МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ «САНКТ-ПЕТЕРБУРГСКИЙ ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР РОССИЙСКОЙ АКАДЕМИИ НАУК» (СПБ ФИЦ РАН)

14-я линия В.О., д. 39, г. Санкт-Петербург, 199178
Тел.: (812) 328-33-11, факс: (812) 328-44-50,
e-mail: info@spcras.ru, web: http://www.spcras.ru
ОКПО 04683303, ОГРН 1027800514411, ИНН/КПП 7801003920/780101001

ОТЗЫВ

официального оппонента на диссертационную работу Андреева Ильи Алексеевича «Исследование методов и алгоритмов обработки текстовой информации социальных сетей в задачах формирования социального портрета пользователя», представленной на соискание ученой степени кандидата технических наук по специальности 05.13.01 – «Системный анализ, управление и обработка информации (информационные технологии и промышленность)»

Актуальность избранной темы

Анализ социальных данных, хранящихся в открытых источниках, с каждым годом становится всё популярнее. Это связано в первую очередь с появлением в конце 90х–начале 00х годов социальных сетей (LiveJournal, ВКонтакте, Одноклассники, YouTube и других). Ежегодно растёт аудитория социальных сетей и вместе с тем растет и количество хранящейся в социальных сетях различного рода информации – в первую очередь аудиовизуальной и текстовой.

Социальные сети можно анализировать с разными целями. Одна из целей – оценка выраженности характеристики человека на основе его высказываний и поведения в социальных сетях, которую, в свою очередь, можно использовать для разных задач – поиск психических отклонений, подбор персонала, мониторинг общественного мнения в разрезе определённой группы людей и т.п. Для этих целей часто применяется анализ текстовых постов и комментариев, оставляемых пользователями. На основании вышеизложенного можно сделать вывод о том, что диссертационная работа посвящена актуальной теме обработки текстовой информации, содержащейся в контенте, публикуемом пользователями социальных сетей.

Оформление, структура и содержание диссертации и автореферата

Диссертация и автореферат оформлены с соблюдением ГОСТ Р 7.0.11-2011 и иных документов, регламентирующих правила оформления диссертационных работ. Язык изложения понятный, соответствует нормам русского языка и имеет научный стиль. Диссертация изложена на 166 страницах, к работе приложен список из 4 приложений. Цель исследования – снижение трудозатрат на построение социального портрета пользователя социальной сети путем автоматизации объединения профилей пользователя в различных

социальных сетях и автоматизированной классификации пользователя по психологической модели «Большой пятёрки», объект – неструктурированные и слабоструктурированные данные социальных сетей, в том числе профилей и постов пользователей.

Во введение описана актуальность исследования и проблемной области, описывается цель и задачи работы, научная новизна, структура и результаты работы.

В первой главе рассмотрено современное состояние методов анализа социальных сетей, упор сделан на методы идентификации пользователей в различных социальных сетях. Кроме того, описана группа методов, связанных с оценкой тональности текста. Приведены различные варианты применения сентимент-анализа, рассмотрена корреляция между сентимент-анализом текста и оценкой общественного мнения. Завершается глава описанием существующих методы построения психологического портрета человека, упор сделан на метод «Большой пятёрки» и его применимости для анализа публичной информации социальных сетей.

Во второй главе рассматривается разработанный алгоритм формирования обучающих выборок для сентимент-анализа. Отличием разработанного алгоритма является использование словаря авторских знаков и расширенного словаря WordNetAffect. Кроме того, описан подход к сопоставлению профилей пользователей в различных социальных сетях. Отличием является гибридизация анализа структурированных и неструктурированных данных. Среди критериев, применяемых в данном подходе, можно выделить критерий схожести лиц, критерий схожести опубликованного текста и критерий совпадения социальных графов. Глава завершается описанием адаптации метода «Большой пятёрки» для определения психологических характеристик пользователя.

Третья глава посвящена разработанному программному комплексу. Комплекс разработан с использованием микросервисной архитектуры, предполагающей использование как комплекса в целом, так и отдельных его частей. Приведена общая схема работы системы, отдельно рассмотрена каждая из его частей. Каждая часть подробно описана, к каждой приведены диаграммы классов, вариантов использования и описаны программные требования для работы.

