Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

«Восстановление профиля пользователя по данным из многих источников»

Автор: Бережко Герман Игоревич _____
Направление подготовки (специальность): 01.03.02 Прикладная математика и информатика
Квалификация: Бакалавр
Руководитель: Фильченков А.А., к.ф.-м.н., доц. каф. КТ _____

К защите допустить
Зав. кафедрой Васильев В.Н., докт. техн. наук, проф. _____

«____» _____ 20_____ г.

Санкт-Петербург, 2017 г.

Студент Бережко Г.И. **Группа** М3438 **Кафедра** компьютерных технологий **Факультет** информационных технологий и программирования

Направленность (профиль), специализация Математические модели и алгоритмы разработки программного обеспечения

Квалификационная работа выполнена с оценкой _____

Дата защиты «____» июня 2017 г.

Секретарь ГЭК _____

Листов хранения _____

Демонстрационных материалов/Чертежей хранения _____

Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

УТВЕРЖДАЮ
Зав. каф. компьютерных технологий
докт. техн. наук, проф.
_____ Васильев В.Н.
«____» _____ 20_____ г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Студент Бережко Г.И. **Группа** М3438 **Кафедра** компьютерных технологий **Факультет** информационных технологий и программирования **Руководитель** Фильченков Андрей Александрович, к.ф.-м.н., доц. каф. КТ,

1 Наименование темы: Восстановление профиля пользователя по данным из многих источников

Направление подготовки (специальность): 01.03.02 Прикладная математика и информатика

Направленность (профиль): Математические модели и алгоритмы разработки программного обеспечения

Квалификация: Бакалавр

2 Срок сдачи студентом законченной работы: «31» мая 2017 г.

3 Техническое задание и исходные данные к работе.

По набору данных из трех социальных сетей (Twitter, Instagram, Foursquare) из трех городов (Сингапур, Нью-Йорк, Лондон) построить модель, позволяющую восстанавливать демографические характеристики пользователя, такие как пол, возраст, уровень образования, род занятий

4 Содержание выпускной квалификационной работы (перечень подлежащих разработке вопросов)

а) Поиск и исследование существующих алгоритмов для решения поставленнойзадачи
б) Разработка и реализация методов классификации, которые помогут в решении поставленной задачи
в) Сравнение результатов с существующими решениями

5 Перечень графического материала (с указанием обязательного материала)

Не предусмотрено

6 Исходные материалы и пособия

а) Aleksandr Farseev, Liqiang Nie, Mohammad Akbari and Tat-Seng Chua. Harvesting Multiple Sources for User Profile Learning: a Big Data Study

б) Daniel Preotiuc-Pietro, Vasileios Lampos and Nikolaos Aletras. An analysis of the user occupational class through Twitter content.

в) Jiwei Li, Alan Ritter, Eduard Hovy. Weakly Supervised User Profile Extraction from Twitter.

7 Календарный план

№№ пп.	Наименование этапов выпускной квалификационной работы	Срок выполнения этапов работы	Отметка о выполнении, подпись руков.
1	Обзор предметной области	31.12.2016	
2	Анализ и реализация существующих решений	10.02.2017	
3	Проработка идеи решения	01.03.2017	
4	Обработка и подготовка данных	14.03.2017	
5	Релизация решения	14.04.2017	
6	Сравнение полученного решения с существующими	01.05.2017	
7	Написание пояснительной записки	25.05.2017	
8	Согласование пояснительной записки с научным руководителем ВКР	02.06.2017	
9	Представление ВКР на кафедре	05.06.2017	

8 Дата выдачи задания: «01» сентября 2016 г.

Руководитель _____

Задание принял к исполнению _____ «01» сентября 2016 г.

Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Студент: Бережко Герман Игоревич

Наименование темы работы: Восстановление профиля пользователя по данным из многих источников

Наименование организации, где выполнена работа: Университет ИТМО

ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

1 Цель исследования: Разработать модель для восстановления демографических характеристик пользователей социальных сетей по данным из многих источников

2 Задачи, решаемые в работе:

а) исследование поставленной задачи;

б) анализ текущих решений и подходов;

в) составление модели, решающую поставленную задачу.

3 Число источников, использованных при составлении обзора: 10

4 Полное число источников, использованных в работе: 31

5 В том числе источников по годам

Отечественных			Иностранных		
Последние 5 лет	От 5 до 10 лет	Более 10 лет	Последние 5 лет	От 5 до 10 лет	Более 10 лет
6	0	0	19	3	3

6 Использование информационных ресурсов Internet: 9

7 Использование современных пакетов компьютерных программ и технологий: Для реализации решения задачи был использован язык программирования Python с дополнительно подключенными библиотеками Pandas для обработки csv-файлов и sklearn для построения и обучения моделей. Используются данные технологии в главах 3 и 5.

8 Краткая характеристика полученных результатов: Была разработана модель, позволяющая определять пол, возраст, уровень образования и род деятельности с высокой точностью.

9 Гранты, полученные при выполнении работы: В рамках данной работы грантов получено не было.

10 Наличие публикаций и выступлений на конференциях по теме работы: Нет.

Выпусник: Бережко Г.И. _____

Руководитель: Фильченков А.А. _____

«____» _____ 20_____ г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В последнее время соцсети всё больше и больше влияют на нашу жизнь и сейчас довольно трудно найти человека, который не пользуется ни одной из социальных сетей, будь то Twitter, Instagram, Foursquare, Facebook или какая-то другая. Более того, все соцсети постоянно растут, привлекая людей различного пола, возраста, с разными семейными статусами, родом занятий и уровнем образования. Быстрый рост различных соцсетей очень тесно связан с ростом популярности смартфонов, потому что человек может публиковать записи, выкладывать снимки и отмечать своё текущее местоположение буквально на ходу. В большинстве случаев, люди пользуются более, чем одной социальной сетью, и поэтому сейчас все соцсети очень связаны. Так, например, при публикации фото в «Instagram» можно автоматически опубликовать её в «Twitter» и «Foursquare». Благодаря этому открывается возможность анализировать пользователей, обладая данными, содержащимися в нескольких социальных сетях. Такие попытки были предприняты в [1-3], но более успешных результатов добились авторы работы [3].

Изучение и восстановление профиля пользователя очень важно и имеет большое значение в различных областях. В качестве данных для восстановления могут служить: возраст, пол, семейное положение, место жительства, род деятельности. Быстрый рост множества социальных платформ позволяет выполнить эту задачу с разных точек зрения. Тем не менее исследовательские усилия по работе над изучением и восстановлением пользовательских профилей из нескольких источников данных достаточно скудны, к тому же нет крупномасштабного набора данных для обучения моделей, позволяющих это сделать.

