

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

**Разработка прототипа экспертной системы по классификации
научных текстов на основе машинного обучения**

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы

направления 02.04.03 Математическое обеспечение и

администрирование информационных систем

факультета компьютерных наук и информационных технологий

Суровягина Дмитрия Павловича

Научный руководитель:

доцент, к. ф.-м. н.

_____ Савина К.П.

подпись, дата

Зав. кафедрой:

доцент, к. ф.-м. н.

_____ Огнева М.В.

подпись, дата

Саратов 2020

ВВЕДЕНИЕ

Актуальность темы. Работа посвящена изучению возможностей интеллектуальной обработки текстов с помощью алгоритмов машинного обучения, в частности — проблеме классификации текстов научной тематики. Широко известно, что научно-технический прогресс увеличил масштабы познавательной деятельности людей. Это выразилось в развитии новых научных направлений, изменении производственных стандартов и появлении инновационных способов обработки больших данных.

Глубокие нейронные сети можно охарактеризовать как технологическую инновацию, отвечающую трем основным критериям инновационного продукта: научно-техническая новизна, промышленная применимость, коммерческий успех. Эта технология реализует усовершенствованный процесс обучения компьютера способам представления данных и успешно решает одну из фундаментальных задач комбинаторной оптимизации — проблему присвоения значения.

В течение последних лет глубокое обучение достигло огромного успеха в решении широкого круга задач, связанных с извлечением полезной информации из изображений, текстов, видео, звуков и других данных. Многие специалисты полагают, что глубокое обучение в последующие несколько лет будет только набирать популярность, поэтому сегодня это один из самых актуальных предметов для изучения в области компьютерных наук.

Цель бакалаврской работы — создание на основе технологии глубоких нейронных сетей прототипа приложения, способного решать задачу многоклассовой классификации текстов научной тематики.

Поставленная цель определила следующие **задачи**:

- 1) изучить основные библиотеки языка Python, используемые для глубокого обучения и обработки естественного языка;
- 2) сконструировать модель рекуррентной нейронной сети, способной решать задачу однозначной многоклассовой классификации текстов;

- 3) сформировать корпус текстовых данных из аннотаций научных статей для последующей его векторизации и обработки с помощью полученной модели;
- 4) классифицировать аннотации научных статей, оценить результаты работы и интегрировать модель классификатора в рабочее веб-приложение.

Методологические основы исследования «Разработка прототипа экспертной системы по классификации научных текстов на основе машинного обучения» представлены в работах Б. Бенгфорта и Р. Билбро [1], Я. Гудфеллоу и И. Бенджио [2], О. Жерона [3], Р. Митчелла [4], С. Николенко и А. Кадурина [5], Дж. Пласа [6], С. Рашки и В. Мирджалили [7], Б. Шардена и Л. Массарона [8], Ф. Шолле [9] и других ученых.

Теоретическая значимость магистерской работы заключается в том, что проблема компьютерной обработки естественного языка с помощью алгоритмов машинного обучения была исследована в новой предметной области, а именно: применительно к текстам научной тематики. Большинство существующих на рынке программного обеспечения приложений и модулей для обработки языка ориентированы на интернет-магазины, новостные ленты и социальные сети, однако в настоящей работе рассматривается возможность анализа научных текстов, что является актуальной задачей в виду экспоненциального роста научных публикаций в последние годы.

Практическая значимость магистерской работы состоит в том, что (а) был сформирован оригинальный обучающий набор из 300 тыс. аннотаций научных статей, полученных из архива электронных публикаций arXiv.org; (б) построена уникальная модель рекуррентной нейронной сети с LSTM-слоем, способная обучиться решению задачи классификации на данном наборе; (в) обученная модель встроена в функциональное веб-приложение, предоставляющее пользователю возможность классифицировать новые аннотации и играющее, таким образом, роль экспертной системы, ориентированной на работу с научными текстами.

Структура и объём работы. Магистерская работа состоит из введения, четырех разделов, заключения, списка использованных источников и десяти приложений. Общий объем работы – 97 страниц, из них 74 страницы – основное содержание, включая 19 рисунков и 1 таблицу, список использованных источников информации – 47 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Компьютерная обработка естественного языка» посвящен задаче классификации текстовых данных и выбору программных инструментов, используемых для обработки естественного языка. Обработка естественного языка — это чрезвычайно сложная задача в компьютерных науках. Символьные структуры языков содержат множество зависимостей, которые варьируются при переходе от одного языка к другому. Однако извлечение значимой информации из текстов на естественном языке — также и крайне важная задача любой интеллектуальной системы, взаимодействующей с человеком. Чтобы эффективно решить эту задачу, ученый, проектирующий приложение обработки естественного языка, разбивает тексты на грамматические формы: предложения, фразы, слова и части речи, используя специальные алгоритмы.