В четвертой главе представлены подтверждающие эффективность результаты проведенных экспериментов по применению разработанных алгоритмов и описаны успешные внедрения разработанного комплекса программ.

Автореферат достаточно и полно отражает содержание диссертационной работы и содержит все необходимые составляющие.

Степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации

Достоверность научных положений, выводов и рекомендаций, сформулированных в диссертации, подтверждена результатами вычислительных экспериментов и результатами

практического использования. Результаты работы были внедрены в НТС ФНПЦ АО «НПО «Марс», в Ульяновской областной спортивной общественной организации «Федерация бадминтона» и ООО «Центр программной инженерии и аналитики «ФаззиЛаб».

Научная новизна и достоверность научных положений, выводов и рекомендаций, сформулированных в диссертации

Научная новизна и достоверность положений, выводов и рекомендаций, сформулированных в диссертации автором, состоит в:

- предложенном алгоритме анализа сентимент-анализа текстовых данных социальных сетей, который отличается интеграцией семантических подходов и методов машинного обучения, использующих новый алгоритм формирования обучающей выборки, отличающийся совместным использованием словарей авторских символов выражения эмоций и ключевых фраз.

- предложенном подходе к сопоставлению профилей пользователей социальных сетях, который отличающимся гибридизацией подходов анализа графической информации, структурированных данных анкет, текстовых данных, а также социальных графов профилей.

- предложенном методе оценки выраженности психологических характеристик пользователя социальных сетей, отличающимся применением метода «Большой пятерки».

Научная и практическая значимость работы

Теоретическая значимость представленной работы заключается в разработанных алгоритмах, подходах и методов обработки данных, хранящихся в неявном виде в социальных сетях. Практическая значимость диссертационной работы заключается в разработке программного комплекса, позволяющего упростить анализ социальных сетей для составления социального портрета человека, который можно использовать в дальнейшем для различных задач, в том числе — подборе персонала в организацию. Разработанный программный комплекс был применён в различных задачах, описанных в данной работе и внедрен в существующие организации ульяновской области.

Полнота изложения материалов диссертации в печатных работах, опубликованных автором

Результаты, полученные во время работы над диссертацией, были представлены на 17 научных конференциях. По результатам работы были опубликованы в 32 статьи, 4 из которых в журналах из «перечня ВАК» и 11 статей в изданиях, индексируемых в Scopus и/или Web Of Science, а также 1 монография.

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, состоящий из нескольких критериев. Один из критериев –

подход к объединению 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 линейная 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», «модулями», «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.

Заключение о соответствии диссертационной работы критериям, установленным Положением о порядке присуждения ученых степеней

В диссертация 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 слабоструктурированнх данных социальных сетей с целью построения социального портрета пользователя. Диссертация Андреева И.А. выполнена на достаточно высоком научном уровне и является завершенной научно-квалификационной работой. Заявленная цель диссертации была достигнута, диссертация имеет применимые научные и практические результаты.

Работа по затронутой тематике соответствует паспорту специальности 05.13.01 – «Системный анализ, управление и обработка информации (информационные технологии и промышленность)»:

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.

Как следует из вышеизложенного, представленная работа удовлетворяет требованиям пп. 9–14 «Положения о присуждении ученых степеней», утверждённого Постановлением Правительства Российской Федерации от 24.09.2013 №842 (в ред. Постановления Правительства РФ от 11.09.2021 №1539), 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)».

Официальный оппонент
Старший научный сотрудник,
руководитель лаборатории теоретических и
междисциплинарных проблем информатики СПб ФИЦ РАН,
Кандидат технических наук (специальность 05.13.19)
12 июля 2022 г. _____ Максим Викторович Абрамов

Личную подпись руки Максима Викторовича Абрамова, старшего научного сотрудника с возложенными обязанностями руководителя лабораторией теоретических и междисциплинарных проблем информатики и автоматизации СПб ФИЦ РАН, кандидата технических наук (специальность 05.13.19), удостоверяю.

Начальник отдела кадров СПб ФИЦ РАН Д.В. Токарев

Контактные данные:

Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН),
199178, г. Санкт-Петербург, 14-я линия В.О., д. 39,
тел: +7 (812) 328-33-11,
факс: +7 (812) 328-44-50,
web-сайт: https://spcras.ru, https://dscs.pro/
e-mail: mva@dscs.pro