество приложений, полученных результатов в различных областях, в частности в области, в которой мы занимаемся – анализа защищенности пользователей в информационных системах от социоинженерных атак, но есть много и других областей.

Хочется отметить, что, поскольку исследование междисциплинарное, соискателю пришлось проделать очень большую работу, которая связана не только с погружением в "не свою" предметную область, но и с тем, что необходимо корректным образом разделить результаты – те, которые относятся к области, по которым защищается диссертация, и смежных областей.

Диссертация достаточно аккуратно оформлена, есть, конечно, небольшие замечания, и они нашли свое отражение в отзыве. Но в целом она хорошо структурирована, правильно оформлена, обладает научной новизной и практической значимостью. Практическая значимость подтверждается, в том числе, актами внедрения в реальные производственные процессы.

С вашего позволения я сосредоточусь на официальной части, которую должен озвучить. Замечания по диссертационной работе:

1. В работе, в частности в пунктах 2.1 и 2.3, представлена онтологическая модель унификации данных профилей различных социальных сетей для объединения в один профиль пользователя, состоящая из нескольких критериев. Один из критериев – подход к объеди-

нению профилей посредством поиска совпадения социальных графов описан недостаточно подробно, представлена только модель.

2. В настоящее время существует большой набор модификаций языковых моделей, каковыми, например, являются RuBERT, ELMo, GloVe. Данные модели могли показать сходную или даже лучшую эффективность при применении их к русскоязычным текстам. В работе же приведены только эксперименты по оценке эффективности алгоритмов сентимент-анализа с использованием только двух языковых моделей - Word2Vec и BERT, при этом их применение не обосновано.

3. В выводах по главе 4 отмечено, что алгоритмы классификации текстов, с целью определения психолингвистических характеристик пользователя социальной сети, основаны на методе случайного леса и методе опорных векторов. Однако эксперименты проводились как минимум на 4 методах: метод опорных векторов, метод случайного леса, наивный Байесовский классификатор и логистическая регрессия.

4. В главе 3 описан разработанный в рамках диссертации комплекс программ. Однако составляющие разработанного комплекса в тексте диссертации не имеют одного общего названия. Встречаются определения «система», «подсистема», «модули», «сервисы» и «микросервисы».

5. По всей диссертации встречаются названия на английском языке. Так, например, на странице 128 представлены объекты классов «Agreebleness», «Openness to experience», «Conscientiousness» и «Neuroticism», хотя далее по тексту и в итоговом предложенном алгоритме данные классы озаглавлены на русском языке.

Указанные замечания не являются определяющими в оценке работы, не снижают высокого уровня диссертационного исследования и не снижают научную и практическую ценность. В диссертации Андреева Ильи Алексеевича на тему «Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя» сформулирована и решена научная задача анализа слабоструктурированных данных социальных сетей с целью построения социального портрета пользователя. Работа по затронутой тематике соответствует паспорту специальности 05.13.01 - "Системный анализ, управление и обработка информации (информационные технологии и промышленность)".

Автореферат отражает основное содержание диссертационной работы, а сама работа удовлетворяет требованиям пунктам 9-14 «Положения о присуждении ученых степеней», утверждённого Постановлением Правительства Российской Федерации, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, а её автор, Андреев Илья Алексеевич, заслуживает присуждения степени кандидата технических наук по специальности 05.13.01 - "Системный анализ, управление и обработка информации (информационные технологии и промышленность)".

(*Отзыв прилагается*).

Председатель

Слово для ответа на замечания оппонента предоставляется соискателю.

Соискатель

С замечаниями оппонента, кандидата технических наук, Абрамова Максима Викторовича, согласен.

Председатель

Кто хочет выступить?

Д.т.н., доцент Наместников А.М.

Уважаемые коллеги, не смотря на то, что тема диссертационного исследования, относится, в том числе, и к моим научным интересам, я задавал не так много вопросов. Это связано с тем, что многие вопросы были заданы на многочисленных научных семинарах, которые проходили на нашей кафедре – Информационные системы. Многие вопросы были поставлены, и на них были получены ответы на заседании научно-технического совета, и я могу сказать, что соискателем была проделана очень большая работа, и очень было сложно тот объем работы, что был выполнен, уместить в границы научного доклада, который мы послушали сегодня.