На сегодняшний день парадигма машинного обучения предоставляет наиболее эффективные алгоритмы для разбиения, предварительного анализа и преобразования текстов. Классификация — это основная и наиболее распространенная форма анализа текста, в основе которой лежит идея изучить существующие связи между экземплярами некоторого множества и целевой категориальной переменной. Цель классификации известна заранее, поэтому ее алгоритмы относятся к области машинного обучения с учителем.

Модель классификации обучают для минимизации ошибки между предсказанными и фактическими категориями данных. После обучения классификационная модель должна быть способна присваивать категории новым экземплярам, опираясь на шаблоны, выявленные в процессе обучения.

Второй раздел «Анализ текстов с помощью глубоких нейронных сетей» посвящен рассмотрению глубоких нейронных сетей как специфического средства для анализа текстов и, в частности, для решения основной задачи работы: классификации аннотаций научных статей. Базовым элементом глубокого обучения является многослойная нейронная сеть прямого распространения (feedforward neural network, FNN). FNN можно понимать как многослойный персептрон, который имеет три слоя: входной, выходной и скрытый. Если такая сеть имеет более одного скрытого слоя, то ее называют глубокой. Количество слоев, присутствующих в модели, называется глубиной обучения. Искусственные нейронные сети, состоящие из множества слоев, сегодня повсеместно используются не только в науке, но и в производстве такими компаниями, как Facebook, Microsoft и Google.

Экспериментально показано, что многослойные персептроны не эффективны при обработке таких данных, в которых порядок следования имеет значение. Особенность последовательности как типа данных состоит в том, что ее элементы не независимы друг от друга, а расположены в определенном порядке. Чтобы решить проблему обработки последовательностей была спроектирована рекуррентная нейронная сеть (Recurrent Neural Network, RNN), которая быстро доказала свою эффективность в решении таких задач, как анализ или порождение текста.

Однако простые рекуррентные сети плохо справляются с ситуациями, когда нужно надолго сохранять информацию: влияние скрытого состояния или входа на последующие состояния рекуррентной сети экспоненциально затухает. В литературе этот феномен называется проблемой затухания градиента. Решения, которые на данный момент предлагаются в индустрии глубокого обучения, состоят в том, чтобы изменить и усложнить архитектуру базового элемента RNN. Вместо единственного числа, на которое влияют все последующие состояния, предлагается сконструировать специального вида ячейку, в которой можно явным образом моделировать «память» модели. Одна из самых популярных конструкций таких ячеек — LSTM (Long

ShortTerm Memory, то есть «долгая краткосрочная память»). Экспериментально показано, что LSTM-ячейки являются весьма гибким и эффективным инструментом для анализа последовательностей.

Третий раздел «Формирование корпуса текстовых данных» посвящен изучению способов формирования специализированного корпуса текстовых данных для машинного обучения. Корпус — это коллекция взаимосвязанных документов на естественном языке. Корпусы могут быть аннотированными (текст или документы могут быть снабжены специальными метками) для алгоритмов обучения с учителем, или неаннотированными для тематического моделирования и кластеризации документов.

Модели интеллектуальной обработки языка, обученные на ограниченной предметной области, часто действуют лучше, чем такие же модели, но обученные на обобщенном корпусе. Дело в том, что в разных предметных областях используется разный язык (специальные термины, выражения и специфический синтаксис), поэтому корпус, специализированный для конкретной области, анализируется точнее, чем обобщенный корпус.

Также в данном разделе рассматриваются методы векторизации текстовых данных. Преобразование текстовых документов в численный вид дает возможность осуществлять их анализ и создавать экземпляры, с которыми смогут работать алгоритмы машинного обучения. В анализе текста экземплярами являются целые документы или высказывания, которые могут иметь самые разные размеры, от коротких цитат до целых книг. Но сами векторы всегда имеют одинаковую длину. Каждое свойство в векторном представлении — это признак. В случае с текстом признаки представляют атрибуты и свойства документов. Признаки документа описывают многомерное пространство признаков, к которому могут применяться методы машинного обучения.

