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

Кафедра информатики и программирования

ПРИМЕНЕНИЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ОБРАБОТКИ АНКЕТНЫХ ДАННЫХ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Волохиной Юлии Алексеевны

Научный руководитель:
ст. преп. кафедры ИиП _____ А.А. Казачкова

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Саратов 2021

ВВЕДЕНИЕ

В современном мире все большее количество процессов переходит в онлайн сферу, так как практически все люди сейчас имеют свободный доступ к интернету. Произошедшая пандемия также подчеркнула необходимость в использовании альтернатив для привычных услуг, которые люди привыкли получать, находясь непосредственно в месте ее оказания. Многие компании начали создавать новые сервисы и добавлять дополнительные возможности, необходимые в сложившейся ситуации, а люди открыли для себя удобные пути решения привычных проблем и задач.

Такой переход на онлайн-альтернативы значительно упрощает использование уже привычных услуг из-за отсутствия лишних действий – можно оформить заказ, оплатить счет, зарегистрироваться где-либо в любое время, вне зависимости от местоположения. Для этого больше не нужно идти в определенное место и заполнять все документы вручную на бумаге, ведь все необходимое всегда есть под рукой – будь то компьютер, ноутбук или смартфон.

Чаще всего при работе с онлайн-сервисами, предоставляющими услуги, человек сталкивается с заполнением полей данными на различных формах и анкетах. В таких случаях существует большая вероятность неправильного ввода в виде опечаток, пропущенных букв и случайном вводе лишних символов, что в дальнейшем может усложнить работу с полученными данными или получить некорректную интерпретацию и итоговый результат. Таким образом, возникает вопрос об обработке и анализе введенных данных для получения необходимого правильного написания.

В работе были рассмотрены методы машинного обучения и выбран один из его разделов – глубокое обучение. За последние годы данное направление начало активно развиваться – на его основе создаются новые технологии, используемые как в крупнейших компаниях, так и в стартапах. Области применения машинного обучения могут быть совершенно

разнообразными, необходимо лишь иметь данные для изучения и дальнейшей работы с ними.

Цель бакалаврской работы – изучение возможностей машинного и глубокого обучений, создание и обучение моделей машинного обучения и модели нейронной сети для классификации анкетных данных, содержащих наименования школ города Саратова.

Поставленная цель определила **следующие задачи**:

1. Изучить и применить различные методы машинного обучения для работы с данными.
2. Провести сравнительный анализ полученных результатов после работы с методами машинного обучения.
3. Изучить общий принцип работы нейронных сетей.
4. Рассмотреть и выбрать сеть для многоклассовой классификации текстовых данных.
5. Подготовить данные для обучения нейронной сети.
6. Расширить (аугментировать) данные, добавив модификации официальных наименований из словаря.
7. Сформировать архитектуру для модели нейронной сети.
8. Обучить модель на данных, подготовленных различным образом.
9. Визуализировать процесс обучения моделей.
10. Применить полученную модель для организации вывода предсказанных значений в Google Таблицах.

Методологические основы машинного и глубокого обучений и работы с нейронными сетями представлены в работах П. Флаха, Х. Бринка, А. Мюллера, П. Дж. Вандера, Я. Гудфеллоу, А.В. Созыкина, Ш. Франсуа, Дж. Паттерсона, С. Хайкина и Р. Тадыусевича.

В теоретической части бакалаврской работы были рассмотрены методы машинного и глубокого обучений для решения задачи классификации текстовых данных, а также изучены способы

предварительной обработки данных и способы встраивания дополнительной функциональности в Google Таблицы.

В практической части бакалаврской работы были созданы модели машинного обучения и модели нейронной сети для определения официального наименования школы по введённому пользователем тексту, далее эти модели были обучены на подготовленных данных. Также была произведена оценка точности предсказания в зависимости от способа подготовки и визуализирован процесс обучения с дальнейшим добавлением полученной модели в работу с Google таблицами.

Структура и объём работы. Бакалаврская работа состоит из введения, 5 разделов, заключения, списка использованных источников и 6 приложений. Общий объем работы – 86 страниц, из них 67 страниц – основное содержание, включая 26 рисунков, список использованных источников информации – 33 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Машинное обучение» посвящен изложению теоретического материала о технологиях, инструментах и методах, используемых в машинном обучении.

Главная цель машинного обучения – это поиск закономерностей и предсказание итога по входным данным, успех анализа которых заключается в наибольшем разнообразии данных, подающихся на исследование.