Я могу сказать, что результатов было получено значительно больше, чем те, которые были озвучены сейчас, поскольку Илья Алексеевич активно принимал участие в хоз. договорной работе, которая выполнялась по заказу НПО Марс, и это, наверное, одна из точек, когда работа зарождалась. Несколько лет назад мы выполняли такую большую работу, был договор с этим предприятием, и работа была настолько сложной, что мы выполняли ее в два этапа. Были, по сути, две работы, было два договора, каждый из которых выполнялся в течение одного года, т.е. два года практической работы, где были получены практические результаты, часть из которых были положены в основу данной диссертационной работы.

Кроме того, я могу сказать, что была проделана очень большая работа, связанная с тем, что для русского языка очень мало так называемых датасетов, на которых можно экспериментировать. И если в англоязычном мире такой проблемы нет, то вот для русского языка есть такая проблема. И одной из задач, которая решалась в данной работе – задача формирования множества размеченных ресурсов, которые, в дальнейшем, использовались для решения задачи классификации, это сама по себе очень большая работа. Есть диссертации, которые только этому и посвящены – формированию обучающей выборки, которая удовлетворяет определенным ограничениям - репрезентативность, сбалансированность и т.д. Поэтому, повторюсь, была проделана большая работа, и я буду голосовать "за".

Д.т.н., профессор Клячкин В.Н.

Я познакомился с этой научной работе на научно-техническом совете. После этого совета я достаточно подробно её прочитал, при том я прочитал не только реферат, но и презентацию, целый ряд замечаний у меня был. Я переписывался с автором диссертации по этому поводу. Я, конечно, не специалист в области психолингвистического анализа, с этой точки зрения я могу опираться только на мнения оппонентов, на отзывы и так далее. Но в том, что работа достаточно

актуальная, я уверен. И самое главное - ее достаточно интересно читать даже неспециалисту в этой области.

В тоже время я хотел бы сказать о ряде замечаний. Передо мной основные итоги. Вот, например, наиболее эффективным алгоритмом является многослойный персептрон и 87% точности. Так заявлять нельзя. Возьмем другую выборку, и окажется, что другая нейронная сеть покажет лучшие результаты. Также следующий пункт - наилучший результат в классификации социальных сетей показал SVM. Даже в вашей задаче это не очень корректное утверждение, потому что вы там перечислили четыре метода, а этих методов - несколько десятков. И random forest - это серьёзный метод, но вполне возможно, что какой-нибудь градиентный бустинг дал бы лучший результат. Т.е. вот так говорить, что этот метод - лучший, нельзя. Про соотношение тестовой и обучающей выборки я уже сказал ранее. Я говорил это не абстрактно, у меня два аспиранта занимались этими вопросами в совершенно разных задачах, и часто получалось, что разброс за счет случайности формирования тестовой выборки оказывается такого же порядка, как это разделение.

Понятно, что любая диссертация имеет свои замечания, недостатки, тем не менее, эти недостатки не носят принципиального характера. В целом, работа интересная, полезная и заслуживающая присуждения степени кандидата наук. Я буду голосовать "за".

д.т.н., доцент Киселев С.К.

Действительно, то замечание, которое я высказывал, не относится к методам, которые Илья Алексеевич исследовал, оно относится к исходным данным. Я уже на научно-техническом совете высказывал мнение, что эти исходные данные можно назвать "грязными", т.е. вот эта выборка, когда люди шли и открыто предоставляли свои социальные сети - не знаю, насколько она репрезентативна. Я еще тогда приводил пример: когда человек может написать какую-то шутливую фразу, поставить смайлик, который не показывает его отношение к фразе, и получить какого-то рода каламбур. И как это анализировать - не совсем понятно. Поэтому чистота исходных данных вызывает сомнения. И когда я говорил про психолингвистику, я бы хотел услышать, что есть какие-то методы, которые позволяют верифицировать эти исходные данные, насколько они адекватные, если уж, выражаться математическим языком. По этому вопросу у меня так и не осталось удовлетворенности, потому что я не сомневаюсь, что сортируется, классифицируется правильно, все, что надо делается, но вот если в исходных данных есть «грязь», то она и в результатах будет. А сами методы сомнений не вызывают. То, что работа действительно актуальна - тут вопросов нет. Я думаю, что эти работы действительно сейчас находятся на острие, они будут развиваться, поэтому такую работу, я думаю, стоит поддержать.