Четвертый раздел «Классификация аннотаций научных статей»

содержит описание авторского приложения классифицирующего аннотации научных статей. С помощью инструмента Baleen и программы-обработчика html-страниц был сформирован предметно-ориентированный корпус из аннотаций, представленных на сайте arXiv.org. Из исходного корпуса в каждую из десяти категорий было отобрано 30 тыс. аннотаций, снабженных меткой своей категории. Итоговый обучающий набор, содержащий 300 тыс. аннотаций с метками, представляет собой файл в формате csv.

Для обучения модели рекуррентной нейронной сети с LSTM-слоем используются такие библиотеки Python, как NumPy, Scikit-Learn и Pandas, фреймворки для глубокого обучения TensorFlow и Keras, а также модуль matplotlib для построения графиков, иллюстрирующих точность обучения. Модель обучается на 240 тыс. аннотаций и проверяется на 60 тыс. на протяжении 12 эпох, достигая точности 91% на этапе обучения и 87% на этапе проверки. Это хороший результат для задачи многоклассовой классификации, который в значительной степени объясняется меньшей подверженностью LSTM проблеме затухания градиента. Увеличить точность модели можно было бы за счет уменьшения количества категорий в обучающем наборе или увеличения числа аннотаций в каждой категории, но, по-видимому, не за счет увеличения числа эпох обучения, потому что модель достигает переобучения на 10 эпохе.

Полученная модель классификации аннотаций встраивается в веб-приложение на основе фреймворка Flask. Листинги соответствующих программ представлены в Приложениях Ж, З, И работы. Скриншоты работающего веб-приложения представлены в Приложении К.

ЗАКЛЮЧЕНИЕ

Таким образом, в работе «Разработка прототипа экспертной системы по классификации научных текстов на основе машинного обучения» были исследованы модели и алгоритмы глубокого обучения, используемые в

современных программных продуктах для анализа и интеллектуальной обработки текстов на естественном языке. Цель работы — создание прототипа приложения, способного решать задачу многоклассовой классификации текстов научной тематики — была достигнута. По результатам работы можно сформулировать следующие выводы.

1. Данные, полученные на естественном языке, несмотря на отсутствие машиночитаемой структуры, не являются случайными. Они подчиняются лингвистическим правилам, которые не только делают эти данные понятными для людей, но и позволяют организовать взаимодействие человека и компьютера.

2. Модели глубокого обучения, специально разработанные для текстов и последовательностей, на сегодняшний день способны формировать понимание естественного языка в простейшей форме, достаточной для решения таких задач, как классификация документов, анализ эмоциональной окраски, идентификация автора и даже получение ответов на вопросы в ограниченном контексте.

3. Для анализа текстов научной тематики необходим большой предметно-ориентированный корпус статей. Пять основных этапов обработки текста (извлечение содержания, выделение абзацев, предложений и лексем, маркировка лексем тегами частей речи и векторизация данных) осуществляются сегодня с помощью эффективных и простых библиотек языка Python.

4. Современное приложение для анализа естественного языка должно содержать модель машинного обучения, способную воспринимать полученные от пользователя данные, открывать новые пространства решений и непрерывно развиваться по мере поступления новой информации. Условия для такого развития можно создать с помощью веб-интерфеса, играющего роль экспертной системы.

Отдельные положения магистерской работы были представлены на конференциях:

- Десятая научная конференция «Presenting Academic Achievements to the World» (Саратов, Саратовский национальный исследовательский государственный университет им. Н.Г. Чернышевского, 16 апреля 2019);
- Первая международная научная конференция «Проблемы и вызовы цифрового общества: тенденции развития правового регулирования цифровых трансформаций» (Саратов, Саратовская государственная юридическая академия, 17–18 октября 2019 г.);

Тезисы выступлений на конференциях были опубликованы в сборниках научных работ:

- Суровягин, Д.П. Интеллектуальная обработка текста с помощью глубоких нейронных сетей: основные проблемы и результаты // Проблемы и вызовы цифрового общества: тенденции развития правового регулирования цифровых трансформаций: сб. научн. тр. по матер. I Междунар. научн.-практ. конф. (Саратов, 17–18 октября, 2019) / Д.П. Суровягин; под. ред. Н.Н. Ковалевой. Саратов: Изд-во ФГБОУ ВО «Саратовская государственная юридическая академия», 2019. с. 87–91.
- Surovyagin, D.P. Using Recurrent Neural Network for Analyze Text Data // Представляем научные достижения миру. Естественные науки: материалы X научной конференции молодых ученых «Presenting Academic Achievements to the World» / Д.П. Суровягин; под ред. С.А. Шиловой и др. Саратов: Саратовский источник, 2020. Вып. 9. с. 124–132.

Основные источники информации:

1. Бенгфорт, Б. Билбро, