


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

Факультет Информационных Технологий и Программирования
 Направление (специальность) Прикладная математика и информатика
 Квалификация (степень) Бакалавр прикладной математики и информатики
 Кафедра Компьютерных технологий Группа 4538

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

Определение хромотипа человека по его активности в социальной сети

Автор квалификационной работы Баев В. А.  (подпись)

Руководитель Фильченков А. А. (подпись)

а) По экономике и организации _____ (подпись)
производства

б) По безопасности жизнедеятельности и экологии _____ (подпись)

В) _____ (подпись)

Зав. кафедрой _____ *Васильев В.Н.* _____ (подпись)

« » 2015 г.

Санкт-Петербург, 2015 г.

Квалификационная работа выполнена с оценкой _____

Дата защиты « _____ » _____ 2015 г.

Секретарь ГАК _____

Листов хранения _____

Чертежей хранения _____

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

Факультет Информационных Технологий и Программирования
Кафедра Компьютерных технологий Группа 4538
Направление (специальность) Прикладная математика и информатика
Квалификация (степень) Бакалавр прикладной математики и информатики

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Студент Баев В. А.
Руководитель Фильченков А. А., кандидат физико-математических наук,
ведущий инженер кафедры ИС

1. Наименование темы Определение хромотипа человека по его активности в
социальной сети

2. Срок сдачи студентом законченной работы 28 мая 2015 г.

3. Техническое задание и исходные данные к работе

Исследовать системы определения хромотипа человека. Провести исследования и
разработать методы определения хромотипа человека, основываясь на его
активности в социальной сети. Разработать программную реализацию
продукта, определяющего хромотип пользователя социальной сети.
Проанализировать результаты классификации и сделать выводы о точности
определения хромотипа.

4. Содержание выпускной работы (перечень подлежащих разработке вопросов)

1. Обзор предметной области. 2. Разработка методов решения задачи. 3. Анализ
полученных результатов.

5. Перечень графического материала (с указанием обязательного материала)

6. Исходные материалы и пособия

1. Rui Guo, Hongzhi Wang, Lucheng Zhong, Jianzhong Li, Hong Gao. Harbinger: An Analyzing and Predicting System for Online Social Network Users' Behavior. 2013.
2. Giannotti F, Cortesi F, Sebastiani T, Ottaviano S. Circadian preference, sleep and daytime behaviour in adolescence. 2001.
3. Hidalgo MP1, Caumo W, Posser M, Coccaro SB, Camozzato AL, Chaves ML. Relationship between depressive mood and chronotype in healthy subjects. 2009.

7. Консультанты по работе с указанием относящихся к ним разделов работы

Раздел	Консультант	Подпись, дата	
		Задание выдал	Задание принял
Экономика и организация производства			
Технология приборостроения			
Безопасность жизнедеятельности и экология			

КАЛЕНДАРНЫЙ ПЛАН

№№ п/п	Наименование этапов выпускной квалификационной работы	Срок выполнения этапов работы	Примечание
1	Ознакомление с предметной областью	15.12.2014	
2	Разработка методов решения задачи	11.03.2015	
3	Программная реализация	15.05.2015	
4	Анализ результатов	15.05.2015	
5	Реализация конечного продукта	25.05.2015	
6	Написание пояснительной записки	28.05.2015	

8. Дата выдачи задания _____ 1 сентября 2014 г.

Руководитель

 _____

Задание принял к исполнению

 _____

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ

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

АННОТАЦИЯ
ПО ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

Студента _____ Баева В.А.
(Фамилия, И., О.)

Факультет _____ Информационных технологий и программирования

Кафедра _____ Компьютерных технологий _____ Группа _____ 4538

Направление (специальность) _____ Прикладная математика и информатика

Квалификация (степень) _____ Бакалавр прикладной математики и информатики

Наименование темы: Определение хронотипа человека по его активности в социальной сети

Руководитель Фильченков А.А., к. ф.-м. н., ведущий инженер каф. ИС
(Фамилия, И., О., ученое звание, степень)

Консультант _____
(Фамилия, И., О., ученое звание, степень)

**КРАТКОЕ СОДЕРЖАНИЕ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
И ОСНОВНЫЕ ВЫВОДЫ**

объем _____ 48 _____ стр., графический материал _____ 0 _____ стр., библиография _____ 58 _____ наим.

- Направление и задача исследований

Целью настоящей работы является определение хронотипа человека, основываясь на его активности в социальной сети.

- Проектная или исследовательская часть (с указанием основных методов исследований, расчетов и результатов)

Определение хронотипа пользователя социальной сети производится с помощью следующих методов: выявление временных промежутков, в течение которых пользователь совершал наибольшее число действий; анализ содержания публичных текстовых сообщений пользователя.

- Экономическая часть (какие использованы методики, экономическая эффективность результатов)

Не рассматривалась

- Характеристика вопросов экологии, техники безопасности и др.

Не рассматривалось

- Является ли работа продолжением курсовых проектов (работ), есть ли публикации

Данная работа не является продолжением курсовых проектов, публикаций на её основе нет.

Практическая ценность работы. Рекомендации по внедрению

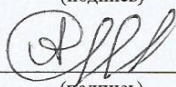
Информация о хронолите человека может быть использована при производстве таргетированной рекламы, так как хронолит обуславливает некоторые увлечения человека и стиль его жизни. Кроме того, хронолит можно принимать во внимание при приеме на работу для составления рабочего графика.

Выпускник



(подпись)

Руководитель



(подпись)

“ ”

2015 г.

Оглавление

Введение	6
Глава 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ.....	8
1.1 Хронотип.....	8
1.2 Машинное обучение	9
1.3 Задача разделения смеси распределений	11
1.4 Социальная сеть Twitter	13
1.5 Постановка задачи	14
1.6 Выводы по главе	15
Глава 2. РАЗРАБОТКА МЕТОДОВ ОПРЕДЕЛЕНИЯ ХРОНОТИПА	16
2.1 Анализ временных интервалов активности	16
2.2 Выделение признаков для анализа текстовых сообщений.....	17
2.2.1 Эмоциональная окраска	17
2.2.2 Склонность к депрессии	18
2.2.3 Склонность к употреблению кофеина	19
2.2.4 Склонность к набору лишнего веса	19
2.2.5 Проблемы со сном	20
2.2.6 Склонность к курению и употреблению алкоголя.....	20
2.3 Выводы по главе	20
Глава 3. СБОР ДАННЫХ.....	22
3.1 Сбор информации о хронотипе.....	22
3.2 Сбор информации о пользователе Twitter.....	25
3.2.1 Моделирование пиков суточной активности	25
3.2.2 Связь хронотипа пользователя и его временной активности.....	27

3.2.3 Формирование обучающей выборки	27
3.3 Выводы по главе	28
Глава 4. ПОСТРОЕНИЕ КЛАССИФИКАТОРА	29
4.1 Алгоритмы классификации.....	29
4.2 Результаты работы классификаторов	30
4.3 Анализ результатов	33
4.4 Выводы по главе	33
Заключение	34
Список литературы	35
Приложения	42
Приложение 1. Тесты для определения хронотипа	42
Тест для прохождения лично.....	42
Тест для прохождения через интернет.....	43
Приложение 2. Строки, формирующие атрибуты классификатора	47

Введение

В последнее время социальные сети становятся все более популярным средством общения. В связи с этим многие исследования направлены на анализ общедоступной информации, предоставляемой социальными сетями [1,2,3]. Анализируя поведение пользователя в социальной сети, можно сделать выводы о различных социальных, психологических, медицинских показателях человека [4,5]. Удобство анализа такого рода информации заключается в том, что данные могут быть извлечены автоматически, без привлечения затрат на ручной сбор информации, что позволяет собирать и обрабатывать существенно большие объемы данных, чем были доступны ранее.

Одним из таких показателей, которым характеризуется человек, является хронотип. Обычно под хронотипом понимают категорию людей в зависимости от времени в течение суток, когда они наиболее активны. Ученые выделяют два хронотипа: «совы» и «жаворонки» [6]. Совы наиболее активны в вечернее время суток, а жаворонки более активны утром, в ранние часы. Хронотип связан с некоторыми увлечениями человека, с его стилем жизни, поэтому знание хронотипа может быть полезно в следующих случаях: при разработке таргетированной рекламы, при планировании оптимального рабочего графика для работников, при принятии решения о приеме человека на определенную должность [7]. Также знания о своем хронотипе помогут подросткам выбрать будущую профессию.

