

МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра математической кибернетики и компьютерных наук

**РАЗРАБОТКА СИСТЕМЫ ОПРЕДЕЛЕНИЯ ЛИЧНОСТНЫХ И
ПРОФЕССИОНАЛЬНЫХ КАЧЕСТВ ЧЕЛОВЕКА ПО ЕГО
ИНФОРМАЦИОННОМУ СЛЕДУ В СОЦИАЛЬНЫХ СЕТЯХ**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 09.03.04 – Программная инженерия
факультета КНиИТ
Фартушновой Яны Сергеевны

Научный руководитель
к. ф.-м. н., доцент _____ А. С. Иванов

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
1 Нейронные сети	5
1.1 Машинное обучение	5
1.1.1 Обучение с учителем	5
1.1.2 Обучение без учителя	5
1.1.3 Глубокое обучение	6
1.2 Нейроны	6
1.3 Функции активации	7
2 Создание и обучение нейронной сети для определения тональности текста	8
2.1 Подготовка данных	8
2.2 Модель Word2vec	8
2.2.1 Описание модели word2vec	8
2.2.2 Обучение модели word2vec	9
2.3 Модель рекуррентной нейронной сети	10
2.3.1 Рекуррентные нейронные сети	10
2.3.2 Сети LSTM	11
3 Разработка автоматизированной системы поддержки принятия решений	12
3.1 Системы поддержки принятия решений	12
3.2 Маркеры профориентации для оценки личностных и профессиональных качеств человека	12
3.3 Разработка интерфейса системы поддержки принятия решений	14
ЗАКЛЮЧЕНИЕ	15
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	16

ВВЕДЕНИЕ

С недавнего времени в некоторых зарубежных странах анализируют активность в социальных сетях при приеме на работу кандидата, ссылку на которые он оставляет в своем резюме. Такой подход стал обретать популярность, так как резюме в настоящее время заполняются по определенному шаблону, в результате чего сложно определить личностные качества и профессиональные способности человека. Психологи проводят множество исследований, изучая модели поведения и речи человека, анализируя отличительные черты на фотографиях и в тексте, которые определяют наличие той или иной черты личности. В результате таких исследований психологи получают так называемые маркеры профориентации, на наличие которых проверяют страницы в социальных сетях кандидата и на основе которых делают выводы о его личностных и профессиональных качествах. Но такой способ требует наличие подготовленных специалистов в этой области, значительных временных ресурсов, требуемых для тщательной проверки, а также точности статистических данных и объективности анализа. Именно поэтому актуальной задачей является создание автоматизированной системы, позволяющей собирать необходимую информацию из социальных сетей кандидата, требуемую для анализа по психологическим маркерам. В качестве анализируемой социальной сети был выбран Instagram.

Целью настоящей работы является разработка автоматизированной системы поддержки принятия решений определения личностных и профессиональных качеств кандидата по информации, извлекаемой из социальной сети Instagram.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Анализ существующих модулей для работы с социальной сетью Instagram;
2. Изучение методов подготовки текстов для их анализа;
3. Анализ работы модели word2vec для преобразования текстов естественного языка в форму, пригодную для анализа;
4. Разработка нейронной сети для определения тональности текста;
5. Реализация автоматизированной системы поддержки принятия решений.

Выпускная квалификационная работа состоит из введения, трех глав, заключения, списка из 25 источников и трех приложений.

В главе «Нейронные сети» рассматривается теоретическая часть используемых технологий для разработки системы.

В главе «Создание и обучение нейронной сети для определения тональности текста» рассматривается алгоритм предварительной подготовки текстовых данных, описываются теоретические аспекты используемых технологий, объясняется использование архитектуры Skip-gram модели word2vec для получения векторных представлений слов, а так же выбор для реализации нейронной сети определения тональности текста технологий рекуррентных нейронных сетей и LSTM-слоев, описываются процессы создания и обучения используемых моделей. Также в главе проиллюстрированы и описаны примеры работы созданных и обученных моделей.