Председатель

Коллеги, я тоже добавлю буквально немного, потому что, мне кажется, всесторонне рассмотрели эту работу. С одной стороны, я поддерживаю высказывания, которые сейчас высказал Сергей Константинович, и они абсолютно правильные. Я боюсь только, что эти сомнения такие работы будут преследовать всегда, потому что это построение

обучающей выборки с точностью до такой науки, как психология. А вот насколько она точна – не знаю. То есть, наверное, всегда будут такие моменты, что всё, что связано с психикой человека, имеет достаточно большую степень неопределенности, и, разумеется, никогда, по крайней мере, для массового использования и первичной обработки текстов никакие тонкие моменты, которые сказываются - виде сарказма или еще чего-то поймать сложно, или, по крайней мере, очень трудоемко, а к точности это прибавит немного. Поэтому мы точно должны понимать, что работа выполнена, по моему впечатлению, добросовестно, но с точностью до тех психологических методик и тех психолингвистических методов, которые были заимствованы и автоматизированы в рамках этой работы.

Еще один момент хочу пояснить. Я знаю характер работы Ильи Алексеевича, он не просто хороший специалист по информационным технологиям, он на самом деле постоянно занят настройкой высоконагруженных систем, поэтому я надеюсь, что его квалификация позволит ему в его будущих работах построить такой инструмент, который будет справляться с миллионами, миллиардами постов непосредственно в рамках эффективного инструментария, потому что у него есть для этого компетенция, которая сейчас даже и не использована, потому что все строилось достаточно камерно в рамках системы поддержки принятия решений для отработки методов. Я вижу перспективы этой работы, поэтому ее поддерживаю, и буду голосовать положительно.

Председатель

Кто еще хочет выступить? Нет желающих? (Нет)

Соискателю предоставляется заключительное слово.

Соискатель

Хочу поблагодарить всех присутствующих, в том числе и за замечания, в будущих работах буду учитывать озвученные замечания. Хочу поблагодарить своего научного руководителя, который мне во всем помогал и консультировал, и членов диссертационного совета за то, что заслушали мою работу и оценили её.

Председатель

Переходим к голосованию.
Прошу голосовать.

(Ученый секретарь организует тайное голосование.)

Председатель

Прошу ученого секретаря озвучить результаты тайного голосования.

Ученый секретарь

Оглашаются результаты тайного голосования.

Председатель

Кто против? (Нет).
Кто воздержался? (Нет).
Результаты голосования утверждаются.

Таким образом, на основании результатов тайного голосования (за - 18 , против - нет, непроголосовавших - нет) диссертационный совет Д212.277.04 при Ульяновском государственном техническом университете признает, что диссертация **Андреева И.А.** содержит новые решения по исследованию методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя, соответствует требованиям, предъявляемым к кандидатским диссертациям (**п.9** "Положения" ВАК), и присуждает **Андрееву Илье Алексеевичу** ученую степень кандидата технических наук по специальности **05.13.01.**

Председатель

У членов Совета имеется проект заключения по диссертации **Андреева И.А.** Есть предложение принять его за основу. Нет возражений? (Нет). Принимается.

Какие будут замечания, дополнения к проекту заключения?

(Обсуждение проекта).

Председатель

Есть предложение принять заключение в целом с учетом редакционных замечаний. Нет возражений? Принимается единогласно.

Заключение объявляется соискателю.

ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА Д 212.277.04,
СОЗДАННОГО НА БАЗЕ
ФГБОУ ВО «УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
ПО ДИССЕРТАЦИИ
НА СОИСКАНИЕ УЧЕНОЙ СТЕПЕНИ КАНДИДАТА НАУК

аттестационное дело № _____
решение диссертационного совета от 14.09.2022 № 15

О присуждении Андрееву Илье Алексеевичу, гражданину Российской Федерации, ученой степени кандидата технических наук.

Диссертация «Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя» по специальности 05.13.01 – Системный анализ, управление и обработка информации (информационные технологии и промышленность) принята к защите 22.06.2022 (протокол заседания № 8) диссертационным советом Д 212.277.04, созданным на базе ФГБОУ ВО «Ульяновский государственный технический университет» (432027, Россия, г. Ульяновск, ул. Северный Венец, д. 32) приказ №678/НК от 18.11.2020 г.