Раньше для решения задач требовалась разработка правил принятия решений самостоятельно. На основании конструкций «if» и «else» происходил жесткий отбор данных, который мог подходить не просто под одну конкретную область, а под конкретную задачу, малейшее изменение которой могло привести к переписыванию большой части кода. Однако сформулировать подходящий для задачи набор правил можно далеко не в любой ситуации. Благодаря машинному обучению, данная проблема нашла решение – теперь можно подать на вход алгоритму большое количество данных, а он уже самостоятельно сможет выделить необходимые для решения признаки

Существует два вида обучения:

1. обучение с учителем – на вход алгоритму пользователь передает данные в формате пар «объект-ответ», после чего алгоритм самостоятельно находит способ получения ответа для каждого объекта из выборки данных. Основная задача – это построение такого алгоритма, который сможет находить ответ и проставлять метки для новых данных, до этого неизвестных

Алгоритмы обучения с учителем:

- метод -ближайших соседей;
- линейные модели;
- наивные байесовские классификаторы;
- деревья решений;
- ансамбли деревьев решений;

- ядерный метод опорных векторов;
- глубокое обучение.

2. обучение без учителя – это вид алгоритмов, которым на вход подаются данные без каких-либо меток, позволяющих предсказать ответы, в них известны только объекты, произведя анализ которых можно получить новые знания.

Второй раздел «Глубокое обучение» описывает один из разделов машинного обучения, содержащий в себе новый подход к поиску представления данных, – глубокое обучение. Он делает упор на изучение последовательных слоев все более значимых представлений. Под глубиной модели в глубоком обучении подразумевается количество слоев, на которые делится модель данных. В то время как другие подходы к машинному обучению используют для изучения только один-два слоя представления данных, современное глубокое обучение может вовлекать в процесс обработки, как десятки, так и сотни последовательных слоев представления, которые могут определяться автоматически под воздействием обучающих данных.

В глубоком обучении подобные многослойные представления изучаются с помощью нейронных сетей – моделей, структурированных в виде слоев, наложенных друг на друга.

Если искусственная нейронная сеть состоит из большого количества нейронов, то самое удобное их представление будет в качестве нескольких слоев, каждый из которых может содержать некоторое число нейронов. Для каждого вида нейронов есть соответствующий вид слоев нейросети:

- входной слой – принимает сигналы из окружающей среды;
- выходной слой – передает сигналы обратно, в окружающую среду;
- скрытые слои – это слои, находящиеся между входным и выходным слоями, которые обрабатывают передающиеся им входные сигналы.

Третий раздел «Применение методов машинного обучения к обработке анкетных данных» посвящен созданию моделей и дальнейшей работе с ними с помощью выбранных алгоритмов машинного обучения для решения задачи классификации текстовых данных.

Для работы были выбраны следующие алгоритмы:

1. Алгоритм KNN, построенный на сходстве объектов и способный выделять среди всех данных k известных объектов, которые похожи на ранее неизвестный объект. Далее на основании классов ближайших соседей происходит принятие решения для нового объекта. Одной из ключевых задач алгоритма считается подбор коэффициента k, количества записей, которые будут считаться близкими.
2. Построение дерева решений. С помощью деревьев решений выстраивается иерархия правил, которая отвечает на заданные вопросы в форме «если …, то …» и приводит к итоговому решению поставленной задачи. Процесс построения деревьев решений заключается в последовательном, рекурсивном разбиении обучающего множества на подмножества с помощью решающих правил в узлах. Этот процесс продолжается до тех пор, пока все узлы в конце всех ветвей не будут признаны листьями. Узел становится листом как естественным образом, при содержании в нем единственного объекта или объектов только одного класса, так и при достижении определенного условия остановки, которое может задаваться непосредственно самим пользователем.
3. Построение случайного леса, где случайный лес – это набор деревьев решений, где каждое дерево немного отличается от остальных. Основная идея этого метода заключается в том, что каждое дерево может довольно хорошо прогнозировать, но, скорее всего, переобучаться на части данных. Поэтому если

построить много таких деревьев, которые будут хорошо работать и переобучаться с разной степенью, то можно будет уменьшить переобучение путем усреднения полученных результатов. В процесс построения деревьев вносится случайность, основная цель которой – это обеспечение уникальности каждого дерева.

После получения и сравнения результатов применения методов классификации между собой, было получено:

- методы построения дерева решений и ансамблей деревьев дают очень похожий результат, который немного превосходит метод - ближайших соседей;
- одним из недостатков является отсутствие исправления опечаток, что дает дополнительные проблемы при токенизации слов;
- пропущенные пробелы не мешают разделению слов, что приводит к увеличению объема словаря не уникальными словами.

Таким образом, получив удовлетворительные результаты при работе с методами машинного обучения, работа с имеющимися данными была продолжена, но используя глубокое обучение, с применением искусственных нейронных сетей для получения более хороших результатов работы.