В главе «Разработка автоматизированной системы поддержки принятия решений» рассмотрены теоретические основы используемых технологий системы поддержки принятия решений, описаны лингвистические маркеры профориентации и их выявление с помощью программных средств, а также создан и описан интерфейс системы. В главе представлен пример работы приложения на конкретном профиле в Instagram и описан функционал.

1 Нейронные сети

1.1 Машинное обучение

Машинное обучение — это методы, используемые для анализа данных, позволяющие аналитической системе обучаться в ходе решения множества схожих по типу задач. Машинное обучение основано на том, что аналитические системы обладают способностью учиться находить закономерности в некотором объеме данных и принимать решения с минимальным участием человека.

В настоящее время алгоритмы машинного обучения применяются практически повсеместно, включая фильтры спама в электронной почте, автомобили с автоматическим управлением, распознавание звуков и изображений, рекомендательные системы и тому подобное, без чего современный человек не может представить комфортную жизнь. [1]

Машинное обучение разделяют на три типа:

1. Обучение с учителем;
2. Обучение без учителя;
3. Глубокое обучение.

1.1.1 Обучение с учителем

Обучение с учителем предполагает наличие определенного набора выборок в качестве входных данных, где ожидаемые выходные данные, или метки, уже известны и находятся в соответствии с элементами выборок. Модели обучаются на основе известных результатов и изменяют внутренние параметры для того, чтобы отобразить найденные закономерности во входных данных. Основная идея такого метода заключается в том, что обученная модель будет принимать на вход данные, похожие на тренировочные, которые она использовала при обучении, и выявленные закономерности позволят легко обрабатывать новую информацию и принимать решения на ее основе.

1.1.2 Обучение без учителя

При обучении без учителя изначально не известны переменные результата. На вход подаются непромаркованные данные неизвестной структуры. Необходимо проанализировать структуру данных, выделить значимую информацию и найти в ней определенные закономерности.

1.1.3 Глубокое обучение

Глубокое обучение может быть как с учителем, так и без него, его отличием от предыдущих видов является анализ больших объемов информации BigData, поэтому глубокое обучение использует нейронные сети.

Глубокое обучение использует иерархическую структуру искусственных нейронных сетей, которые построены аналогично мозгу человека со связанными узлами нейронов, образующими сеть. Такой подход позволяет осуществлять анализ данных нелинейным способом. Каждый из уровней нейронной сети получает на вход набор данных, внутри уровня происходит обработка полученных данных и извлечение некоторой информации, затем обработанные данные в качестве выходных передаются на следующий слой, в котором. Таким образом происходит каждый последующий уровень обрабатывает информацию, полученную из предыдущего уровня, до тех пор, пока не придет очередь последнего уровня, который на выходе делает предсказание.

1.2 Нейроны

Искусственная нейронная сеть — математическая модель, целью которой является моделирование аналитических механизмов, осуществляемых человеческим мозгом. Искусственная нейронная сеть представляет собой систему взаимодействующих между собой нейронов.

Нейронная сеть состоит из нескольких слоев нейронов, каждый из которых отвечает за распознавание одного из признаков или критериев: форма, цвет, размер и так далее.

Нейронные сети применяются в сложных задачах, для решения которых необходимо применение аналитических вычислений, подобных тем, что осуществляет человеческий мозг. [2]

Нейрон — вычислительная единица, которая производит над получаемой информации какие-либо вычисления и передает ее дальше. Существует три типа нейронов: входной, скрытый и выходной.

Если необходимо использовать большое количество нейронов, то используют понятие слой. Таким образом входные нейроны составляют входной слой, который получает информацию, скрытые нейроны составляют n скрытых слоев для обработки полученной информации, а выходные нейроны — выходной слой, который предоставляет результат вычислений.

У каждого нейрона есть два основных параметра:

1. входные данные,
2. выходные данные.

Под входными данными нейрона подразумевается суммарная информация всех связанных с ним нейронов предыдущего слоя, которая внутри нейрона нормализуется, проходит через функцию активации и в качестве выходных данных передается на следующий слой.

Нейроны входного слоя обычно не выполняют никаких преобразований над входными данными, внутри него используется линейная функция активации, и полученные сигналы передаются на вход нейронам скрытого слоя.