Целью настоящей работы является разработка и реализация методов определения хронотипа человека, основываясь на его активности в социальной сети. Определение хронотипа будет производиться на основе анализа временных интервалов активности в течение суток и текста общедоступных публичных сообщений пользователя. Для этого требуется решить следующие задачи:

- сбор данных для обучающей выборки,

- установление связи между хронотипом человека и его активностью в социальной сети,
- разметка данных с помощью модели смеси двух гауссиан,
- выделение гипотез о связи хронотипа и прочих социально-медицинских показателей,
- разработка способа тестирования предложенных гипотез у пользователей социальной сети,
- обучение классификатора, анализирующего текст сообщений для тестирования гипотез.

Глава 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

1.1 Хронотип

Хронотип человека — это характеристика, отражающая значения некоторых физических показателей человека, в зависимости от времени суток [8]. Этими показателями могут быть температура тела, уровень определенных гормонов в организме, частота сердечных сокращений, способность к умственной или физической деятельности и многие другие. Обычно это понятие рассматривается в более узком смысле: под хронотипом понимают классификацию людей, в зависимости от времени суток, когда они наиболее активны. Ученые выделяют два основных хронотипа: «совы» и «жаворонки». Основная особенность «сов» заключается в том, что они поздно просыпаются и поздно ложатся спать. «Жаворонки» же, наоборот, предпочитают раннее пробуждение и не любят засиживаться до поздней ночи. Таким образом, «совы» наиболее работоспособны во второй половине дня, а «жаворонки» — в первой [9].

Существуют различные тесты, позволяющие определить хронотип человека. Один из таких тестов разработан в Мюнхенском университете Людвиг-Максимилиана (МСТQ) [10]. Данный тест наиболее популярен среди существующих, многие проекты используют его для исследований, собрана база из более чем двадцати пяти тысяч человек. Кроме того, существует альтернативный тест Automated Morningness-Eveningness Questionnaire (AutoMEQ), который также используется в медицинских и психологических исследованиях [11]. Долгое время считалось, что причисление человека к одному из этих типов происходит на основании привычки или особенностей в распорядке дня. Недавно выяснилось, что хронотип имеет генетическую природу [8], т.е. люди предрасположены к какому-то хронотипу, но зачастую им приходится менять его ввиду сложившихся обстоятельств.

1.2 Машинное обучение

Машинное обучение — раздел искусственного интеллекта, изучающий алгоритмы, способные обучаться [12].

Традиционно задача машинного обучения ставится следующим образом: дано множество объектов и множество ответов. Существует зависимость между объектами и ответами, которая заранее неизвестна. Известно конечное множество пар из объекта и ответа, которое называется обучающей выборкой. Требуется на основе обучающей выборки восстановить зависимость между объектами и ответами, то есть предъявить алгоритм, который для любого объекта будет выдавать ответ. Для измерения точности ответов вводится определенный функционал качества.

Алгоритмы машинного обучения можно разделить на два больших класса: обучение с учителем и обучение без учителя. Выделенные классы отличаются тем, что при обучении с учителем для обучающей выборки известны объекты и ответы, а при обучении без учителя — только объекты.

Классификация — раздел машинного обучения, решающий следующую задачу. Задано конечное множество объектов, для которых заранее известны метки классов. Требуется предъявить алгоритм, который будет классифицировать любой объект, принадлежащий исходному множеству, с определенной точностью [13].

Обучение модели происходит следующим образом: выделяется обучающая выборка, для которой известны классы, к которым принадлежит каждый объект. Алгоритм машинного обучения запускается на данной выборке и подстраивает свои параметры определенным образом под выборку. После этого алгоритм может работать на других входящих данных, для которых заранее неизвестны классы объектов, и классифицировать их.

Для **оценки качества модели** традиционно используют перечисленные ниже показатели. Данные показатели вычисляются для каждого класса выборки и строятся на основе базовых величин:

- истинно-положительные результаты классификатора (True Positives)
- ложно-положительные результаты классификатора (False Positives)
- ложно-отрицательные результаты классификатора (False Negatives)
- истинно-отрицательные результаты классификатора (True Negatives)

Точность — величина, показывающая, какая часть объектов, найденных классификатором, действительно принадлежит данному классу. Точность определяется следующей формулой:

$$P(\text{precision}) = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}.$$

Полнота — величина, показывающая, какая часть объектов данного класса найдена классификатором. Полнота определяется следующей формулой:

$$R(\text{recall}) = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.$$

На практике удобно оценивать качество работы алгоритма с помощью **матрицы неточностей**, которая состоит из N строк и N столбцов, где N — количество классов для классификации. Столбцы матрицы отвечают за истинный класс, а строки матрицы отвечают за класс, предсказанный алгоритмом. Элемент матрицы $A_{i,j}$ показывает, сколько объектов класса j алгоритм классифицировал как объект класса i . Диагональные элементы матрицы показывают число верно классифицированных объектов. С помощью матрицы неточностей можно вычислять точность и полноту для класса c следующим образом:

$$P(\text{precision})_c = \frac{A_{c,c}}{\sum_{i=1}^N A_{c,i}}.$$

$$R(\text{recall})_c = \frac{A_{c,c}}{\sum_{i=1}^N A_{i,c}}.$$

Итоговая точность алгоритма вычисляется как среднее арифметическое точности для всех классов. Полнота вычисляется аналогично.

Введем показатель, совмещающий в себе показатели точности и полноты. **F-мера** — величина, определяемая как взвешенное среднее гармоническое точности и полноты:

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}, \alpha \in [0, 1].$$

F₁-мера — величина, являющаяся средним гармоническим точности и полноты. F₁-мера определяется следующей формулой:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}.$$

Для оценки статистической связи между случайными величинами рассмотрим **коэффициент корреляции**. Коэффициент корреляции Пирсона случайных величин X и Y вычисляется по формуле:

$$r_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y},$$

где cov_{XY} — ковариация случайных величин X и Y , σ_X — среднеквадратичное отклонение случайной величины X .

1.3 Задача разделения смеси распределений

В данном разделе будет рассмотрена задача разделения смеси вероятностных распределений.

Нормальное распределение (распределение Гаусса) — вероятностное распределение, функция плотности вероятности которого является функцией Гаусса:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x - \mu)^2}{-2\sigma^2}\right),$$

где μ — математическое ожидание распределения, а σ — среднее квадратичное отклонение распределения [14].

Стандартное нормальное распределение — частный случай нормального распределения, где $\mu = 0$, $\sigma = 1$ (рис. 1).

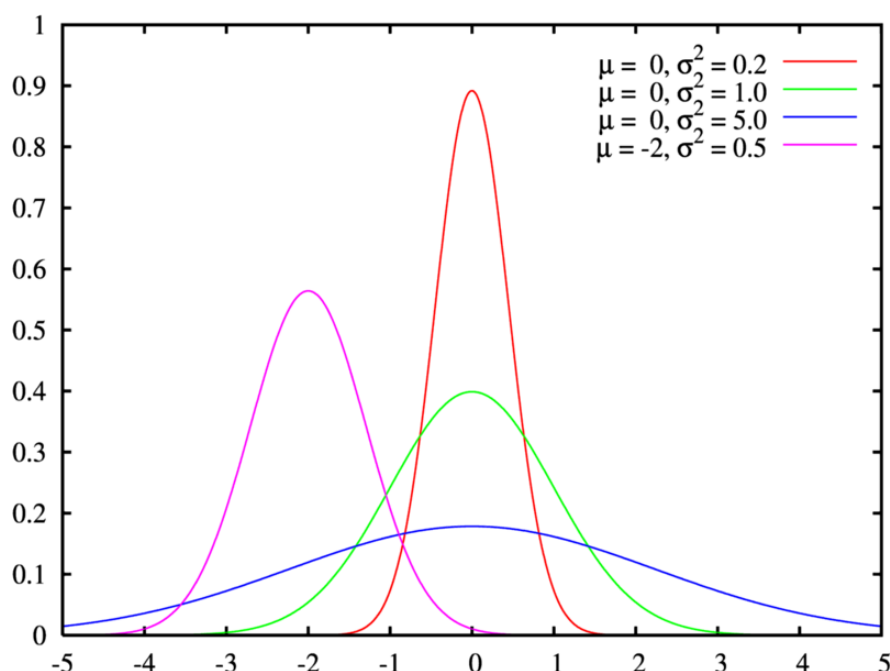


Рисунок 1. Плотность вероятности нормального распределения.