Соискатель Андреев Илья Алексеевич, 24 апреля 1994 года рождения.

В данный момент соискатель обучается на 4 курсе аспирантуры в ФГБОУ ВО «Ульяновский государственный технический университет», работает старшим преподавателем кафедры «Информационные системы» в ФГБОУ ВО «Ульяновский государственный технический университет». Диссертация выполнена в ФГБОУ ВО «Ульяновский государственный технический университет».

Научный руководитель – кандидат технических наук (05.13.12 Системы автоматизации проектирования), Мошкин Вадим Сергеевич, доцент кафедры «Информационные системы» ФГБОУ ВО «Ульяновский государственный технический университет».

Официальные оппоненты:

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

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

дали положительные отзывы на диссертацию.

Ведущая организация

Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича» в своем положительном отзыве, подписанном заведующим кафедры инфокоммуникационных систем, к.т.н., Зарубиным Антоном Александровичем, профессором кафедры инфокоммуникационных систем, д.т.н., Гольдштейном Борисом Соломоновичем, доцентом кафедрой инфокоммуникационных систем, к.т.н. Елагиным Василием Сергеевичем и утвержденном Шестаковым Александром Викторовичем, и.о. проректора по научной работе, д.т.н., с.н.с., указала, что диссертация является законченной научно-квалификационной работой, выполненной на актуальную тему. Научные результаты, полученные в диссертации, обладают новизной и направлены на решение научной задачи, которая имеет большое значение для развития методов решения задач системного анализа.

Соискатель имеет 32 опубликованные работы, все они опубликованы по теме диссертации, в том числе статьи в изданиях, рекомендованных ВАК 4 работы, 11 работ в изданиях, индексируемых в Scopus и/или Web Of Science.

Наиболее значимые научные работы по теме диссертации:

1. Андреев, И. А. Алгоритм психолингвистического анализа текстовых данных социальных сетей с применением модели «Большая пятёрка» / Андреев И. А., Ярушкина Н. Г., Мошкин В. С. // Онтология проектирования. – 2022. – Т. 12, №1 (43). – С. 82–92. (лично соискателем – 9 страниц).

2. Андреев, И. А. Комбинирование статистического и лингвистического методов для извлечения двухсловных терминов из текста / Андреев И. А, Башаев В. А., Клейн В. В., Ярушкина Н. Г. // Авто-

матизация процессов управления. – 2013. – № 4 (34). – С. 61–70. (лично соискателем – 6 страниц).

3. Andreev, I.A. Solving the problem of determining the author of text data using a combined assessment // Yarushkina N.G., Moshkin V.S., Andreev I.A. // Proceedings of 8th International Conference «Fuzzy Systems, Soft Computing and Intelligent Technologies 2020 (FSSCIT 2020)» , CEUR WS Proceedings, Vol-2782, pp. 112-118. (лично соискателем – 5 страниц).

4. Andreev, I. The Sentiment Analysis of Unstructured Social Network Data Using the Extended Ontology SentiWordNet / Andreev, I., Moshkin V., Yarushkina N. // IEEE, 12th International Conference on Developments in eSystems Engineering (DeSE) – Kazan, Russia – 2019 – pp. 576-580. (лично соискателем – 4 страницы).

5. Andreev, I. Approaches to sentiment analysis of the social network text data / Andreev, I. Yarushkina N., Moshkin V. // Proceedings of the Data Science Session at the VI International Conference on Information Technology and Nanotechnology (DS-ITNT 2020) – Vol. 2667. – pp. 198–202. (лично соискателем – 3 страницы).

На диссертацию и автореферат поступили 4 отзыва, все отзывы положительные, в отзывах содержатся следующие замечания:

1. ФГБОУ ВО «Ульяновский институт гражданской авиации имени главного маршала авиации Б.П. Бугаева» (г. Ульяновск). Отзыв подписан доцентом кафедры «Организации аэропортовой деятельности и информационных технологий» Чоракаевым О.Э. (к.т.н., специальность 05.13.12).

Замечания:

•В автореферате можно заметить следующий недостаток: приведены эксперименты по оценке эффективности алгоритмов классификации тональности с использованием только двух языковых моделей – Word2Vec и BERT, при этом не встречается обоснования их применения. В настоящее время существует большой набор модификаций языковых моделей (ELMo, GloVe, TinyBERT), которые могли бы показать высокую эффективность в данной задаче.