Задача восстановления профиля пользователя, используя данные из нескольких социальных сетей непроста, так как какие-то сервисы заточены

только для видео (YouTube), какие-то только для фото (Instagram), а в каких-то в тексте могут содержаться и мультимедиа данные (Twitter). Еще одна проблема заключается в трудности собрать данные пользователей. Из-за соображений соглашений приватности, с каждого аккаунта пользователя можно собрать только какое-то ограниченное количество информации, которой в очень редких случаях бывает достаточно. Другой проблемой является то, что даже после того как данные будут собраны, большой трудностью окажется возможная нехватка данных, так как не все пользователи указывают аккаунты своих социальных сетей в каком-то одном месте. Так же проблемой является малое количество достоверной (ground-truth) информации о пользователе, что значительно затрудняет обучение моделей.

Целью данной работы является улучшить и усовершенствовать текущие модели и придумать новые для более точного и полного восстановления профиля.

Для достижения поставленной цели был выработан ряд задач:

- изучение и анализ существующего набора данных;
- поиск и изучение существующих решений задачи восстановления данных пользователя по данным социальных сетей;
- разработка собственной модели для восстановления профиля пользователя по данным из многих источников;
- сравнение эффективности предложенного алгоритма с уже существующими.

Новизна данной работы заключается в улучшении качества профилирования, что тем самым поднимет точность результатов и позволит лучше манипулировать ими на практике.

Знание о демографических особенностях пользователей представляет большой интерес для коммерческого сектора. На основе технологий профилирования компании могут прогнозировать поведение разных типов

своих клиентов. Затем маркетинговые стратегии могут быть адаптированы к аудитории, подходящей к этим типам. Примерами практики профилирования в маркетинге являются карты лояльности клиентов, управление взаимоотношениями с клиентами в целом и персонализированная реклама.

В контексте приема на работу, профилирование может быть полезным для отслеживания сотрудников в различных соцсетях путем оценки факторов, влияющих на найм кандидатов.

Так же предложенные алгоритмы могут быть полезны для различного администрирования и составления статистик. Так как очень часто из-за отсутствия каких-то демографических данных бывает невозможно составить полную статистику какого-либо сообщества в социальной сети, а с помощью предложенной модели это станет осуществимо.

ГЛАВА 1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1. Социальные медиа

Социальные медиа — веб-ресурсы, созданные для общения пользователей в сети [1]. Наиболее ярким примером социальных медиа являются социальные сети.

Используя ленты социальных сетей можно получить много информации о человеке: куда он ходит, какие книги он читает, какую музыку слушает. Эта же работа направлена на восстановление демографических характеристик таких как пол, возраст, род занятий, уровень образования. Существует множество работ, которые оценивают зависимость некоторых характеристик от социальной активности. В одной из последних работ – в работе [31] проведен сравнительный анализ выбора рода занятий выпускников ведущих вузов на основе данных социальной сети «Вконтакте».

Социальные сети можно косвенно разделить по типу контента, который они содержат. Так, например, Twitter – это социальная сеть, которая служит для общедоступного обмена сообщениями. Этим самым она очень удобна для составления текстовых выборок данных для различных исследований. Поэтому социальная сеть Twitter используется в качестве источника данных во многих работах. Так, например, в работе [2] Twitter используется для изучения демографических особенностей пользователей. В работе [3] на основе этой социальной сети изучается поведение пользователей. Работа [5] занимается определением рода деятельности пользователей, а в работе [6] описан способ определить уровень образования пользователя.

Другая популярная сеть – Instagram, в которой сейчас зарегистрировано около полумиллиарда пользователей и загружено порядка пятидесяти миллиарда снимков. Instagram является наилучшим источником медиаданных. Кроме того, Instagram является cross-linking social media – из Instagram можно

автоматически размещать записи в других социальных сетях. Но так как алгоритмов по работе с изображениями несколько меньше, чем алгоритмов по работе с текстом и из-за сложности скачивать фотографии из Instagram и к тому же из-за соображений приватности, то не так много работ, которые используют Instagram как источник медиаданных. В работе [7] предприняты попытки изучить тип контента, который размещается пользователями, а по нему также и тип пользователей, которые пользуются этой социальной сетью.

Другая социальная сеть, которая содержит в себе еще один тип данных – это Foursquare. Foursquare – это социальная сеть с возможностями геопозиционирования. Благодаря этому, она является самым лучшим источником геоданных. Так, в работе [8] производятся попытки описать поведенческие шаблоны для жителей по всему миру, в зависимости от их «чекинов» (геометки, оставляемые пользователями). В работе [9] описан алгоритм определения места жительства пользователя с точностью до 50-ти километров.

1.2. Демографические характеристики

К демографическим характеристикам относятся такие вещи, как пол, возраст, уровень образования, доходы, жильё, размер семьи, род занятий, мобильность и другие. В данной работе речь пойдет о четырех из них: пол, возраст, уровень образования и род деятельности.

Существуют различные социальные сети и в некоторых из них регистрация невозможна без указания, например, возраста, для того, чтобы ограничить доступ к некоторым ресурсам этой социальной сети. Где-то необходимо указывать пол, что чаще всего делается для того, чтобы не предлагать мужской части аудитории контент, связанный, например, с косметикой, а женской – с автомобилями. Некоторые характеристики являются опциональными: место учебы или работы, род деятельности, уровень образования. Пол, возраст, род деятельности и уровень образования очень

полезны для описанных во введении случаях. Каждая из этих характеристик может быть использована для таргетирования рекламы. Возраст, род занятий и уровень образования будет полезен для тех, кто работает в сфере по подбору персонала. Пол и возраст может быть полезен для администрирования сообществ в социальных сетях.

Многие пользователи социальных сетей скрывают большинство своих демографических признаков, поэтому данная работа будет полезна для их восстановления.

1.3. Описание имеющихся решений

Из существующих решений было найдено только одно, которое использует данные из нескольких социальных сетей. Это решение описано в работе [10]. Оно основано на составлении ансамбля классификаторов и совмещении их в один. Для составных классификаторов используется классификатор, основанный на методе опорных векторов, Random Forest классификатор и наивный Байесовский классификатор. После этого вводится оценка определения пола и возраста и с помощью алгоритма «Random Restart Hill Climbing» ищутся гиперпараметры, при которых достигается максимум этой формулы. Но в этой работе данный авторами алгоритм применен только к полу и возрасту. Тем не менее он показывает хорошие результаты. Для пола точность 87,8%, а для возраста 50,9%. Поэтому этот алгоритм можно использовать в качестве опорной точки.

Но кроме этого, в работе [9], как упоминалось ранее, был описан метод определения рода занятий пользователя по социальной сети Twitter. Определяется это с точностью 52,7%. В дальнейшем приведенную модель можно сравнивать с этим результатом.

В работе [10] приводится алгоритм определения уровня образования по социальной сети Twitter. Точность определения составляет 74,3%. Так же, как и

с родом деятельности, приведенную модель можно сравнивать с именно этим результатом.

Выводы по главе 1

Задача по восстановлению демографических данных пользователя по данным из многих источников всё ещё остается мало исследованной. Целью данной работы является придумать новую модель, которая будет способна превзойти текущие результаты по определению таких демографических данных, как пол, возраст, род деятельности и уровень образования. В работе определены задачи, которые помогут это сделать, изучена предметная область и приведено сравнение предложенной модели с текущими результатами.