ЕМ-алгоритм — алгоритм, который применяется для определения оценок максимального правдоподобия некоторых параметров модели, когда модель зависит от скрытых параметров [15]. Алгоритм итерационный, каждая его итерация представляет собой два определенных шага. На первом шаге (Е-шаг, expectation) определяется значение функции правдоподобия, причем скрытые переменные являются наблюдаемыми на данном шаге. На втором шаге (М-шаг, maximization) определяется оценка максимального правдо-

подобия, увеличивая ожидаемое правдоподобие, вычисляемое на первом шаге. Далее Е-шаг использует полученное значение на следующей итерации. Итерации алгоритма продолжаются до достижения требуемой точности. Зачастую ЕМ-алгоритм применяется для разделения смеси нормальных распределений [16].

1.4 Социальная сеть Twitter

Twitter — социальная сеть для обмена короткими сообщениями. Часто подобный формат взаимодействия называют «микроблоггингом». Проект был создан в 2006 году в Сан-Франциско, на текущий момент имеет 140 миллионов активных пользователей и 500 миллионов зарегистрированных пользователей [17]. Особенностью данного сервиса является ограничение на длину текстовых сообщений в 140 символов. Кроме того, Twitter особенно интересен для научных исследований тем, что сообщения пользователя доступны для чтения любым другим пользователям (за исключением приватных аккаунтов), вследствие чего можно собрать обширную базу сообщений. Сервис предусматривает элемент социального взаимодействия «подписка». Пользователь может подписаться на обновления любого пользователя. На рисунке 2 приведен пример ленты сообщений пользователя.

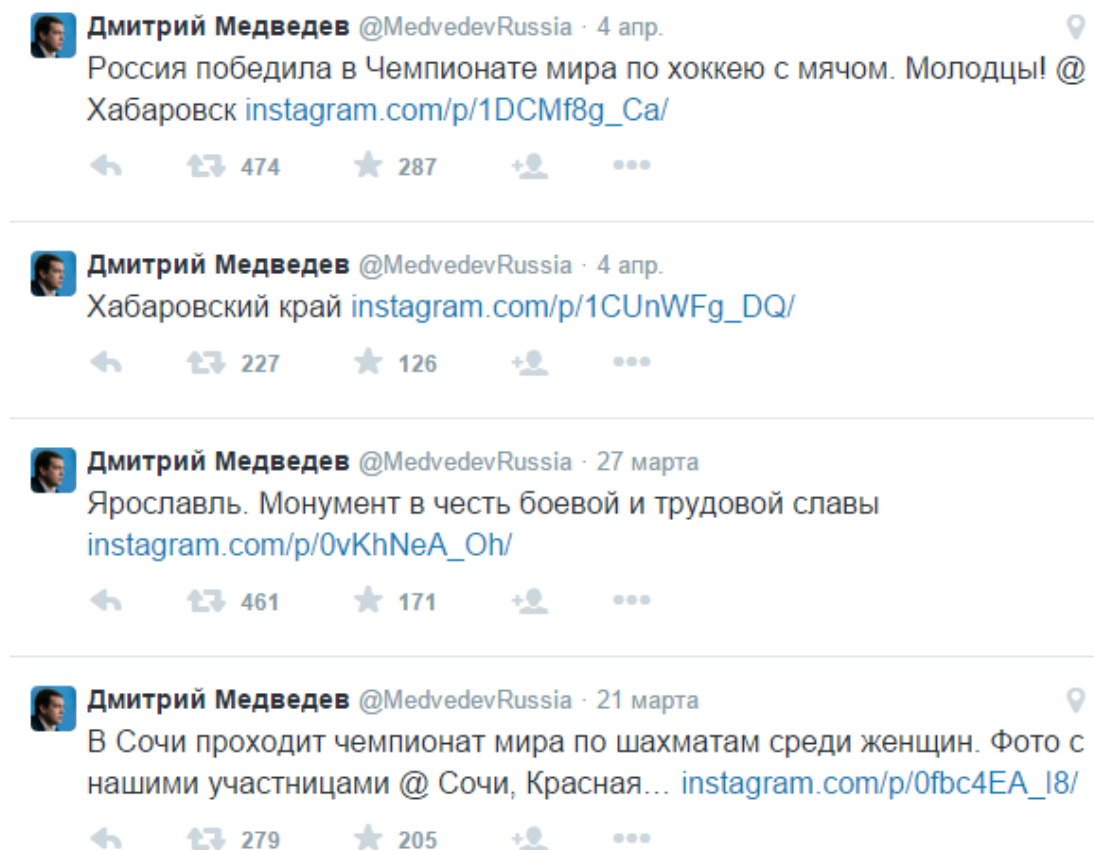


Рисунок 2. Лента сообщений пользователя¹.

Имя пользователя (`screen_name`) — уникальное имя пользователя социальной сети Twitter. Имя пользователя отображается в упоминаниях пользователей в виде «@`screen_name`». Также личная страница пользователя доступна по адресу: https://twitter.com/screen_name.

Статус пользователя (`status`) — сообщение, опубликованное пользователем на домашней странице. Также используется термин «твит» (`tweet`).

Ретвит (`retweet`) — сообщение, созданное в результате повторной отправки уже существующего сообщения.

1.5 Постановка задачи

В качестве исходных данных используются публичные статусы англоязычных пользователей социальной сети Twitter.

¹ <https://twitter.com/MedvedevRussia>

Необходимо научиться классифицировать хромотип пользователя социальной сети, анализируя его социальную активность.

В настоящее время существует система «Harbinger» [18], которая анализирует активность обновления статуса пользователя в течение суток и предсказывает дальнейшую активность пользователя. Система использует данные о времени в течение дня, когда пользователь публикует сообщение. Система моделирует дневную активность (обновление статусов) с помощью смеси двух гауссиан, используя ЕМ-алгоритм для разделения компонент смеси. Одна компонента смеси отвечает за дневной или вечерний пик активности. Недостатком данной системы является то, что она не предсказывает хромотип пользователя, а только предсказывает суточную активность.

1.6 Выводы по главе

Хромотип человека может быть полезен для выявления особенностей его стиля жизни, увлечений. Так, например, совы чаще любят ночную жизнь, посещение ночных клубов, ресторанов, увлекаются астрономией. Также сведения о хромотипе человека могут быть полезны для определения того, стоит ли заниматься сменной работой или работой, связанной с постоянными сдвигами во времени (пилот самолета, стюардесса). Целью исследования является применение методов машинного обучения для определения хромотипа пользователя социальной сети Twitter.

Глава 2. РАЗРАБОТКА МЕТОДОВ ОПРЕДЕЛЕНИЯ ХРОНОТИПА

В данном разделе будут рассмотрены теоретические подходы к решению поставленной задачи.

2.1 Анализ временных интервалов активности

Существуют исследования, подтверждающие связь хронотипа и активности человека в течение суток. Предполагается, что хронотип пользователя социальной сети также связан с его социальной активностью в течение суток. Активность будет скачкообразной и будет иметь два ярко выраженных пика: дневной и вечерний. Анализируя амплитуды данных пиков, их расположение во времени и расположение относительно друг друга можно сделать вывод о принадлежности пользователя к определенному хронотипу. В случае анализа пользователей социальной сети невозможно определить хронотип с помощью методов социальных и психологических опросов, потому что это дорогостоящее мероприятие, требующее большого количества времени, а данные, которыми мы располагаем, имеют весьма внушительные размеры. Следовательно, при анализе социальной сети под хронотипом пользователя будем понимать время, в течение которого он наиболее активен (утро или вечер). Для того, чтобы пользоваться таким определением, необходимо проверить гипотезу о связи хронотипа пользователя и его времени активности в социальной сети. Следует провести опрос среди пользователей социальной сети для определения их хронотипа. Далее следует проанализировать их временную активность и сделать выводы о связи с хронотипом. Если будет выявлена связь, то разметка данных будет происходить на основании анализа пиков суточной активности.