Четвертый раздел «Применение методов глубокого обучения к обработке анкетных данных» посвящен созданию и обучению моделей искусственной нейронной сети на текстовых данных, обработанных различным способом.

После подготовки данных и их разбиения на обучающую, проверочную и тестовую выборки происходил процесс обучения моделей – сначала на исходных данных, потом – на данных с предварительной обработкой.

Если сравнивать первую и последнюю модели, можно увидеть, как поменялись значения качества обучения и ошибки. Если в первой модели итоговое качество модели на обучающем наборе данных было равно 0.9903, то во второй дополненной модели – уже 0.9993, а на тестовом наборе –

0.9954 и 0.9974. Аналогично и уменьшилось значение ошибок на всех наборах данных.

Также была написана функция предсказания, принимающая на вход строку, содержащую любую вариацию написания школы, и выдающую по ней итоговый результат с правильным написанием.

Пятый раздел «Организация работы модели с Google Таблицами» описывает работу с полученной моделью нейронной сети с помощью дополнительных платформ для удобства в использовании.

Google Apps Script – это платформа разработки от компании Google, созданная для разработки приложений. С ее помощью можно быстро и легко создавать проекты, взаимодействующие со всеми сервисами этой компании, такими как почта, календарь, диск, таблицы, документы.

Также для взаимодействия работы таблиц через Apps Script с полученной моделью нейронной сети была использована облачная платформа Heroku, основанная на управляемой контейнерной системе. Она поддерживает многие современные языки программирования, в том числе Python, Java, PHP.

Таким образом в Google Таблицах был создан новый пункт меню, по нажатию на который происходит запуск работы нейронной сети, которая выводит предсказанные значения с правильным написанием школ для выделенного диапазона ячеек.

ЗАКЛЮЧЕНИЕ

В результате выполнения данной работы были получены следующие результаты:

1. Изучены основные понятия из теории машинного и глубокого обучения, а также технологии и возможности их применения в различных областях современного мира.
2. Рассмотрены и применены на практике несколько методов машинного обучения с произведением дальнейшего сравнительного анализа полученных результатов среди всех моделей.
3. Рассмотрен такой универсальный метод глубокого обучения как нейронная сеть, применяющийся к задачам из различных сфер жизни.
4. Был создан собственный набор данных для решения поставленной задачи, с его последующей обработкой и подготовкой к применению.
5. Было произведено обучение моделей нейронной сети на данных, подготовленных различным образом, с их дальнейшим сравнительным анализом и визуализацией обучения.
6. И в заключительном этапе обученная модель нейронной сети была применена к задаче предсказания правильных результатов на основании введенных значений.
7. После этого было также оформлено использование полученной модели нейронной сети для работы с данными, используя Google таблицы.

Таким образом, в результате изучения новых технологий с различными возможностями их применения были созданы различные модели машинного обучения, а также создана и обучена модель нейронной сети, способная предсказывать исправленные варианты написания школ города Саратова на основании введенных данных.

Полученная модель нейронной сети была встроена в работу с удобным вспомогательным сервисом Google таблицы, что является большим преимуществом для ее использования простыми пользователями. С помощью

Google Apps Script и платформы Heroku было дополнено приложение, способное запускать работу модели, которая считывает выделенные поля таблицы, передает полученные значения в модель и выдает новые, правильные написания для каждого поля.

Основные источники информации:

1. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. — М.: «ДМК Пресс», 2015. — 400 с.
2. Бринк Хенрик, Ричардс Джозеф, Феверолф Марк. Машинное обучение. — СПб.: «Питер», 2017. — 336.
3. Андреас Мюллер. Сара Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. — Москва: Издательский центр "Гевисста", 2017. — 393 с.
4. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — СПб.: «Питер», 2018. — 576 с.
5. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. — 2-е изд., испр. — М.: ДМК Пресс, 2018. — 652 с.
6. А.В. Созыкин. «Обзор методов обучения глубоких нейронных сетей». — «Вестник ЮУрГУ», 2017. — 59 с.
7. Шолле Франсуа. Глубокое обучение на Python. — СПб.: «Питер», 2018. — 400с.
8. Паттерсон Дж., Гибсон А. Глубокое обучение с точки зрения практика. — М.: «ДМК Пресс», 2018. — 418 с.
9. Хайкин С. Нейронные сети: полный курс, 2-е издание. — М.: Издательский дом «Вильямс», 2006. — 1104 с.
10. Р. Тадыусевич. Элементарное введение в технологию нейронных сетей. — М.: «Телеком», 2011. — 408 с.