ГЛАВА 2. ЭКСПЕРИМЕНТАЛЬНАЯ СРЕДА

В данной главе приведено описание данных, по которым будет происходить обучение моделей, описание алгоритмов классификации, которые используются, а так же описаны технологии, которые применяются.

2.1. Описание данных

Предоставленный датасет покрывает всевозможные области: текстовые, визуальные и локационные данные. Они были собраны из трех регионов, в которых активность пользователей в приложении «Foursquare» максимальна: в Сингапуре, Нью-Йорке и Лондоне. Из-за соображений приватности, некоторые данные предоставлены не в оригинале, а сразу в качестве некоторых извлеченных признаков. Чтобы не было проблемы сличения пользователей из нескольких соцсетей, при сборе данных, они собирались только с тех аккаунтов, где упоминались другие их соцсети. Поэтому в приоритете были «cross-linking» социальные сети. Так, в Twitter были найдены записи, в тексте которых есть упоминание «swarmapp.com» или «instagram.com» для получения твитов, которые были опубликованы из этих приложений. Далее, по таким твитам можно перейти в Instagram и далее перейти на профиль пользователя, чтобы скачать его картинки. Для чекинов содержится ссылка на полную информацию о чекине, а далее – ссылка на профиль, откуда можно взять другие данные.

Из соображений приватности каждому пользователю был присвоен один уникальный идентифицирующий номер, единый для всех социальных сетей.

2.1.1. Twitter

Данные из социальной сети Twitter представлены в следующем виде:

- _id – анонимизированная строка, идентифицирующая пользователя
- time_utc – время твита в формате UTC
- text – текст твита

2.1.2. Instagram

Данные из социальной сети Instagram представлены в следующем виде:

- _id – анонимизированная строка, идентифицирующая пользователя
- time_utc – время публикации изображения в формате UTC
- concepts – распределение признаков изображения, построенное на 1000 концептов ImageNet. Описание всех признаков приведено в отдельном файле. Представляет собой массив вещественных чисел.
- caption – текст поста

2.1.3. Foursquare

Данные из социальной сети Foursquare представлены в следующем виде:

- _id – анонимизированная строка, идентифицирующая пользователя
- time_utc – время чекина
- venue_id – идентификационная строка места чекина.
- venue_name – название места чекина.
- venue_location – координаты чекина.
- categories – массив идентифицирующих строк категорий места. Описания категорий приведены в отдельном файле.
- shout – текст подписи к чекину.

2.1.4. Facebook

Facebook был использован для восстановления ground-truth данных. Для сличения пользователей Facebook-Foursquare-Twitter было использовано Fousqare REST API. У 65% пользователей был указан пол, но возраст указан только у 20% пользователей. Поэтому он был установлен благодаря данным, которые пользователи оставляют о своих школах, колледжах и университетах.

Так же были взяты данные об уровне образования и занимаемой должности для дальнейшего обучения.

2.1.5. Общая статистика

Общее количество твитов, чекинов и публикаций в Instagram по каждому городу представлены в табл. 1.

Таблица 1. Статистика данных

Город	**Число пользователей**	**Число твитов**	**Число чекинов**	**Число изображений**
Сингапур	7,023	11,732,489	366,268	263,530
Лондон	5,503	2,973,162	127,276	65,088
Нью-Йорк	7,957	5,263,630	304,493	230,752

В таблице 2 приведена наполненность данных. То есть сколько пользователей имеют соответствующие демографические признаки.

Таблица 1. Наполненность данных

Город	**Возраст**	**Пол**	**Работа**	**Образование**
Сингапур	602	4691	846	1629
Лондон	0	3248	966	874
Нью-Йорк	0	1356	721	591

Было выделено 4 возрастные группы: < 20 лет, 20-30, 30-40, >40. В таблице 3 приведено процентное соотношение каждой возрастной группы относительно тех, у кого известен возраст.

Таблица 3. Распределение возрастных групп

Возрастная группа	**Процент**

< 20	25,7%
20-30	64,3%
30-40	8,5%
>40	1,5%

В таблице 4 представлено распределение по полу

Таблица 4. Половое распределение

Пол	**Процент**
Женский	40%
Мужской	60%

Было выделено 4 группы для определения уровня образования: учащиеся школы, колледжа, вуза и выпускники. В таблице 5 приведено их распределение.

Таблица 5. Распределение уровня образования

Уровень образования	**Процент**
college	32.77%
school	24.63%
graduate	1.58%
undergraduate	41.01%

В представленном датасете была выделена 21 группа родов занятий, опираясь на [22]. Процентное распределение приведено в таблице 6.

Таблица 6. Распределение рода занятий

Род занятий	**Процент**
protective service	4.89%
education, training, and library	3.63%
sales and related	6.71%
farming, fishing and forestry	0.55%

office and administrative support	7.57%
healthcare support	2.4%
life, physical, and social science	0.47%
production	1.77%
construction and extraction	0.47%
community and social service	0.39%
business and financial operations	0.75%
transportation and material moving	2.76%
archetecture and engineering	5.36%
personal care and service	4.18%
management	30.47%
healthcare practitioners and technical	0.55%
building and grounds cleaning and maintenance	0.35%
legal	1.85%
food preparation and service related	4.22%
arts, design, entertainment, sports, and media	16.38%
computer and mathematical	4.18%

2.1.6. Вывод

Из приведенные таблиц можно сделать вывод, что данные распределены более или менее равномерно, за исключением тех, кто уже выпустился из университета, их всего лишь 1,69% и некоторых специальностей, которых заметно явное преобладание: сфера искусства и менеджмента. Так же преобладающее большинство имеет возраст до 30 лет, что обусловлено небольшим распространением социальных сетей среди тех, кому больше тридцати лет. Исходя из этого, можно сделать вывод, что на таких данных можно в дальнейшем обучать модели.

2.2. Описание технологий

Для достижения поставленной задачи был использован язык программирования Python версии 3 и следующие библиотеки:

- scikit-learn [11] – библиотека, содержащая в себе множество алгоритмов машинного обучения, в том числе и алгоритмы классификации, необходимые для данной задачи
- pandas [12] – библиотека, которая позволяет работать с файлами в формате csv

2.3. Алгоритмы классификации

В качестве алгоритмов классификации в данной работе используются следующие алгоритмы:

- метод k ближайших соседей (kNN) – простейший метрический классификатор, который основывается на оценивании сходства между объектами. Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки [13]. В библиотеке scikit-learn данный метод представлен классом KNeighborsClassifier;
- метод опорных векторов (SVM) – линейный классификатор, который основывается на том, что находит разделяющую гиперплоскость [14]. В библиотеке scikit-learn данный метод представлен классом svm.SVC;
- наивный байесовский классификатор (Naive Bayes) – классификатор, который основывается на принципе максимума апостериорной вероятности [15]. В библиотеке scikit-learn данный метод представлен классами GaussianNB, MultinomialNB, BernoulliNB;
- случайный лес (Random Forest) – алгоритм, который заключается в использовании ансамбля решающих деревьев.