2.2 Выделение признаков для анализа текстовых сообщений

Для определения хронотипа пользователя будем анализировать публичные текстовые сообщения (статусы). Существует ряд гипотез о связи хронотипа и прочих социально-медицинских показателей. Алгоритм классификации выявляет вероятность наличия тех или иных показателей у пользователя, анализируя текст публичных сообщений, после чего классифицирует данного пользователя. Некоторые из гипотез, используемых в классификаторе, приведены ниже.

2.2.1 Эмоциональная окраска

Хронотип может быть связан с эмоциональной окраской сообщений [19]. Для определения эмоциональной окраски можно анализировать число позитивных и негативных смайлов в текстах сообщений. В исследовании [20] анализировалась социальная сеть Twitter, смайлы были использованы в качестве меток для определения позитивного или негативного отношения автора к написанному. В результате был построен наивный байесовский классификатор [21], показатель точности которого составил 81%. В таблице 1 приведены примеры позитивных и негативных смайлов.

Таблица 1. Позитивные и негативные смайлы.

Позитивные	Негативные
:)	:(
:-)	:-(
:~))	:~((
=)	=(
;)~	:~'(
^_~	t_t
:~D	=\

Кроме того, можно анализировать отдельные слова и определять эмоциональную окраску сообщения исходя из эмоциональной окраски отдельных слов. Эмоциональная окраска слова оценивается по шкале негативного и позитивного отношения в некотором диапазоне, например, в диапазоне $[-5; 5]$ баллов, как это предложено в корпусе слов AFINN [22], состоящем из 2477 размеченных английских слов. Например, слово «breathtaking» имеет оценку «5», а слово «catastrophic» имеет оценку «-4» в данном наборе слов.

Таким образом, будем использовать в качестве признаков классификатора число позитивных и негативных смайлов в сообщениях пользователя. Также признаком классификатора будет суммарная оценка слов в сообщениях по шкале, используемой в корпусе AFINN. Коэффициент корреляции между эмоциональной окраской и целевым признаком составил 0,61.

2.2.2 Склонность к депрессии

Хронотип может быть связан со склонностью человека к депрессии [23]. Для анализа депрессивности сообщений можно подсчитывать число слов, входящих в списки депрессивной лексики [24]. Данные списки публикуются на веб-сайтах, темой которых является лечение депрессии и других психических расстройств [25]. Кроме того, анализируя число ретвитов, подписчиков, пользователей, за обновлениями которых следит данный пользователь, упоминаний других пользователей, можно сделать выводы об уровне социального взаимодействия, можно понять, использует ли человек Twitter как интернет-дневник или как средство общения с друзьями. Используя данную информацию, можно определить, насколько человек замкнут или общителен, ведь зачастую эти признаки являются симптомами депрессии [24].

Таким образом, будем использовать в качестве признаков классификатора число слов, содержащихся в списках депрессивной лексики, число ретвитов и упоминаний других пользователей, содержащихся в сообщениях

пользователя, число упоминаний слов «психотерапевт», «антидепрессант», число упоминаний названий лекарственных средств, являющихся антидепрессантами. Коэффициент корреляции между депрессивностью сообщений и целевым признаком составил 0,59.

2.2.3 Склонность к употреблению кофеина

Хронотип может быть связан с пристрастием человека к употреблению кофе [26]. Большинство сов испытывают проблемы с ранним пробуждением и чувствуют себя уставшими рано утром. В связи с этим, они употребляют кофе для того, чтобы взбодрить свой организм. Для выявления данного признака можно анализировать частоту употребления слов «кофе», «кофеин» и прочих слов, связанных с данной тематикой.

Таким образом, будем использовать в качестве признаков классификатора число употреблений слов «кофе», «кофеин», «кофейня», «арабика», «эспresso», «латте», «американо», «капучино», «мокко», «глясе», «фраппе», «энергия» [27]. Коэффициент корреляции между склонностью к употреблению кофе и целевым признаком составил 0,3.

2.2.4 Склонность к набору излишнего веса

Хронотип может быть связан с излишним весом пользователя [28]. Нередко вес человека связан с его пристрастием ко вкусной пище. Также поздние приемы пищи могут спровоцировать набор веса. Поэтому можно анализировать упоминания различных калорийных блюд, в том числе в позднее время, в сообщениях пользователя.

Таким образом, будем использовать в качестве признаков классификатора число употреблений названий блюд, десертов, слов «вес», «еда», «пища», «диета». Коэффициент корреляции между склонностью к набору веса и целевым признаком составил 0,21.

2.2.5 Проблемы со сном

Хронотип может быть связан с различными расстройствами сна, такими как бессонница, тревожный сон, сонливость в течение дня, усталость, неравномерность сна [26,29]. Эти признаки можно выявлять с помощью анализа частоты употребления пользователем таких слов, как «сон», «бессонница».

Таким образом, будем использовать в качестве признаков классификатора число употреблений слов «сон», «бессонница», «усталость», «ночь», «утро», «спать». Коэффициент корреляции между проблемами со сном и целевым признаком составил 0,32.

2.2.6 Склонность к курению и употреблению алкоголя

Хронотип может быть связан со склонностью к курению и употреблению алкоголя [30,31]. Считается, что вечерние типы людей чаще являются курильщиками и в больших дозах употребляют алкогольные напитки.

Таким образом, будем использовать в качестве признаков классификатора число употреблений названий алкогольных напитков, слов «сигареты», «сигары», «зажигалка», «спички», «алкоголь», «пьян», «зависимость». Коэффициент корреляции между склонностью к курению и употреблению алкоголя и целевым признаком составил 0,24.

2.3 Выводы по главе

На основе анализа предметной области было выделено 10 признаков, которые будут использоваться при построении классификатора:

- позитивные и негативные смайлы,
- суммарная оценка слов по шкале AFINN,
- депрессивные слова,
- ретвиты,
- упоминания,
- слова, связанные с кофе,

- слова, связанные с излишним весом,
- слова, связанные с проблемами со сном,
- слова, связанные с алкоголем и курением.

Глава 3. СБОР ДАННЫХ

В данной главе будут рассмотрены используемые особенности сбора данных для последующего анализа и построения классификаторов.

3.1 Сбор информации о хронотипе

Для обучения классификатора требуется получить выборку данных, содержащую информацию о хронотипах пользователей социальной сети. Хронотип человека обычно определяется с помощью некоторых тестов [32], содержащих вопросы о том, в какое время человек чаще всего просыпается, ложится спать, когда чувствует прилив сил, в какое время суток он наиболее конфликтен, насколько много времени ему требуется, чтобы уснуть или встать с кровати после пробуждения, насколько тяжело он переносит разницу между часовыми поясами в путешествии и т.д. Данные тесты позволяют с определенной долей уверенности сделать выводы о принадлежности человека к тому или иному хронотипу.

Проблема применения подобных тестов заключается в том, что вовлечение в подобные тесты большого числа людей может быть весьма затруднительным, ведь проведение тестирования требует временных затрат, финансовой поддержки и штата социологов, готовых опрашивать добровольцев. Исходя из этого, будем считать, что невозможно опросить большое количество людей традиционным способом.

Для решения данной проблемы воспользуемся методом так называемой «грязной разметки» [33,34]. Суть метода заключается в том, что по причине невозможности или дороговизны осуществления реальной разметки данных, разметка производится на основе некоторого признака, который может заменить реальную разметку. Обычно при построении модели для определенных данных считается, что метки классов расставлены верно. В случае «грязной разметки» можно считать, что метки классов зашумлены. То есть провести точную разметку невозможно, и мы размечаем данные более доступным, но

не всегда точным способом. Например, в работе [20] для определения эмоциональной окраски сообщений были введены метки, основывающиеся на содержащихся в сообщениях смайлах.

Для того чтобы получить корректно размеченные данные, необходимо было провести опрос для определения хронотипа среди активных пользователей социальной сети Twitter. Были применены следующие способы сбора информации:

- личный опрос жителей Санкт-Петербурга о хронотипе,
- личное тестирование жителей Санкт-Петербурга для определения хронотипа,
- публикация и распространение теста для определения хронотипа в сети Интернет,
- публикация и распространение опроса о хронотипе пользователей Twitter (на русском и английском языках),
- публикация опросов и тестов в сообществах, связанных с социологическими исследованиями и социальной сетью Twitter,
- поиск упоминаний хронотипа пользователя в публичных сообщениях социальных сетей.