Внутри скрытого слоя происходит основная обработка данных и используется нелинейная функция активации.

Нейроны выходного слоя необходимы для вывода результата работы сети, но внутри него также могут выполняться некоторые вычисления. [3]

1.3 Функции активации

Функция активации нейрона определяет выходной сигнал на основании входного сигнала или набора входных сигналов. Функция активации отображает вычисленные значения в диапазоны, например, от 0 до 1 или от -1 до 1, в зависимости от выбранной функции.

Функции активации делятся на два основных типа:

1. Линейные функции активации,
2. Нелинейные функции активации.

Именно свойство нелинейности и отличает нейронные сети от большинства других моделей машинного обучения. Для того, чтобы данное свойство присутствовало, необходимо обеспечить нелинейность нейронов. Таким образом, выбор функций активации оказывает особенное влияние при проектировании нейронных сетей.

В большинстве случаев для экономии временных и вычислительных ресурсов функция активации задается сразу для всего слоя, а не отдельно для каждого нейрона.

2 Создание и обучение нейронной сети для определения тональности текста

2.1 Подготовка данных

Для успешной обработки текстовых данных программа должна получать на вход предварительно подготовленный и обработанный корпус текстов. Предварительная обработка состоит из последовательного выполнения следующих действий:

1. Разбиение текста на предложения.
2. Удаление знаков препинания и токенизация. С помощью регулярных выражений слова приводятся к нижнему регистру, удаляются знаки препинания и символы, не входящие в русский алфавит, происходит замена буквы «ё» на «е».
3. Лемматизация слов или стемминг. Это позволяет анализировать не разные формы одного слова, а производить поиск семантически близких слов. Под стеммингом подразумевается поиск основы слова, которая не всегда совпадает с корнем. Лемматизация — это процесс приведения слова к начальной форме. Для лемматизации в данной работе используется pymorphy2. Pymorphy2 — морфологический анализатор, разработанный на языке программирования Python. Он выполняет лемматизацию и анализ слов, способен осуществлять склонение по заданным грамматическим характеристикам слов.
4. Удаление «стоп-слов». Под «стоп-словами» подразумеваются такие, которые не несут особой смысловой нагрузки: частицы, предлоги, союзы, местоимения, числительные.

Итогом текущего раздела является функция подготовки текста, в которой выполняется токенизация, удаление знаков препинания, цифр, небуквенных символов и символов, не входящих в кириллицу, затем просматривается каждое слово, выполняется его лемматизация и добавляется к результирующей строке.

2.2 Модель Word2vec

2.2.1 Описание модели word2vec

После предварительной подготовки необходимо перевести полученные слова естественного языка в форму, которую способна принять на вход и

проанализировать нейронная сеть. Для этой цели выберем модель для анализа семантики естественных языков word2vec, разработанную в 2013 году группой исследователей компании Google под руководством Томаша Миколова.

Word2vec – набор алгоритмов для расчета векторных представлений слов, получающий на вход корпус текста и создающий на выходе набор векторов слов.

Принцип работы данной модели основан на идее о том, что лингвистические единицы, встречающиеся в похожих контекстах, являются семантически близкими, или означают похожие вещи.

В word2vec содержится два типа алгоритмов: на основе моделей CBOW (Continuous Bag of Words) и Skip-gram. Алгоритмы, основанные на CBOW пытаются предсказать слово по заданному контексту, и для хороших результатов применяются при выборке более ста миллионов слов. В архитектуре Skip-gram предсказывается контекст на основе заданного слова, и такая модель может показывать хорошие результаты на выборке менее ста миллионов слов. Для реализации модели была выбрана архитектура Skip-gram, которая, хоть и обладает меньшей скоростью обучению, но лучше работает на относительно маленьких выборках и учитывает редкие слова.