2.4. Методы оценки классификаторов

Для оценки работы алгоритмов в данной работе используется понятие F-мера. Эта метрика состоит из двух других – точности (precision) и полноты (recall).

Достоверность (accuracy) — число, показывающее в какой степени исследование измеряет то, что должно быть измерено. То есть это отношение верно определенных классов к общему количеству определяемых объектов. При недостаточной достоверности метода появляется смещение или возникает систематическая ошибка, то есть если число объектов одного класса существенно больше числа объектов другого, то такая метрика будет иметь высокий результат на тех моделях классификации, которые плохо классифицирует небольшие классы, но при этом хорошо различают объекты из больших.

Для решения описанной проблемы и вводятся такие понятия, как точность (precision) и полнота (recall).

Точность (precision) – число, показывающее долю верно классифицированных объектов класса среди всех объектов, которые классификатор отнес к данному классу.

Полнота (recall) – число, показывающее долю верно классифицированных объектов класса среди всех объектов этого класса.

Для подсчета этих метрик существует 2 способа: составление таблицы контингентности (contingency table) и составление матрицы неточностей (confusion matrix).

В первом способе для каждого класса составляется своя таблица. Она имеет следующий вид: (таблица 7)

Таблица 7. Таблица контингентности

		Экспертная оценка	
		Положительная	**Отрицательная**
Оценка системы	**Положительная**	TP	FP
	Отрицательная	FN	TN

Данная таблица показывает сколько раз классификатор принял правильное, а сколько неправильное решение по объектам определенного класса.

- TP – истинно-положительное решение (объект из этого класса и классификатор смог это определить);
- TN – истинно-отрицательное решение (объект не из этого класса и классификатор его к нему не определил);
- FP – ложно-положительное решение (объект не из этого класса, но классификатор его к нему определил);
- FN – ложно-отрицательное решение (объект из этого класса, но классификатор его к нему не определил).

Тогда при использовании этого способа две предложенные метрики считаются следующим образом:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Второй способ, заключающийся в построении марицы неточности заключается в том, что строится таблица $N \times N$ (N – число классов), где столбцы обозначают действительный класс объектов, а строки – классы, которые возвращает классификатор. При классификации объекта

инкрементируется число, стоящее на пересечении строки класса, который вернул классификатор и столбца класса, к которому действительно относится объект. Тогда точность и полнота, исходя из формул выше, считаются следующим образом:

$$Precision_c = \frac{A_{c,c}}{\sum_{i=1}^{N} A_{c,i}}$$

$$Recall_c = \frac{A_{c,c}}{\sum_{i=1}^{N} A_{i,c}}$$

В этих формулах $A_{i,j}$ – элемент, стоящий на пересечении строки i -ой строки и j -ого столбца, c – рассматриваемый класс, а N – общее количество классов.

F-мера является средним гармоническим между точностью и полнотой. Она позволяет сбалансировать две этих метрики, так как нельзя улучшить одну без ухудшения другой. Формула для F-меры:

$$F = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}$$

В качестве метода оценки полученного классификатора используется алгоритм k-fold cross-validation при k=10. Каждый запуск алгоритма состоит из десяти итераций, на каждой из которых вся выборка делится на 10 одинаковых по размеру частей и 9/10 отводится на обучение модели, а 1/10 – для проверки. Для каждого метода алгоритм запускался 50 раз, чтобы минимизировать случайную ошибку, которая могла быть получена рандромизированным выбором тренировочных данных.

2.5. Методы сокращения размерности признакового пространства

Очень часто размерность пространства признаков слишком высока, и для небольших выборок это может оказаться проблемой. Для решения этой

проблемы существуют алгоритмы, которые позволяют сократить размерность признакового пространства и не уменьшить качество классификации, а иногда даже, напротив, увеличить его.

Одним из подходов в машинном обучении является так называемый алгоритм выбора признаков (feature selection), который выбирает какое-то подмножество признаков. Разделяют несколько видов данного подхода. В данной работе использовался wrapper-метод, а именно Recursive Feature Elimination (RFE). В библиотеке sickit-learn он представлен классом RFE.

Выводы по главе 2

В данной главе были изучены данные, по которым будут обучаться модели, рассмотрены и определены алгоритмы, по которым это будет выполняться, описаны методы уменьшения размерности пространства признаков, а также приведены метрики для оценивания предложенных классификаторов.

ГЛАВА 3. ПРИМЕНЕНИЕ ПРОСТЕЙШИХ МОДЕЛЕЙ

В данной главе описаны результаты исследований текущих решений по определению пола, возраста, рода занятий и уровня образования.

3.1. Метод k ближайших соседей

Метод k ближайших соседей – самый простой метод классификации. В качестве признаков на основе работ [3, 10, 17] из всего множества твитов берутся признаки, полученные при помощи латентного размещения Дирихле (LDA) [19]. Для каждого пользователя получен вектор из 50-ти чисел, каждое характеризующее какую-либо тему. Также были извлечены LIWC (Linguistic Inquiry and Word Count) [20, 21]. Для каждого пользователя был составлен вектор из 71 LIWC признака.

Для сети Instagram для каждого концепта изображения был посчитан его удельный вес – то есть доля картинок пользователя с этим концептом.

Для сети Foursquare для каждой категории места были посчитаны удельные веса среди всех чекинов данного пользователя.

В качестве k бралось 10% ото всех пользователей с соответствующей указанной демографической характеристикой.

Результаты работы с оценкой F-меры приведены в таблице 1.

Таблица 1. Результаты работы kNN

Характеристика	**Foursquare-признаки**	**LWIC**	**LDA**	**Instagram-признаки**
Пол	0,555	0,543	0,601	0,601
Возраст	0,163	0,354	0,301	0,217
Род деятельности	0,171	0,105	0,158	0,083
Образование	0,410	0,387	0,415	0,391

Можно заметить, что стандартные наборы признаков довольно плохо подходят классификации методом k ближайших соседей.

3.2. Наивный байесовский классификатор

Наивный Байесовский классификатор используется во многих работах, например в [10, 16]. В качестве признаков были взяты те же самые признаки, что и при методе kNN. Результаты работы алгоритма приведены в таблице 2.

Таблица 2. Результаты работы наивного байесовского классификатора

Характеристика	**Foursquare-признаки**	**LWIC**	**LDA**	**Instagram-признаки**
Пол	0,575	0,640	0,653	0,631
Возраст	0,185	0,392	0,343	0,233
Род деятельности	0,171	0,105	0,158	0,105
Образование	0,487	0,413	0,501	0,481

3.3. Метод опорных векторов

Метод опорных векторов – один из наиболее частовстречающихся классификатором, так как отличается своей простотой и мощностью. Его используют в работах [5, 10]. В качестве признаков были использованы те же самые, что и в байесовском классификаторе. Результаты приведены в таблице 3.