2. Филиал ФГБОУ ВО «Национальный исследовательский университет «МЭИ» (г. Смоленск). Отзыв подписан ведущим программистом лаборатории Луферовым В.С. (к.т.н., специальность 05.13.01).

Замечания:

- Одним из пунктов, вынесенных автором в качестве научной новизны, является методика объединения профилей пользователей из различных социальных сетей. В экспериментах показано, что текущего набора алгоритмов сравнения данных профилей, примененных последовательно, достаточно для объединения профилей 100 человек (всего 204 профиля) разных социальных сетей. Однако, данная методика может показать иной уровень эффективности при обработке данных более крупной выборки.
- В списке публикаций встречаются опечатки и ошибки форматирования.

3. ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации» (г. Москва). Отзыв подписан доцентом департамента анализа данных и машинного обучения Факультета информационных технологий и анализа больших данных Андрияновым Н.А. (к.т.н., специальность 05.13.18).

Замечания:

• В описании алгоритма формирования обучающей выборки Шаг №4 включает в себя предобработку данных. Однако предобработка – это достаточно важный шаг, детальное описание которого в автореферате опущено.

• В таблице 1 и далее приводятся метрики точности разных моделей. Однако отсутствует информация о распределении классов в обучающей и тестовой выборках, без которой сложно судить о качестве работы модели, даже если точность составляет 87%.

• В таблице 2 в качестве алгоритма классификации указана «Линейная регрессия». Вероятно, автор имел в виду логистическую регрессию?

4. ФГАОУ ВО «Казанский (Приволжский) федеральный университет» (г. Казань). Отзыв подписан доцентом кафедры информационных систем Невзоровой О.А. (к.т.н., специальность 05.13.17).

Замечания:

•Недостатком автореферата является тот факт, что автором не указан объем применяемых им словарей: ключевых фраз на базе тезауруса WordNet-Affect и авторских символов выражения эмоций, которые используются при генерации обучающей выборки

Выбор официальных оппонентов и ведущей организации обосновывается их высокой компетенцией, научными достижениями и наличием публикаций в соответствующей отрасли наук.

Диссертационный совет отмечает, что на основании выполненных соискателем исследований:

проведено сравнение современных интеллектуальных методов анализа текстовых данных, их возможностей и ограничений в рамках психолингвистического и сентимент-анализа ресурсов социальных сетей;

разработан алгоритм формирования обучающей выборки, состоящей из открытых русскоязычных текстовых ресурсов социальных сетей, классифицированных по семи эмоциям, отличающийся совместным использованием словарей авторских символов выражения эмоций и ключевых фраз;

предложен подход к определению психологических характеристик пользователя социальных сетей посредством психолингвистического анализа текстовых сообщений в социальных сетях с использованием методов машинного обучения.

Теоретическая значимость исследования обоснована тем, что:

доказаны положения о необходимости развития алгоритмов классификации неструктурированных данных социальных сетей по классам тональности, вносящие вклад в развитие моделей и методов обработки текстовой информации, которые основаны на совместном применении искусственных нейронных сетей и семантических подходов.

Применительно к проблематике диссертации результативно (эффективно, то есть с получением обладающих новизной результатов)

использован комплекс существующих методов машинного обучения с учителем на основе рекуррентных, свёрточных, двунаправленных рекуррентных нейронных сетей, семантических методов и онтологических моделей;

изложены основные научные положения, гипотезы и рекомендации, позволяющие повысить эффективность построения социального портрета пользователя социальной сети на основе психолингвистического анализа текстовых данных пользователя;

раскрыты принципиальные ограничения методов машинной обработки, заключающиеся в необходимости обучения с подкреплением класси-

фикатора, в задачах определения психологических характеристик человека путем обработки коротких текстов социальных сетей;

изучены алгоритмы формирования обучающей выборки, алгоритмы сентимент-анализа, базирующиеся на использовании лингвистических словарей и методах машинного обучения, подходы к объединению профилей пользователей и алгоритмы построения психологического портрета;

проведена модернизация существующих алгоритмов формирования обучающей выборки текстов социальных сетей, подходов к сопоставлению профилей пользователей, алгоритма анализа эмоциональной окраски русскоязычных текстовых данных социальных сетей в системах поддержки принятия решений для обеспечения качественной и эффективной оценки профиля пользователя.