Предполагается, что человек готов потратить немного времени для прохождения личного тестирования и больше времени для прохождения тестирования через интернет. Для личного тестирования был составлен тест, приведенный в приложении 1, построенный на основании одного из предложенных способов определения хронотипа [35]. При выборе теста учитывались такие показатели, как число вопросов, время, необходимое для прохождения теста, сложность вопросов. Для тестирования через интернет использовался более объемный тест, взятый с медицинского портала [36], текст которого также приведен в приложении 1.

По результатам опроса были получены данные о хронотипах 320 человек, 41% из которых оказались жаворонками, а 59% — совами (рис. 3).

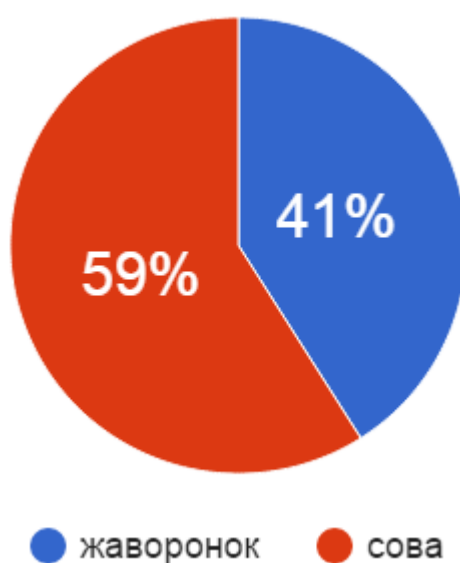


Рисунок 3. Хронотипы опрошенных пользователей.

Возраст опрошенных варьировался в диапазоне 18–35 лет (рис. 4). По статистике, собранной в 2015 году [37], 85% пользователей социальной сети Twitter младше 40 лет. Исходя из этого, выборку можно считать репрезентативной.

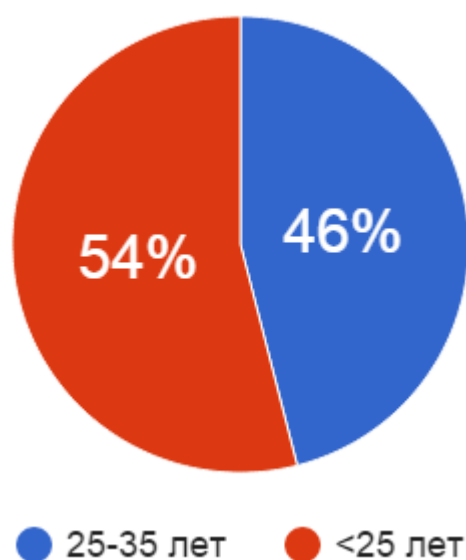


Рисунок 4. Возраст опрошенных пользователей.

Среди опрошенных добровольцев было примерно одинаковое число мужчин и женщин.

3.2 Сбор информации о пользователе Twitter

Для разметки большого количества реальных данных требуется предложить способ автоматической разметки данных. Выдвинем гипотезу о том, что хромотип человека связан со временем его активности в социальной сети в течение суток. Суточная активность представляет собой дневной и вечерний пики, анализируя амплитуды и положения во времени которых, можно сделать вывод об активности человека в течение дня [18]. Для моделирования суточной активности воспользуемся алгоритмом разделения смеси гауссиан.

3.2.1 Моделирование пиков суточной активности

Суточная активность пользователя моделируется смесью двух гауссиан с помощью ЕМ-алгоритма [16]. Алгоритм итеративно вычисляет параметры компонент смеси распределений, с каждым шагом вычисляя более точное приближение параметров. Шаги ЕМ-алгоритма для разделения смеси нормальных распределений представлены ниже.

Пусть вектор параметров $\theta = (\omega_1, \omega_2, \mu_1, \mu_2, \sigma_1, \sigma_2)$, где ω_i — скрытые параметры, μ_i — матожидание компоненты смеси, σ_i — среднее квадратичное отклонение компоненты смеси.

Пусть плотность распределения задается следующей формулой:

$$p_j(x) = N(x; \mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right).$$

Е-шаг алгоритма:

$$g_{ij} = \frac{\omega_j N(x_i; \mu_j, \sigma_j)}{\sum_{s=1}^2 \omega_s N(x_i; \mu_s, \sigma_s)},$$

М-шаг алгоритма:

$$\omega_j = \frac{1}{m} \sum_{i=1}^m g_{ij},$$

$$\mu_j = \frac{1}{m\omega_j} \sum_{i=1}^m g_{ij}x_i,$$

$$\sigma_j^2 = \frac{1}{m\omega_j} \sum_{i=1}^m g_{ij} (x_i - \mu_j)^2, j = 1, 2;$$

m — размер выборки.

Пример моделирования суточной активности пользователя изображен на рисунке 5.

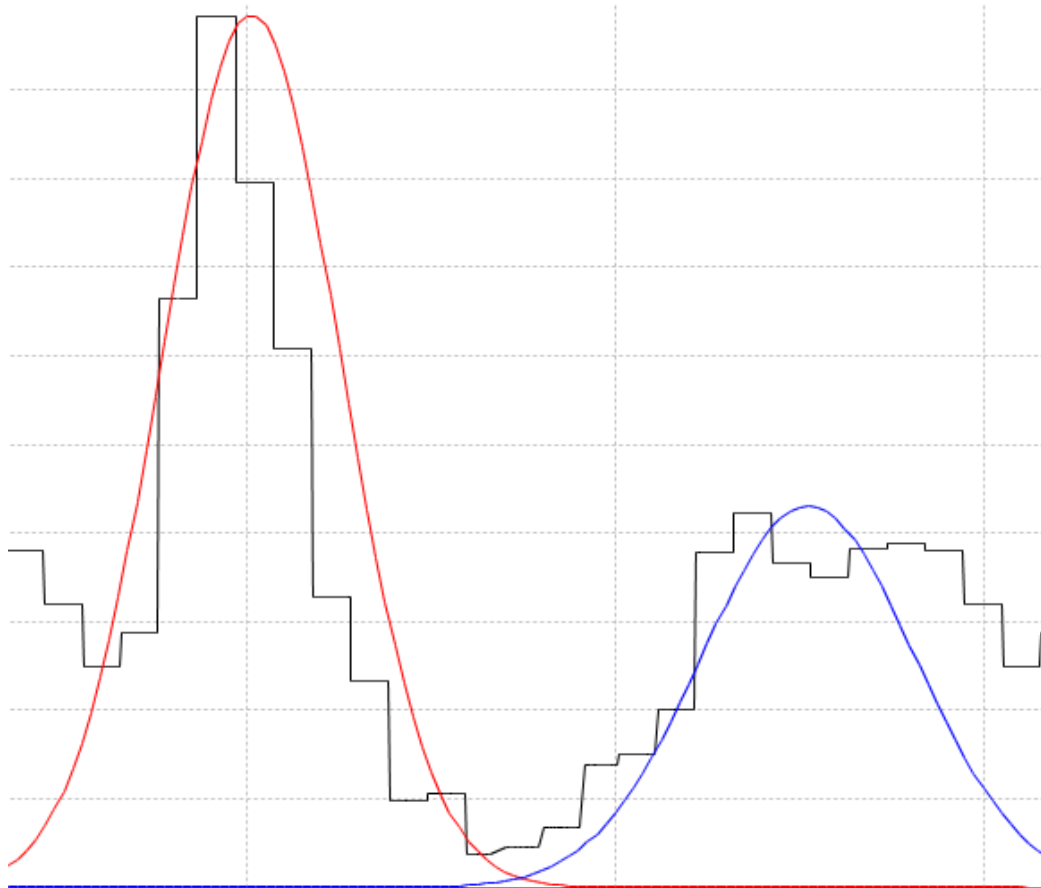


Рисунок 5. Суточная активность пользователя.

3.2.2 Связь хронотипа пользователя и его временной активности

В результате тестирования были опрошены 320 человек. После фильтрации пользователей с приватными профилями и пользователей, в профиле которых не указаны данные о локальном времени, осталось 230 аккаунтов Twitter.