Таблица 3. Результаты работы классификатора на основе метода опорных векторов

Характеристика	**Foursquare-признаки**	**LWIC**	**LDA**	**Instagram-признаки**
Пол	0,581	0,590	0,595	0,581

Возраст	0,251	0,254	0,260	0,254
Род деятельности	0,173	0,155	0,138	0,115
Образование	0,427	0,434	0,551	0,475

Полученные результаты получились примерно такими же, как и при использовании наивного байесовского классификатора.

3.4. Random forest

Метод Random forest находит своё применение в работе [5]. В ней так же выделяются эвристические признаки текстовых данных. Среди них, например, среднее количество смайликов в сообщениях, частота использования хештегов, частота использования слов с повторяющимися буквами, которые могут указывать на возраст пользователя. Результаты применения метода классификатора на основе Random forest приведен в таблице 4.

Таблица 4. Результаты работы классификатора на основе Random forest

Характеристика	**Foursquare-признаки**	**LWIC**	**Эвристические признаки**	**LDA**	**Instagram-признаки**
Пол	0,649	0,716	0,685	0,788	0,784
Возраст	0,306	0,407	0,463	0,357	0,366
Род деятельности	0,393	0,216	0,353	0,255	0,299
Образование	0,587	0,514	0,521	0,498	0,563

Из таблицы можно заметить, что эвристические признаки играют очень существенную роль, так как все 4 характеристики удалось повысить после их введения. Так же существенно влияют Foursqare и Instagram признаки. Из этого можно сделать вывод, что в дальнейшем можно обдумать еще больше

эвристических признаков, а так же дополнить пространство признаков для упомянутых двух социальных сетей.

3.5. Ансамбль классификаторов

В приведенных выше способах учитывались только данные из какой-то одной социальной сети. Поэтому композиция результатов может существенно увеличить точность классификации. Единственная работа, в которой рассматривается объединение нескольких моделей – это работа [5]. В ней используется подход late-fusion.

В машинном обучении при нескольких источниках разделяют два метода: early fusion и late fusion. В early fusion подходе все типы признаков объединяются в один и отправляются на вход классификатору, как единое целое. В late fusion подходе же каждый тип признаков запускается на отдельном классификаторе, а после этого результаты, полученные каждым из них объединяются для получения наилучшего результата.

В работе [5] используются Random forest классификаторы, так как каждый такой классификатор на всех наборах признаков выдал самый высокий результат. Далее для объединения используется следующий подход. Для каждой демографической характеристики определяется оценка Score, которая считается по следующей формуле: $Score(l)=\sum_{i=0}^{k}\frac{P(l)_i\times d_i\times w_i\times l_i}{k}$, где k – количество классификаторов, $P(l)_i$ – достоверность модели, d_i – количество данных того типа, который обрабатывает классификатор, деленое на среднее число таких данных по всем пользователей (например, если есть 2 пользователя, у которых 1 и 3 твита, то их значения d_i соответственно равны 1/2 и 3/2), w_i – оценка модели (F-мера), l_i – мощность модели. Для нахождения оптимального и оправданного набора этих l_i используется алгоритм восхождения к вершине со случайным выбором (Stochastic Hill climbing with Random Restart), который используется для поиска максимальной оценки. Результаты работы ансамбля из

пяти Random Forest классификаторов, а именно составленных по LIWC, LDA, Instagram-признакам, Twitter-признакам и эвристическим признакам текста приведены в таблице 5.

Таблица 5. Результаты работы ансамбля RF-классификаторов

Характеристика	**F-мера**
Пол	0,878
Возраст	0,509
Род деятельности	0,498
Уровень образования	0,659

Как можно заметить, результаты, полученные при применении предложенного подхода на порядок выше, чем те, которые дают ранее представленные модели.

Выводы по главе 3

В данной главе было дано описание всех наиболее известных алгоритмов и описаны признаки, по которым можно строить модели. Результаты проделанной работы по реализации методов показывают, что наибольшее влияние оказывают правильно подобранные эвристические признаки, а также наиболее высокие результаты показывает классификатор, являющийся объединением других.

ГЛАВА 4. ВЫДЕЛЕНИЕ ПРИЗНАКОВ

Модели, рассмотренные ранее, напрямую зависят от составленного пространства признаков. В данной главе для каждой социальной сети выделяется ряд признаков, по которым в дальнейшем будут обучены модели.

4.1. Twitter

Как было описано в 3, для твитов были использованы алгоритмы LDA и LIWC, по которым получились свои наборы признаков. Так как они показывают хороший результат работы, то их выгодно оставить.

Как выяснилось в главе 3, на качество классификации существенно влияют различные эвристические характеристики текста, поэтому определим их.

В работах [3,5,6,23] используется среднее количество хештегов, которые пишет пользователь в сообщении. Это может указывать на пол, возраст или уровень образования [23]. Считается такая метрика для каждого пользователя по формуле $\frac{\sum_{i=0}^{N} htcount(t_i)}{N}$, где $htcount(t_i)$ – количество хештегов в твите.

В наше время в ход всё чаще и чаще вступают сленговые слова. Существует целый сайт, который содержит описание всех сленговых выражений и их описаний – urbandictionary.com [24]. Для составления модели был составлен список из самых популярных 40 сленговых слов с этого сайта и посчитано среднее количество твитов для каждого пользователя, в которых они применяются. Список слов [25]: bail, buck, skin of, creep, coucho potato, cram, crash, for real, asshole, going dutch, cold shoulder, hype, hyped, knock, pig out, tight, trash, uptight, wrap, zonked, pants, jonesing, aol, on point, arse, asshat, muff, meep, mosher, manboobs, reem, tee-bag, top, zit, netflix and chill, gooch, gilf. Данная

метрика для каждого пользователя была посчитана по формуле $\frac{\sum_{i=0}^{N} scount(t_i)}{N}$, где $scount(t_i)$ – количество слов из списка в твите.

С ростом социальных сетей, увеличивается количество сокращений, которые пользователи используют для скорости написания твитов. Например, в английском есть довольно длинное выражение «as soon as possible», которое переводится как «так скоро, насколько это возможно» и вместо него используется короткая форма – «ASAP». Так же существуют и цифровые замены, например, «too fast for you» сокращается до «2F4U». Такие признаки, как и использование сленговых выражений, могут указывать искомые демографические характеристики. Были извлечены самые популярные сокращения и акронимы [26] и к ним применен подход, описанный в работе [27]. Список сокращений: AEAP, ALAP, ASAP, ASL, B3, BFF, BM&Y, BRB, BRT, CTN, CUS, CWOT, CYT, EM?, EOD, F2F, IDC, IDK, ILU, ILY, IRL, IMU, JK, JC, JTLYK, KFY, L8R, NAZ, NP, NVM, OATUS, NIMBY, NP, OIC, PAL, PAW, PIR, RIP, RU, SO, TMI, XOXO, W8, WYWH. Список акронимов: 2F4U, AAMOF, ACK, AFAIK, AFAIR, BTT, BTW, C&P, CYS, EOBD, EOD, EOT, HF, HTH, IDK, MMW, N/A, OMG, POV, SPOC, TIA, TYT, WTH, YAM, TBA, TQ, NOOB, N00B, TTYL. Так же отдельно были посчитаны частоты употребления сокращений научных степеней (такие, как Ph.D, B.N, Th.D. и другие), а так же временные сокращения (am, pm, Mon, Dec и другие). Данные акронимы и сокращения были помещены в один вектор и для каждого элемента было посчитано общее количество раз, которое это сокращение или акроним были употреблены по отношению ко всему количеству твитов данного пользователя. То есть формула будет иметь следующий вид: $\frac{\sum_{i=0}^{N} acount(t_i)}{N}$, где $acount(t_i)$ – количество сокращений и акронимов в определенном твите.