2.2.2 Обучение модели word2vec

Корпус текстов предварительно обрабатывается, а затем подается на вход модели, затем поочередно выполняются следующие действия:

1. Происходит считывание корпуса текстов, затем подсчитывается частота встречаемости в нем каждого слова.
2. Полученные слова образуют словарь, которые сортируются по вычисленной частоте встречаемости. Словарь обладает определенным размером, поэтому из него исключаются редкие слова.
3. Обработка происходит по субпредложению (это может быть как предложение исходного текста, так и целый абзац), модель проходит по ним окном определенного размера. Размер окна – максимальная длина между текущим словом и словом, которое предсказывается.
4. Данные, находящиеся в текущем окне попадают на вход нейронной сети прямого распространения.

Архитектура Skip-gram состоит из трех слоев:

1. Входной слой, который принимает одно слово в формате one-hot, то есть

входящее слово кодируется бинарным вектором, в котором единица стоит в месте, соответствующем позиции этого слова в словаре.

2. Embedding-слой, представляющий собой матрицу размерности $N \times P$, где N – размер словаря, а P – гиперпараметр, подбираемый эмпирическим способом.
3. Выходной слой размерности $N \times 1$, где N – размер словаря, с функцией активации softmax. [4]

В качестве модели word2vec была выбрана реализация, предоставляемая библиотекой Gensim, так как она предоставляет удобные инструменты для многопоточной обработки больших корпусов текстов, что позволяет производить обучение быстрее.

Для модели были выбраны следующие гиперпараметры:

1. Размерность векторного пространства – 300;
2. Размер окна – 10;
3. Константа субсэмплирования – 0.00006;
4. Количество эпох – 5

Модель была создана и обучена с помощью встроенных методов библиотеки Gensim.

В качестве обучающей выборки был взят корпус коротких текстов на русском языке на основе постов twitter Юлии Рубцовой [5].

2.3 Модель рекуррентной нейронной сети

2.3.1 Рекуррентные нейронные сети

Рекуррентные нейронные сети работают с последовательностями и списками. Они используют уже полученную информацию для решения будущих задач. В некоторых случаях для решения поставленной задачи требуются только последние из полученных данных. [6] Рекуррентные нейронные сетиправляются с задачами, в которых расстояние между необходимой информацией и местом ее использования для решения следующей задачи небольшое.

Но существуют ситуации, которые требуют больше контекстных данных, тогда расстояние между предыдущей информацией и местом, в котором она требуется увеличивается. С учетом увеличения такого расстояния связь между информацией теряется. Теоретически с этим можно справиться, если исследовать все зависимости и правильно подобрать необходимые параметры сети,

но существуют исследования, доказывающие фундаментальные ограничения рекуррентных нейронных сетей, а в следствие и неспособность решать такие задачи на практике. [7]

2.3.2 Сети LSTM

LSTM (long short-term memory, долгая краткосрочная память) — тип рекуррентной нейронной сети, который способен обучаться долгосрочным зависимостям.

LSTM созданы специальным образом для решения проблемы долгосрочных зависимостей. Они запоминают информацию в течение длительных периодов времени, поэтому их практически не нужно обучать.

Архитектура всех рекуррентных сетей имеет форму цепочки, состоящей из повторяющихся в цикле модулей сети. Если в стандартных нейронных сетях такие повторяющиеся модули имеют простую структуру, к примеру, только один слой, то в LSTM повторяющийся слой может состоять из четырех связанных слоев.

Именно LSTM будет использована в данной работе, но для начала нужно осуществить обработку данных. В качестве тренировочной и тестовой выборок так же использован корпус коротких текстов на русском языке на основе постов twitter Юлии Рубцовой. Текстовый корпус нужно обработать с помощью описанной выше функции предобработки текста.

Полученный обработанный корпус необходимо разделить на тренировочную и тестовую выборки, а затем разбить текст на токены, преобразовать этот текст в последовательность чисел в соответствии с полученным словарем для подачи на вход нейронной сети. Далее для того, чтобы на вход модели позитивные и негативные данные подавались в перемешку, перемешаем их. [8]

Далее необходимо загрузить предварительно обученную модель word2vec для того, чтобы инициализировать матрицу Embedding слоя весами, полученными при обучении модели word2vec.

После того, как необходимые данные подготовлены и обработаны, нужно создать и обучить модель рекуррентной нейронной сети с использованием рекуррентного слоя LSTM.