Для оценки связи между хронотипом пользователя и суточной активностью в социальной сети был рассчитан коэффициент корреляции двух случайных величин (хронотипа и пика суточной активности), значение которого составило 0,69, что говорит о наличии связи между случайными величинами. Следовательно, показана правомерность использования информации о пиках активности вместо реального хронотипа пользователя.

3.2.3 Формирование обучающей выборки

Для анализа публичных сообщений необходимо собрать следующую информацию о пользователях социальной сети Twitter, имея данные об именах пользователей:

- уникальный идентификационный номер пользователя,
- язык, используемый пользователем,
- географическое местоположение пользователя,
- смещение часового пояса пользователя относительно UTC [38],
- последние публичные сообщения пользователя.

Также для формирования обучающей выборки необходимо иметь список имен пользователей.

Сбор данных из социальной сети Twitter происходил с помощью библиотеки Twitter4J [39], которая оперирует методами Twitter API [40]. Имена пользователей для обучающего множества выбирались случайным образом среди последних публикуемых сообщений.

Для обучающей выборки были отобраны случайным образом 1000 англоязычных аккаунтов социальной сети Twitter. Также для каждого пользователя были собраны последние 400 сообщений. Данные были размечены, основываясь на информации о пиках активности, для дальнейшего обучения классификатора. По результатам разметки 54% аккаунтов были помечены хронотипом сова, 46% — жаворонок.

3.3 Выводы по главе

В качестве данных для анализа требуется провести опрос с целью выявления реального хронотипа, собрать информацию о временной активности пользователей, проверить гипотезу о связи хронотипа и суточной активности, чтобы пользоваться автоматической разметкой. Также требуется собрать информацию о содержании публичных текстовых сообщений пользователей.

Глава 4. ПОСТРОЕНИЕ КЛАССИФИКАТОРА

В данной главе будут рассмотрены основные особенности построения классификатора. Классификатор занимает центральное место приложения, которое характеризуется следующими этапами работы: сбор данных о пользователях социальной сети Twitter, разметка выборки с помощью анализа пиков временной активности, обучение классификатора на полученной выборке с целью дальнейшего предсказания хронотипа других пользователей.

4.1 Алгоритмы классификации

Обучающая выборка была разделена на две части: для обучения и для оценки качества классификатора. 66% исходных данных были использованы для обучения модели, 34% использовались для тестирования полученной модели. При формировании частей выборки имена пользователей были случайным образом перемешаны, данные о сообщениях пользователей были собраны после разделения выборки на две части. В качестве атрибутов для классификатора использовались частоты встречаемости определенных слов, исходя из гипотез о связи хронотипа и некоторых социально-медицинских показателей, описанных в разделе 2.2. В качестве алгоритмов классификации были использованы нижеперечисленные алгоритмы из библиотеки машинного обучения Weka [41].

Метод k ближайших соседей — метрический алгоритм классификации, основным принципом которого является то, что объект относится к тому классу, который наиболее распространен среди соседей данного объекта [42].

Наивный байесовский классификатор — классификатор, в основе которого лежит применение теоремы Байеса [43]. Является частным случаем байесовского классификатора, использует дополнительное предположение о независимости атрибутов объектов [21].

Многослойный перцептрон — многослойная искусственная нейронная сеть, состоящая из нескольких узлов входного слоя, нескольких скрытых

слоев и одного выходного слоя [44]. Одним из самых популярных алгоритмов обучения такой сети является алгоритм **обратного распространения ошибки** [45].

Сеть радиальных базисных функций — нейронная сеть, использующая в качестве функции активации радиальные базисные функции. Такие сети имеют следующие отличительные особенности: содержат один скрытый слой, нейроны скрытого слоя задаются нелинейной функцией активации, веса входного и выходного слоев равны единице [46].

KStar — алгоритм ленивого обучения, использующий энтропию в качестве меры расстояния [47].

Алгоритмы бустинга представляют собой ансамбль алгоритмов машинного обучения, в котором каждый новый классификатор стремится минимизировать недостатки текущей композиции классификаторов [48]. Процедура построения композиции классификаторов является жадной. Основная идея бустинга состоит в том, чтобы из множества простых некачественных классификаторов получить новый хороший классификатор. Наиболее популярным видом бустинга является бустинг над деревьями решений. Рассмотрим такие алгоритмы бустинга, как AdaBoost [49], LogitBoost [50], Random Forest [51].

4.2 Результаты работы классификаторов

Для оценки полезности использования метода «грязной разметки» в данном исследовании классификаторы были построены на двух выборках данных: данные с реальной разметкой, полученные при проведении опроса о хронотипе, и данные, разметка которых была сделана автоматически на основе информации о пиках суточной активности пользователя.

В таблице 2 приведены оценки качества работы алгоритмов классификации, где в качестве метки класса использовались пики активности.

Таблица 2. Результаты работы алгоритмов классификации при использовании автоматической разметки.

Алгоритм	F_1 -мера
k NN	0,63
NaiveBayes	0,62
MultilayerPerceptron	0,73
RBFNetwork	0,58
KStar	0,66
AdaBoost	0,74
LogitBoost	0,68
RandomForest	0,70

В таблице 3 приведена матрица неточностей для классификатора AdaBoost.

Таблица 3. Матрица неточностей для классификатора AdaBoost.

	Сова	Жаворонок
Сова	132	35
Жаворонок	52	121

Если же тренировать и тестировать алгоритмы классификации, используя реально размеченные данные по результатам опроса, получим следующие показатели F_1 -меры для построенных классификаторов, приведенные в таблице 4.

Таблица 4. Результаты работы алгоритмов классификации при использовании ручной разметки.

Алгоритм	F_1 -мера
k NN	0,56
NaiveBayes	0,58

MultilayerPerceptron	0,65
RBFNetwork	0,7
KStar	0,54
AdaBoost	0,67
LogitBoost	0,71
RandomForest	0,70

В таблице 5 приведена матрица неточностей для классификатора LogitBoost.

Таблица 5. Матрица неточностей для классификатора LogitBoost.

	Сова	Жаворонок
Сова	32	8
Жаворонок	14	24

Исходя из приведенных таблиц, лучшие показатели F_1 -меры составили 0,74 для автоматической разметки и 0,71 для ручной разметки данных.

Предложенные классификаторы определяют некоторый демографический показатель (хронотип). Для того чтобы оценить качество полученного решения, рассмотрим результаты исследований в области определения прочих демографических показателей. В работе [52] определялась принадлежность жителей США к определенной расе (афроамериканцы или нет) на основе анализа активности в социальной сети Twitter, показатель F_1 -меры составил 0,65. В работе [52] также определялась симпатия пользователя социальной сети Twitter к партии демократов или республиканцев в США, показатель F_1 -меры составил 0,91. Исходя из этого, заключим, что различные демографические показатели определяются с разной степенью успешности и полученный результат работы классификатора соответствует прочим результатам в данной области.

4.3 Анализ результатов

Учитывая то, что данная область малоизучена, и сейчас хронотип определяется только с помощью социологических и медицинских исследований, возможность определения хронотипа автоматизированным способом может помочь в обработке больших объемов данных. Также если учитывать исследования о связи хронотипа и социально-медицинских показателей, то данные о хронотипе могут помочь в предсказании некоторых болезней, психических расстройств. Кроме того, возможность автоматизации определения хронотипа может подтолкнуть медицинское сообщество на углубление дальнейших исследований в данной теме. В условиях отсутствия аналогичных работ, результаты работы классификатора можно считать удовлетворительными. В дальнейшем можно ввести показатель уверенности классификатора и считать достоверными результаты, учитывая определенный порог уверенности. Следовательно, можно автоматически определять хронотип пользователя, а в случае неуверенности классификатора проводить ручной анализ на определенных данных.

4.4 Выводы по главе

Были протестированы различные алгоритмы построения классификаторов, показатель F_1 -меры равен 0,71 для набора данных, размеченного вручную. Учитывая то, что значения различных демографических показателей определяются с разной степенью точности, можно считать полученную точность удовлетворительной. Точность алгоритмов ставится под сомнение ввиду малоизученности способов определения хронотипа. Однако, данное исследование может послужить толчком к развитию области и помочь в смежных областях.