Значение полученных соискателем результатов исследования для практики подтверждается тем, что:

разработан и внедрен в эксплуатацию новый программный комплекс психолингвистического и сентимент-анализа открытых текстовых русскоязычных данных профилей пользователей социальных сетей. Применение разработанного программного комплекса сократило временные затраты в задачах подбора персонала на основании результатов психолингвистического анализа и анализа эмоциональной окраски текстовых ресурсов профилей соответствующих пользователей. В абсолютных значениях экономия составила более 14 часов для каждой итерации подбора персонала;

в рамках проекта «Система интеллектуального поиска и анализа Интернет-СМИ в социальных сетях» в задачах поиска пользователей с учетом параметров достигнуто среднее сокращение времени поиска необходимых профилей на 40%;

определены границы применимости разработанных методов и алгоритмов психолингвистического и сентимент-анализа социальных сетей;

создан и применен программный комплекс классификации профилей пользователей по психологическим характеристикам и анализу высказываний об объектах реального мира в форме текстовых сообщений;

представлены предложения по дальнейшему совершенствованию моделей и алгоритмов психолингвистического и сентимент-анализа текстовых данных социальных сетей.

Оценка достоверности результатов исследования выявила:

для экспериментальных работ результаты получены на основе корректного использования методов анализа неструктурированных ресурсов и современного программного обеспечения, подтверждены вычислительными экспериментами и результатами практического применения в лабораторных и производственных условиях ФГБОУ ВО Ульяновский государственный технический университет, УОСОО «Федерация бадминтона», ООО «ФаззиЛаб» и ФНПЦ АО "НПО"Марс";

теория построена на известных научных данных, которые в полной мере согласуются с ранее опубликованными данными по теме диссертационного исследования;

идея базируется на анализе и обобщении передового опыта и практических исследованиях ряда российских и зарубежных исследователей по теме диссертации;

использованы сравнения авторских данных и данных, полученных ранее по рассматриваемой тематике для схожего набора данных на английском языке;

установлено качественное совпадение результатов, полученных автором, с опубликованными ранее результатами аналогичных исследований в области психолингвистического и сентимент-анализа текстовых данных социальных сетей;

использованы современных методики сбора и обработки текстовой информации, обеспечивающие воспроизводимость и достоверность результатов.

Личный вклад соискателя состоит в: анализе научных и патентных источников по теме диссертационной работы, обработке и интерпретации аналитической информации, разработке моделей, методов и алгоритмов, планировании и проведении экспериментов, формулировке выводов, внедрении полученных результатов. Все основные исследования проведены лично автором, либо при его непосредственном участии.

В ходе защиты диссертации было высказано следующее критическое замечания – при формировании обучающей выборки не учитываются стилистические особенности русскоязычных текстовых ресурсов, влияющих на процесс классификации текстов по классам эмоциональной окраски.

Соискатель Андреев И.А. ответил на задаваемые ему в ходе заседания вопросы и привел собственную аргументацию: для формирования обучающей выборки использовались короткие текстовые сообщения, которые отличались использованием одного авторского символа выражения эмоций, что снижало вероятность влияния стилистических особенностей на результат классификации текстов с использованием полученной обучающей выборки.

На заседании 14.09.2022 диссертационный совет принял решение

за решение научной задачи формирования алгоритмов и методов психолингвистического и сентимент-анализа русскоязычных текстовых данных социальных сетей, имеющей значение для развития технической отрасли знаний, присудить Андрееву И.А. ученую степень кандидата технических наук.

При проведении тайного голосования диссертационный совет в количестве 18 человек, из них 7 докторов наук по специальности рассматриваемой диссертации, участвовавших в заседании, из 23 человек, входящих в состав совета, дополнительно введены на разовую защиту 0 человек, проголосовали: за 18, против нет, непроголосовавших членов нет.

Защита окончена. Есть ли замечания по процедуре защиты? (Нет).

Поздравляет соискателя с успешной защитой. Благодарит членов совета и всех участников за внимание.

Заседание объявляется закрытым.

Председатель Совета
д.т.н., профессор Н.Г. Ярушкина

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