Результатом текущего раздела является обученная модель LSTM-сети для определения тональности текста.

3 Разработка автоматизированной системы поддержки принятия решений

3.1 Системы поддержки принятия решений

Система поддержки принятия решений — компьютерная автоматизированная система, оказывающая помощь людям, которым необходимо принимать объективные решения в сложных условиях, для полного анализа исследуемой области.

Система поддержки принятия решений используется для поддержки решений, принимающих во внимание множество различных показателей и критериев, в сложной информационной среде. Решение таких задач требует обработки и анализа большого объема данных и объективности результатов, чего не достичь без использования современных вычислительных приборов.

Для анализа в системе поддержки принятия решений используются различные методы:

1. информационный поиск;
2. интеллектуальный анализ данных;
3. поиск знаний в базах данных;
4. рассуждение на основе прецедентов;
5. эволюционные вычисления и генетические алгоритмы;
6. нейронные сети и другие. [9]

Часть из перечисленных методов была разработана с использованием искусственного интеллекта. Если в основе системы поддержки принятия решений лежат методы искусственного интеллекта, то ее называют интеллектуальной системой поддержки принятия решений.

3.2 Маркеры профориентации для оценки личностных и профессиональных качеств человека

В настоящее время проводится множество исследований в области влияния личностных качеств человека и его текстовым, графическим, видео и аудиоконтентом, информацией о количестве друзей, интересных сообществах, комментариях и других активностях в социальных сетях. Большинство исследований проводится на иностранных языках, в русскоязычных источниках только начинает развиваться данное направление.

Для извлечения требуемой информации из социальной сети Instagram

будем использовать модуль Instaloader для языка программирования Python, разработка которого организована на сервисе GitHub.

Психологами из Саратовского государственного университета были предоставлены лингвистические маркеры для анализа информации из социальных сетей. В настоящей работе будет использована часть из предоставленных лингвистических маркеров:

1. Использование в тексте названий кинофильмов различных жанров.

Для анализа использования названий кинофильмов необходимо составить пожанровые словари, на основании которых будет происходить анализ. После этого будем проверять публикации пользователя на наличие соответствий.

Каждый файл со словарем должен называться русским названием описанного в нем жанра. Имя файла в данном случае будет использоваться в качестве ключа в словаре, а его содержание — в качестве списка значений.

Затем будем искать соответствия элементам списков значений каждого жанра в каждой публикации. Таким образом, получим словарь, в котором ключами будут являться жанры кинофильмов, а значениями — количество употреблений названий фильмов каждого из жанров.

Словари можно добавлять, редактировать, удалять, в том числе добавлять словари с названиями компьютерных игр, музыкальных композиций, событийными фразами и тому подобным.

2. Частотность употребления местоимения «я» на странице пользователя.

Для определения частотности местоимения «я» будем использовать среднее количество употреблений на одну публикацию пользователя.

3. Частота публикаций пользователя.

Для получения частоты необходимо вычислить среднее время между публикациями записей. Для этого вычислим разницу между каждыми соседними публикациями и найдем среднее.

4. Использование URL-адресов.

Для подсчета количества используемых в публикации URL-адресов будем с помощью регулярных выражений заменять ссылки на метку «URL» и подсчитывать их число.

5. Количество отметок «нравится» на публикациях.

Для получения ответа на данный маркер определим среднее количество отметок «нравится» на публикациях пользователя.

6. Эмоциональная окраска излагаемого текста.

С помощью нейронной сети, созданной для определения тональности текста, проанализируем публикации пользователя и разделим их на две группы по эмоциональной окраске: позитивные и негативные. Для этого необходимо выполнить предварительную обработку текста и перевести его в форму, требуемую для передачи на вход нейронной сети.

7. Реакция сообщества (эмоциональная окраска комментариев к изложенному сообщению).

По описанному выше принципу выполняем предварительную обработку текста, а затем анализируем комментарии к каждой публикации и определяем их эмоциональную окраску. Таким образом, позитивные высказывания показывают поддержку сообщества, а негативные — наоборот.

3.3 Разработка интерфейса системы поддержки принятия решений

Интерфейс проекта будет создан с использованием модуля PySide2, который является проектом для создания привязки Qt к языку программирования Python.