Сейчас при составлении текстов микроблогов молодежь и люди, не вовлеченные в серьезные занятия, уделяют небольшое внимание знакам препинания. Но, напротив, старшие и те, у кого есть образование следят за этим [28]. Были выделены следующие знаки, доступные на английской клавиатуре:

- «!@#$%^&*()_+=-,./?\';:|»

Но при этом нужно разграничивать людей, которые применяют знаки препинания неосмысленно. Например, есть много твитов, в которых прослеживаются подобные конструкции: «some text.....», «some text !!!!!!!», «some text ??????». Поэтому для метрики, оценивающей использование знаков препинания были составлены следующие формулы:

- $\frac{\sum_{i=0}^{N} count(s, t_i)}{N}$ — среднее количество знаков в твитах;
- $\frac{\sum_{i=0}^{N} [count(s, t_i) > 0]}{N}$ — среднее количество твитов со зкаком;
- $\frac{\sum_{i=0}^{N} unique(t_i)}{N}$ — среднее число уникальных знаков в твитах;
- $\frac{\max_{i}(unique(t_i))}{S}$ — «вес» самого наполненного разными знаками твита.

Так же очень важным элементом, показывающим «серьезность» пользователя, является наличие в конце предложения точки, вопросительного или восклицательного знака. Для каждого пользователя была посчитана доля таких твитов по формуле $\frac{\sum_{i=0}^{N} endsymb(t_i)}{N}$, где $endsymb(t_i)$ – 1 или 0 в зависимости от того, заканчивается ли i-ый твит на один из трех знаков.

Кроме того, было посчитано среднее количество смайликов и слов, выражающих эмоции, для чего была использована система SentiWordNet. Использование подобных элементов переписки может указывать на возраст, пол и уровень образования.

Так же было посчитано среднее количество упоминаний других пользователей (обозначается как @имя_пользователя). Это тоже может указывать на любой из демографических признаков, так как каждому человеку присущ свой стиль использования таких особенностей социальных сетей.

Довольно очевидным является тот факт, что пользователи очень часто используют слова как-то связанные с их работой. Поэтому для каждого рода деятельности были выделены по 10 связанных слов. Для этого был использован сервис Semantic link [30]. Эти наборы представлены в таблице 1. Тем самым для каждого пользователя получился вектор размерности 200, где каждый элемент равен среднему количеству упоминай пользователем этого слова и его производных. Таблица соответствия приведена в приложении 1.

В итоге были выделены следующие наборы признаков:

- LIWC;
- LDA;
- Сокращения, сленговые слова, смайлики, частота знаков препинания эмоциональные слова, количество хештегов и упоминаний других пользователей были объединены в один набор;
- Упоминания слов, связанных с родом занятий.

4.2. Instagram

В качестве одного из признаков для сети Instagram был выбран средний вес концептов картинок. Данный признак представляет собой вектор из тысячи

элементов, в котором i-ый элемент посчитан по формуле $\frac{\sum_{j=0}^{N} w_{ij}}{N}$, где w_{ij} – вес i-го признака в j-ой фотографии.

В предоставленных данных для каждой фотографии есть время. Это очень важная характеристика, так как разные люди делают снимки определенных вещей в разное время. Например, люди, работающие в ресторанах будут фотографировать связанные с ними вещи в течение всего дня, а люди, которые приходят туда в качестве посетителей, будут делать подобные снимки вечером. Поэтому для каждого концепта была отобрана половина всех фотографий каждого пользователя, в которой содержатся снимки с максимальными весами этого концепта. После этого было посчитано среднее время суток этих снимков. Чтобы описать как это было сделано, сначала опишем проблему. Допустим, у нас есть 2 времени: 11 вечера (то есть 23 часа) и 3 ночи. Очевидно, что среднее время – 2 часа ночи, но если посчитать просто среднее, то получим 13, то есть час дня, что не имеет никакого смысла. Для этого изначально были посчитаны 2 величины: среднее «дневное» время и среднее «вечернее», а именно с 0 до 12 и с 12 до 24. Путь это t_d и t_n соответственно. После этого если $(24-t_n)+t_d>12$, то среднее время – $\frac{(t_n+t_d)}{2}$, иначе – $\frac{(t_d+(24-t_n))}{2}$.

Каждый пост в инстаграме содержит подпись. При правильном выделении признаков оттуда, они вместе с концептами фотографий могут дать мощный способ определения демографических признаков. Мощным способом для обработки текстовых данных, как показала глава 2 оказался LDA, поэтому используется алгоритм, который был применен для текста твитов и им дополняется массив средних весов концептов фотографий 50-ю весами топиков, полученных при LDA.

В итоге были выделены следующие наборы признаков:

• Средний вес концептов фотографий вместе с весами топиков, полученных при LDA;

• Для каждого концепта среднее время суток, в которые определенный концепт имеет максимальный вес.

4.3. Foursquare

В качестве одной из метрик из сети Foursquare была взята частота посещений каждой категории. Каждый чекин пользователя представлен списком категорий, которые характеризует это место. Всего категорий 592, поэтому каждый пользователь был представлен в виде вектора из 592-ух элементов, каждый из которых вычисляется по формуле $\frac{\sum_{j=0}^{N}[c_{ij}]}{N}$, где $[c_{ij}]$ – это 0 или 1 в зависимости от того, присутствует ли i-ая категория в j-ом чекине.

Вторым признаком, подобно тому, что описан в разделе 4.2, было взято среднее время посещения определенной категории. Это было сделано таким же методом с делением на «дневное» и «вечернее».

Каждый чекин сопроваждается подписью. Так как они тесно связаны с самим местом, то их можно совместить с метрикой, характеризующей непостредственно место. Для этого был применен LDA, поэтому алгоритм из главы 2 для выделения топиков текстов твитов был применен для описаний к чекинам и совмещен и вектором частот посещений мест;

В итоге были выделены следующие наборы признаков:

• Средняя частота посещения каждой категории мест вместе с весами топиков, полученных при LDA;

• Среднее время суток посещения каждой категории мест;

Выводы по главе 4

В этой главе были представлены признаки, которые помогут определять пол, возраст, уровень образования и род деятельности. Многие из них, например, анализ времени постов и чекинов, некоторые признаки, связанные с пунктуацией, а так же сокращения и модные слова, не были представлены в существующих работах, а были получены путем анализа предметной области.