Заключение

Данная работа направлена на разработку методов определения хроно-типа пользователя социальной сети. Были собраны данные об активности пользователей социальной сети Twitter. Также было проведено социологическое исследование с целью выявить связь между хронотипом человека и его активностью в социальной сети, были получены положительные результаты. Были найдены свидетельства о связи хронотипа и определенных социально-медицинских показателей, были разработаны методы выявления данных показателей посредством анализа сообщений пользователей. Были построены классификаторы, предсказывающие хронотип человека. Также автоматизированная классификация людей по хронотипу поможет обрабатывать большие объемы данных для медицинских исследований, связанных с хронотипом и стимулировать интерес к углублению знаний в областях науки, связанных с хронотипом. На основе этого можно заключить, что все задачи исследования выполнены, а цель достигнута.

Список литературы

1. Mohsen Jamali, Hassan Abolhassani. IEEE/WIC/ACM International Conference on Web Intelligence // Different Aspects of Social Network Analysis. Hong Kong. 2006.
2. Alexander Pak, Patrick Paroubek. Proceedings of the International Conference on Language Resources and Evaluation // Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Malta. 2010.
3. Butts C.T. Social Network Analysis with sna // Journal of Statistical Software, Vol. 24, No. 6, 2008.
4. Scott J. Social Network Analysis. Los Angeles: SAGE, 2012.
5. Stephen P Borgatti, Martin G. Everett, Jeffrey C. Johnson. Analyzing Social Networks. Los Angeles: SAGE, 2013.
6. Phillips M.L. Circadian rhythms: Of owls, larks and alarm clocks // Nature, No. 458, 2009. pp. 142-144.
7. Thomas Kantermann, Myriam Juda, Céline Vetter, Till Roenneberg. Shift-work research: Where do we stand, where should we go? // Sleep and Biological Rhythms, Vol. 8, No. 2, 2010. pp. 95-105.
8. Jessica Rosenberg, Ivan I. Maximov, Martina Reske, Farida Grinberg, N. Jon Shah. "Early to bed, early to rise": Diffusion tensor imaging identifies chronotype-specificity // NeuroImage, 2013.
9. Chronotype definition [Электронный ресурс] URL: <http://www.medilexicon.com/medicaldictionary.php?t=17540> (дата обращения: 4.Декабрь.2014).
10. MCTQ [Электронный ресурс] URL: https://www.bioinfo.mpg.de/mctq/core_work_life/core/introduction.jsp?language=eng (дата обращения:

20.Май.2015).

11. MEQ [Электронный ресурс] URL: <http://www.cet-hosting.com/limesurvey/?sid=61524> (дата обращения: 19.Май.2015).
12. Bishop C. Pattern Recognition and Machine Learning. New York: Springer, 2007.
13. Классификация [Электронный ресурс] URL: <http://www.machinelearning.ru/wiki/index.php?title=Классификация> (дата обращения: 22.Май.2015).
14. Нормальное распределение [Электронный ресурс] URL: https://ru.wikipedia.org/wiki/Нормальное_распределение (дата обращения: 23.Май.2015).
15. Trevor Hastie, Robert Tibshirani, Jerome Friedman. The EM Algorithm // In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009. pp. 272-279.
16. ЕМ-алгоритм [Электронный ресурс] URL: <https://ru.wikipedia.org/wiki/ЕМ-алгоритм> (дата обращения: 18.Февраль.2015).
17. Twitter [Электронный ресурс] URL: <https://twitter.com/> (дата обращения: 20.Май.2015).
18. Rui Guo, Hongzhi Wang, Lucheng Zhong, Jianzhong Li, Hong Gao. DASFAA // Harbinger: An Analyzing and Predicting System for Online Social Network Users' Behavior. Bali. 2014.
19. Ottoni GL, Antonioli E, Lara DR. Circadian preference is associated with emotional and affective temperaments. // Chronobiology International, 2012.
20. Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision, 2009.

21. Trevor Hastie, Robert Tibshirani, Jerome Friedman. Statistical Decision Theory // In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009. pp. 20-21.
22. AFINN [Электронный ресурс] URL: <http://neuro.imm.dtu.dk/wiki/AFINN> (дата обращения: 7.Июнь.2015).
23. Maria Paz Hidalgo, Wolnei Caumo, Michele Posser, Sônia Beatriz Cocco, Ana Luiza Camozzato, Márcia Lorena Fagundes Chaves. Relationship between depressive mood and chronotype in healthy subjects // Psychiatry and Clinical Neurosciences, 2009. pp. 283-290.
24. Munmun De Choudhury, Michael Gamon, Scott Counts, Eric Horvitz. Predicting Depression via Social Media.
25. Words about depression [Электронный ресурс] URL: <https://adarkersshadeofblue.wordpress.com/2011/08/20/99-words-about-depression/> (дата обращения: 21.Май.2015).
26. Giannotti F, Cortesi F, Sebastiani T, Ottaviano S. Circadian preference, sleep and daytime behaviour in adolescence // Journal of sleep research, 2002.
27. Glossary of coffee terms // Grumpy Mule. URL: <http://www.grumpymule.co.uk/coffee-tour/glossary-of-terms> (дата обращения: 2.Июнь.2015).
28. Culnan E, Kloss JD, Grandner M. A prospective study of weight gain associated with chronotype among college freshmen // Chronobiology International, 2013.
29. Jeanne Sophie Martin, Marc Hébert, Élise Ledoux, Michaël Gaudreault, Luc Laberge. Relationship of Chronotype to Sleep, Light Exposure, and Work-Related Fatigue in Student Workers // Chronobiology International, 2012.

30. Urbán R, Magyaródi T, Rigó A. Morningness-eveningness, chronotypes and health-impairing behaviors in adolescents // Chronobiology International, Vol. 28, No. 3, 2011. pp. 238-247.
31. Wittmann M, Paulus M, Roenneberg T. Decreased psychological well-being in late 'chronotypes' is mediated by smoking and alcohol consumption // Substance use misuse, Vol. 45, No. 1-2, 2010. pp. 15-30.
32. Andrei Zavada, Marijke C. M. Gordijn, Domien G. M. Beersma, Serge Daan, Till Roenneberg. Comparison of the munich chronotype questionnaire with the horne-o'stberg's morningness-eveningness score // Chronobiology International, Vol. 22, No. 2, 2005. pp. 267-278.
33. Victor S. Sheng, Foster Provost, Panagiotis G. Ipeirotis. Knowledge Discovery and Data Mining // Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. 2008.
34. Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar. Learning with Noisy Labels,.
35. Голубь, жаворонок или сова [Электронный ресурс] URL: <http://happydiva.ru/time-management/71-golub-zhavoronok-ili-sova-4-testa> (дата обращения: 22.Май.2015).
36. Хронотипы человека - онлайн тест [Электронный ресурс] URL: <http://hematologiya.ru/test-na-opredelenie-khronotipa.html> (дата обращения: 23.Май.2015).
37. Number of Twitter users in the United States as of January 2015, by age group URL: <http://www.statista.com/statistics/398152/us-twitter-user-age-groups/> (дата обращения: 9.Июнь.2015).
38. UTC time [Электронный ресурс] URL: http://www.worldtimeserver.com/current_time_in.UTC.aspx (дата обращения: 21.Май.2015).

39. Twitter4J [Электронный ресурс] URL: <http://twitter4j.org/en/index.html>
(дата обращения: 21.Май.2015).
40. Twitter API [Электронный ресурс] URL: <https://dev.twitter.com/overview/api>
(дата обращения: 20.Май.2015).
41. Weka [Электронный ресурс] URL: <http://www.cs.waikato.ac.nz/ml/weka/>
(дата обращения: 22.Май.2015).
42. Trevor Hastie, Robert Tibshirani, Jerome Friedman. k-Nearest-Neighbor Classifiers // In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009. P. 463.
43. Чернова Н.И. Формула Байеса // В кн.: Теория вероятностей. Новосибирск. 2009. С. 38-39.
44. Хайкин С. Многослойный персептрон // В кн.: Нейронные сети. Полный курс (перевод с английского). Москва: Вильямс, 2006. С. 219.
45. Хайкин С. Алгоритм обратного распространения // В кн.: Нейронные сети. Полный курс (перевод с английского). Москва: Вильямс, 2006. С. 225-240.
46. Хайкин С. Сети на основе радиальных базисных функций // В кн.: Нейронные сети. Полный курс (перевод с английского). Москва: Вильямс, 2006. С. 341.
47. John G. Cleary, Leonard E. Trigg. K*: An Instance-based Learner Using an Entropic Distance Measure, 1995.
48. Trevor Hastie, Robert Tibshirani, Jerome Friedman. Boosting and Additive Trees // In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009. P. 337.
49. Yoav Freund, Robert E. Schapire. Thirteenth International Conference on

Machine Learning // Experiments with a New Boosting Algorithm. San Francisco. 1996.