PySide2 предоставляет доступ к программе QtDesigner, в которой можно создать собственный интерфейс. Для разработки системы необходимо три формы:

1. Первая форма нужна для ввода логина и пароля, требуемых для входа в личный или рабочий Instagram-аккаунт.
2. Вторая форма необходима для ввода идентификатора профиля и количества последних публикаций, анализ которых требуется провести.
3. В третьей форме будут отображаться результаты, полученные в результате анализа Instagram-профиля пользователя. Таким образом, получаемая информация будет содержать данные об имени пользователя, количестве его публикаций, подписчиков и подписок, частоте публикаций, среднем числе отметок «нравится», числе позитивных и негативных публикаций и реакции сообщества, частоте употребления местоимения «Я», количестве используемых в публикациях ссылок и употреблении названий кинофильмов различных жанров.

ЗАКЛЮЧЕНИЕ

В ходе выполнения настоящей работы были выполнены следующие задачи:

1. Проанализированы существующие модули, используемые для извлечения информации из социальной сети Instagram;
2. Изучены методы, используемые для предварительной подготовки текста, и была проведена соответствующая обработка используемого корпуса текстов для дальнейшего анализа;
3. Проанализирована работа модели word2vec для преобразования текстов естественного языка в векторную форму, и модель была обучена на основе корпуса коротких текстов на русском языке на основе постов twitter Юлии Рубцовой;
4. Рассмотрен принцип работы рекуррентных нейронных сетей и сетей LSTM и разработана нейронная сеть с использованием слоев LSTM для анализа тональности текстов;
5. Реализован интерфейс автоматизированной системы поддержки принятия решений.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 *C. Рашка, В. Мирджалили. Python и машинное обучение / В. Мирджалили С. Рашка.* — Москва: Вильямс, 2019. — С. 656.
- 2 *Николенко С. Архангельская Е., Кадурин А. Глубокое обучение. Погружение в мир нейронных сетей / Кадурин А. Николенко С., Архангельская Е.* — Питер, 2020. — С. 480.
- 3 Как работает нейронная сеть: алгоритмы, обучение, функции активации и потери [Электронный ресурс]. — URL: <https://neurohive.io/ru/osnovy-data-science/osnovy-nejronnyh-setej-algoritmy-obuchenie-funkcii-aktivacii-i-poteri> (Дата обращения 26.03.2021). Загл. с экрана. Яз. рус.
- 4 Word2Vec: как работать с векторными представлениями слов [Электронный ресурс]. — URL: <https://neurohive.io/ru/osnovy-data-science/word2vec-vektornye-predstavlenija-slov-dlya-mashinnogo-obuchenija/> (Дата обращения 10.05.2021). Загл. с экрана. Яз. рус.
- 5 Корпус коротких текстов на русском языке на основе постов твиттер RuTweetCorp Ю. Рубцовой [Электронный ресурс]. — URL: <https://study.mokoron.com/> (Дата обращения 08.05.2021). Загл. с экрана. Яз. рус.
- 6 Введение в рекуррентные нейронные сети [Электронный ресурс]. — URL: https://proprogs.ru/neural_network/vvedenie-v-rekurrentnye-neyronnye-seti (Дата обращения 01.05.2021). Загл. с экрана. Яз. рус.
- 7 LSTM — нейронная сеть с долгой краткосрочной памятью [Электронный ресурс]. — URL: <https://neurohive.io/ru/osnovy-data-science/lstm-nejronnaja-set/> (Дата обращения 14.05.2021). Загл. с экрана. Яз. рус.
- 8 Обзор систем анализа тональности текста на русском языке [Электронный ресурс]. — URL: <https://moluch.ru/archive/47/5951/> (Дата обращения 12.05.2021). Загл. с экрана. Яз. рус.
- 9 Система поддержки принятия решений [Электронный ресурс]. — URL: <https://cyberleninka.ru/article/n/sistema-podderzhki-prinyatiya-resheniy/viewer> (Дата обращения 12.04.2021). Загл. с экрана. Яз. рус.