ГЛАВА 5. КЛАССИФИКАТОР НА ОСНОВЕ ПОЛУЧЕННЫХ ПРИЗНАКОВ

В данной главе описывается итоговый классификатор для определения пола, возраста, рода деятлеьности и уровня образования. Так приведены результаты его работы.

5.1. Выбор классификатора

В качестве классификатора для восстановления демографических признаков был использован классификатор, описанный в главе 3. Для этого были составлены составные классификаторы. Как было показано в главе 3, самыми лучшими классификаторами оказались классификаторы, которые основываются на Random Forest. Для каждого набора признаков, описанных в главе 4 была применена k-fold кросс-валидация при k=10 для поиска оптимального количества деревьев в Random Forest. Результаты приведены в таблице 1.

Таблица 1. Количество деревьев для Random Forest

Набор признаков	**Количество деревьев**
LIWC (Twitter)	120
LDA (Twitter)	70
Общие эвристические признаки (Twitter)	75
Связанные с работой слова (Twitter)	70
Концепты фото + LDA (Instagram)	170
Средние времена (Instagram)	90
Категории мест + LDA (Foursquare)	160
Средние времена (Foursquare)	95

Результаты работы RF-классификаторов в качестве оценки F-меры на разных наборах признаков представлены в таблице 2.

Таблица 2. Оценки определения характеристик итоговыми классификаторами

Способ	**Пол**	**Возраст**	**Род занятий**	**Уровень образования**
LIWC (Twitter)	0,716	0,407	0,353	0,415
LDA (Twitter)	0,788	0,357	0,321	0,398
Общие эвристические признаки (Twitter)	0,773	0,499	0,356	0,681
Связанные с работой слова (Twitter)	0,770	0,461	0,703	0,715
Концепты фото + LDA (Instagram)	0,791	0,463	0,693	0,651
Средние времена (Instagram)	0,650	0,403	0,617	0,591
Категории мест + LDA (Foursquare)	0,692	0,398	0,659	0,585
Средние времена (Foursquare)	0,671	0,458	0,631	0,573

На основе приведенных выше классификаторов, применяется метод, описанный в главве 3 где для каждого классификатора, и ищется его «мощность», а в качестве результата каждой демографической характеристики по всем берется тот ответ, у которого больше суммарная «мощность».

Ниже приведена оценка F-меры для итогового классификатора:

- Пол – 0,895
- Возраст – 0,553
- Род занятий – 0,727
- Уровень образования – 0,829

5.2. Значимость признаков

Для получившихся признаков была измерена их значимость относительно взаимной информации. Взаимная информация (mutual information) — функция двух случайных величин, которая описывает количество информации, содержащееся в одной случайной величине относительно второй:

$I(X;Y)=H(X)-H(X|Y)$ (здесь H – энтропия)

В приложении 2 для каждого набора признаков представлены лучшие признаки (те, у которых взаимная иформация больше).

5.3. Сокращение размерности признакового пространства

В качестве алгоритма для уменьшения пространства признаков был взят, как и описывалось в главе 2.5, алгоритм RFE. Он был запущен для каждого классификатора и для каждой демографической характеристики. Результаты работы представлены в таблице 3. В ней к.п. обозначает количество признаков, а F – F-мера.

Таблица 3. Результаты сокращения размерности

Способ	**Пол**		**Возраст**		**Род занятий**		**Уровень образования**	
	к.п.	**F**	**к.п.**	**F**	**к.п.**	**F**	**к.п.**	**F**
LIWC (Twitter)	45	0,719	51	0,408	63	0,351	49	0,417
LDA (Twitter)	31	0,795	39	0,359	48	0,322	30	0,403
Общие эвристические признаки (Twitter)	53	0,778	61	0,503	70	0,358	55	0,683
Связанные с работой слова (Twitter)	83	0,773	103	0,471	171	0,711	141	0,723
Концепты фото + LDA (Instagram)	613	0,792	659	0,469	370	0,695	639	0,658

Средние времена (Instagram)	481	0,654	517	0,408	189	0,619	467	0,592
Категории мест + LDA (Foursquare)	383	0,697	461	0,400	517	0,663	505	0,589
Средние времена (Foursquare)	298	0,675	400	0,461	483	0,637	353	0,577
Композиция классификаторов	0,898		0,561		0,735		0,832	

Как видно из таблицы, сокращение признакового пространства не только позволило уменьшить размерность, но и увеличить результат определения демографических характеристик.

5.4. Сравнение с аналогами

Как упоминалось раньше, существует только одна работа, работающая с данными из многих источников. По определению же рода занятий и уровня образования были выделены 2, где показаны лучшие результаты. Предложенный в работе классификатор показывает большую эффективность для всех признаков. Они приведены в таблице 4.

Таблица 4. Сравнение с аналогами

Характеристика	**Существующее решение**	**Предложенное решение**
Пол	0,878	0,898
Возраст	0,509	0,561
Род занятий	0,527	0,735
Уровень образования	0,743	0,829

Выводы по главе 5

В данной главе описан итоговый классификатор и представлены результаты его работы. Кроме того, был использован алгоритм уменьшения размерности признакового пространства, что помогло улучшить результаты.

Также было выполнено сравнение с существующим решением, которое показало, что предложенный в работе метод имеет большую эффективность в решении поставленной задачи.

ЗАКЛЮЧЕНИЕ

Во-первых, был произведен полный обзор предметной области, в том числе поиск существующих решений и реализация наиболее эффективных методов, предложенных в этих работах.

Во-вторых, был произведен анализ имеющихся данных, определены алгоритмы построения моделей.

В-третьих, был выполнен тщательный отбор признаков социальной активности пользователей в трех социальных сетях и выделены наиболее подходящие для определения демографических характеристик.

В-четвертых, был описан итоговый классификатор, который показал свою эффективность по сравнению с текущими решениями. Кроме того, были произведены запуски алгоритмов, позволяющих уменьшить пространство признаков.