50. Jerome Friedman, Trevor Hastie, Robert Tibshirani. Additive logistic regression: a statistical view of boosting // Annals of Statistics, Vol. 28, No. 2, 2000. pp. 337-407.
51. Breiman L. Random Forests // Machine Learning, Vol. 45, No. 1, 2001. pp. 5-32.
52. Marco Pennacchiotti, Ana-Maria Popescu. Fifth International Conference on Weblogs and Social Media // A Machine Learning Approach to Twitter User Classification. Barcelona. 2011.
53. Hidalgo MP, Caumo W, Posser M, Coccaro SB, Camozzato AL, Chaves ML. Relationship between depressive mood and chronotype in healthy subjects. 2009.
54. Yiming Yang, Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization // ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1997. pp. 412-420.
55. Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, Zheng Chen. International World Wide Web Conference // Demographic Prediction Based on User's Browsing. New York. 2007.
56. Кормен, Томас Х., Лейзерсон, Чарльз И., Ривест, Рональд Л., Штайн Клиффорд. Алгоритмы: построение и анализ, 2-е издание. Пер. с англ. Издательский дом "Вильямс", 2010.
57. Машинное обучение [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение (дата обращения: 22.Май.2015).

58. Машинное обучение [Электронный ресурс] URL: http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение (дата обращения: 22.Май.2015).

Приложения

Приложение 1. Тесты для определения хронотипа

Тест для прохождения лично

- 1) Утром вы с просыпаетесь с трудом:
 - a) почти всегда (3),
 - b) иногда (2),
 - c) редко (1),
 - d) довольно редко (0).
- 2) Если бы вы могли выбирать, то в какое время бы вы предпочли лечь спать:
 - a) после часа ночи (3),
 - b) между одиннадцатью и часом (2),
 - c) между десятью и одиннадцатью (1),
 - d) до десяти (0).
- 3) В первый час после пробуждения вы выберете:
 - a) плотный завтрак (0),
 - b) легкий завтрак (1),
 - c) вареное яйцо (2),
 - d) чашку кофе (3).
- 4) В какое время вы часто ссоритесь с семьей и коллегами?
 - a) до полудня (1),
 - b) после полудня (0).
- 5) Вам легко отказаться от:
 - a) кофе утром (2),
 - b) кофе вечером (0).
- 6) Легко ли вам поменять предпочтения в еде?
 - a) легко (0),
 - b) без особых проблем (1),

- c) трудно (2),
 - d) невозможно (3).
- 7) Если вы планируете ответственное мероприятие на утро следующего дня:
- a) вы ляжете спать больше, чем на 2 часа раньше обычного (3),
 - b) вы ляжете спать раньше на 1-2 часа (2),
 - c) ляжете немного раньше (1),
 - d) ляжете как обычно (0).

Результаты:

0-6 баллов: жаворонок

7-11 баллов: голубь

12-18 баллов: сова

Тест для прохождения через интернет

- 1) Если у вас нет планов на следующий день, в какое время вы ляжете спать?
- a) 20:00-21:00,
 - b) 21:00-22:30,
 - c) 2:30-0:00,
 - d) 0:00-1:30,
 - e) 1:30-3:00.
- 2) Если вам нужно выполнить физическую работу по дому, какое время вы для этого выберете?
- a) 8:00-10:00,
 - b) 11:00-13:00,
 - c) 15:00-17:00,
 - d) 19:00-21:00.
- 3) Если завтра у вас будет неожиданный выходной, то в какое время вы ляжете спать?
- a) как обычно,

- b) на час позже обычного,
 - c) на два часа позже обычного,
 - d) намного позже.
- 4) Насколько вы должны устать, чтобы лечь в кровать в 21 час?
- a) очень сильно,
 - b) относительно сильно,
 - c) немного,
 - d) обычно ложусь в это время.
- 5) Насколько легко вам встать с кровати утром?
- a) очень тяжело,
 - b) достаточно тяжело,
 - c) относительно легко,
 - d) с удовольствием просыпаюсь и встаю.
- 6) Во сколько вы встаете в выходной день?
- a) 5:00-6:30,
 - b) 6:30-8:00,
 - c) 8:00-9:30,
 - d) 9:30-11:00,
 - e) позже 11:00.
- 7) Вы отнесете себя к «утреннему» или «вечернему» типу?
- a) определенно к «утреннему»,
 - b) скорее к «утреннему»,
 - c) смешанный тип,
 - d) ближе к «вечернему»,
 - e) определенно к «вечернему».
- 8) В какое время дня вы чувствуете максимальный прилив сил?
- a) 0:00-5:00 или 20:00-23:00,

- b) 4:00-8:00,
- c) 7:00-10:00,
- d) 10:00-16:00,
- e) 15:00-21:00.

9) Если вы легли немного позже, чем обычно, а на следующий день нет никаких дел, во сколько вы проснетесь?

- a) как обычно,
- b) немного подремлю,
- c) проснусь как обычно, но потом засну,
- d) проснусь позже, чем обычно.

10) Оцените вашу степень бодрости в первые 30 минут после того, как проснетесь

- a) сильная вялость,
- b) небольшая вялость,
- c) относительно деятелен,
- d) очень деятелен.

11) Насколько вы способны проснуться рано без будильника

- a) всегда так делаю,
- b) иногда получается,
- c) редко получается,
- d) всегда просыпаюсь с будильником.

12) Насколько вы голодны после пробуждения

- a) нет аппетита,
- b) слабый аппетит,
- c) относительно хороший аппетит,
- d) слона бы съел.

13) Как вы смотрите на пробежку по утрам в 7 часов утра?

- a) с легкостью,
- b) с трудом, но можно,
- c) невозможно.

14) Когда у вас период наибольшей работоспособности?

- a) 8:00-10:00,
- b) 11:00-13:00,
- c) 15:00-17:00,
- d) 19:00-21:00.

15) В какое время вы чувствуете, что готовы уснуть?

- a) до 21 часов
- b) до 22 часов,
- c) до 0:30,
- d) до 2 часов,
- e) после 2 часов.

Приложение 2. Строки, формирующие атрибуты классификатора

abandon	disheartened	insecure
achy	dismal	irrational
afraid	distractable	irritable
agitated	distraught	isolated
agon	distressed	lonely
alone	doomed	lousy
anguish	dreadful	low
antisocial	dreary	melancholy
anxious	edgy	insecure
breakdown	emotional	irrational
brittle	empty	irritable
broken	excluded	isolated
catatonic	exhausted	lonely
consumed	exposed	lousy
crisis	fatalistic	low
crushed	forlorn	melancholy
crying	fragile	miserable
defeated	freaking	moody
defensive	gloomy	morbid
dejected	grouchy	needy
demoralized	helpless	nervous
desolate	hopeless	nightmarish
despair	hurt	oppressed
desperate	inadequate	overwhelmed
despondent	inconsolable	pain
devastated	injured	paranoid

pessimistic	anxi	haha
reckless	alone	good
rejected	nervous	happy
resigned	pain	:(
sadness	sad	:'(
self-conscious	psychos	t_t:-(
self-disgust	pessimist	:((
shattered	suicid	((((
sobbing	antidepressant	((
sorrowful	psychotherap	cry
suffering	passiv	bad
suicidal	@	night
tearful	rt	morning
touchy	weight	coffee
trapped	food	caffe
uneasy	fat	espress
unhappy	drugs	arabic
unhinged	alcohol	
unpredictable	smok	
upset	insomni	
vulnerable	tired	
wailing	sleep	
weak	:)	
weepy	:-(
withdrawn	:))	
woeful)))	
wounded))	
wretched	:D	
depress	xD	