Таким образом, были выполнены все поставленные задачи и достигнута цель работы. Как её итог, был получен классификатор, который эффективно решает поставленную задачу.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Балуев Д.Г. Политическая роль социальных медиа как поле научного исследования // Образовательные технологии и общество (EDUCATION TECHNOLOGY & SOCIETY). Т. 16(2). 2013. С. 604-616.
2. M. Duggan and J. Brenner. The demographics of social media users. 2013.
3. R. Ottoni, D. Las Casas, J. P. Pesce, W. Meira Jr, C. Wilson, A. Mislove, and V. Almeida. Of pins and tweets: Investigating how users behave across image-and text-based social networks. 2014.
4. URL: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html (Дата обращения: 15.04.2017)
5. Daniel Preotiuc-Pietro, Vasileios Lampos and Nikolaos Aletras.An analysis of the user occupational class through Twitter content. 2013.
6. Jiwei Li, Alan Ritter, Eduard Hovy. Weakly Supervised User Profile Extraction from Twitter. 2015.
7. Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati. What We Instagram: A First Analysis of Instagram Photo Content and User Types. 2014.
8. A Noulas, S Scellato, C Mascolo, M Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. 2014.
9. T. Pontes, M. Vasconcelos, J. Almeida. We know where you live: privacy characterization of foursquare behavior. 2012.
10. Aleksandr Farseev, Liqiang Nie, Mohammad Akbari and Tat-Seng Chua. Harvesting Multiple Sources for User Profile Learning: a Big Data Study. 2015.
11. URL: http://scikit-learn.org/ (Дата обращения: 01.03.2017)
12. URL: http://pandas.pydata.org/ (Дата обращения: 12.12.2016)

13. URL: http://www.machinelearning.ru/wiki/index.php?title=KNN (Дата обращения: 08.01.2017)

14. URL: http://www.machinelearning.ru/wiki/index.php?title=SVM (Дата обращения: 08.01.2017)

15. URL: http://www.machinelearning.ru/wiki/index.phptitle=Байесовский_классификатор (Дата обращения: 08.01.2017)

16. Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins and Mahesan Niranjan. Trendminer: An Architecture for Real Time Analysis of Social Media Text. 2012.

17. Deitrick W., Miller Z., Valyou B., Dickinsoosn B., Munson T., Hu W. Gender Identification on Twitter Using the Modified Balanced Winnow // Communications and Network. 2012. Vol. 4. P. 189–195.

18. James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. The Development and Psychometric Properties of LIWC. 2015.

19. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. 2003.

20. Y.R. Tausczik, J.W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. 2010.

21. Deitrick W., Miller Z., Valyou B., Dickinsoosn B., Munson T., Hu W. Gender Identification on Twitter Using the Modified Balanced Winnow

22. URL: http://dpeaflcio.org/programs-publications/issue-fact-sheets/the-professional-and-technical-workforce/ (Дата обращения: 06.04.2017)

23. Schwartz A.H., Eichstaedt J.C., Kern M.L., Dziurzynski L., Ramones S.M., Agrawal M., Shah A., Kosinski M., Stillwll D., Seligman M.E.P., Ungar L.H. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. 2013.

24. URL: http://urbandictionary.com (Дата обращения: 21.04.2017)

25. A. Peckham. Urban dictionary: Fularious street slang defined. 2009.

26. URL: http://www.smart-words.org/abbreviations/text.html (Дата обращения: 03.04.2017)

27. S.Yeates. Automatic Extraction of Acronyms from Text. 1999. (Дата обращения: 03.04.2017)

28. Lin J. Automatic Author Profiling of Online Chat Logs // Master's thesis. Naval Postgraduate School. Monterey. 2007. P. 265.

29. S. Kiritchenko, X. Zhu and S. M. Mohammad. "Sentiment Analysis of Short Informal Texts. 2014.

30. URL: http://semantic-link.com/ (Дата обращения: 04.04.2017)

31. Павлова О.Н., Казин Ф.А., Бутаков Н.А. "Сравнительный анализ профильности трудоустройства выпускников ведущих вузов России на основе данных социальной сети «Вконтакте»" // Журнал "Университетское управление: практика и анализ" №3, 2017.

ПРИЛОЖЕНИЯ

Приложение 1. Соответствие рода деятельности и связанных слов

Род занятий	Связанные слова
protective service	Constables, policing, constabulary, brutality, precept, policemen, uniformed, demostrators, lapd, gendarmerie
education, training, and library	Accridit, vocational, imparting, curricula, tertiary, mandatory, schooling, educators, school, university
sales and related	Encourage, copy, illegal, sell, sale, copies, retaires, resale, limit, ticket, marketing
farming, fishing and forestry	Farm, crop, forestry, livestock, sheep, cow, fallow, wheat, milk, meat
office and administrative support	Zip, money, box, table, elective, post, supervising, auditor, desk, audit
healthcare support	Neurology, gynecology, obstetrics, urology, radiology, phychology, pergant ,veterinary, pathalogy, internship
production	Biofuel, tonne, ethanol, obie, alumnia, biomass, airframes, barrel, oil, gas
construction and extraction	Solvent, shale, bitumen, aqueous, situ, hydracarbon, extract, solubility, gravel, brine
community and social service	Stratification, mores, sociologist,

	networking, societal, conservatism, taboo, cognition, liberalism, worker
business and financial operations	Acumen, venture, deal, money, buck, enterprise, sell, transaction, bank, cost
transportation and material moving	Car, travel, transport, export, import, hub, modal, rail, fuels, train, airplane
archetecture and engineering	Urbanism, architecture, museum, calc, mips, cathedral, gotic, draft, modern
personal care and service	Inpatient, hospice, clinic, patient, hourly, ridership, commuter, ambulance, customer, passenger
management	Lifecycle, asset, workflow, automate, outsourcing, analytics, outscore, consult, resource
healthcare practitioners and technical	Fouls, vocational, technique, innovation, managerial, diploma, unicode, knockout, mastery, metal
building and grounds cleaning and maintenance	Skyscraper, deco, facade, brick, renovate, remodel, cornice, edifice, architect, roof
legal	Law, rule, lagal, brief, court, police, judge, juridical, tender, suit
food preparation and service related	Bistro, restaurant, buffet, dinner, pizza, culinary, cafe, fries, motel, vegetarian, lunch
arts, design, entertainment, sports, and media	Deco, expressionism, cubism, museum, curator, exhibite, concept, picasso, paint,

	pencil, color
computer and mathematical	Mainframe, acm, hacker, code, program, turing, net, laptop, processor, calc

Приложение 2. Лучшие признаки классификаторов

Общие эвристические признаки (Twitter)		
Пол	Частота использования сокращений	0,531
	Частота применения эмоциональных слов	0,417
	Частота появления хештегов	0,401
	Употребление смайликов	0,351
	Использование конечных знаков препинания	0,317
Возраст	Сленговые слова	0,635
	Сокращения	0,517
	Употребление смайликов	0,413
	Частота повторяющихся знаков препинания	0,381

	Частота появления хештегов	0,328
Уровень образования	Сокращения научных степеней	0,631
	Остальные сокращения	0,583
	Хештеги	0,478
	Упоминания пользователей	0,391
	Частота одиночных знаков препинания	0,329
Род деятельности	Аббревиатуры	0,643
	Сокращения	0,571
	Упоминания пользователей	0,481
	Хештеги	0,363
	Сленговые слова	0,311
Связанные с работой слова (Twitter)		
Пол	art	0,516
	technik	0,440
	maid	0,481
	hospital	0,363
	children	0,215
Возраст	museum	0,501

	deal	0,403
	money	0,351
	travel	0,298
	restraunt	0,243
Уровень образования	university	0,631
	school	0,582
	teacher	0,403
	lecture	0,398
	mark	0,304
Род деятельности	money	0,535
	deal	0,382
	museum	0,301
	lawyer	0,291
	office	0,263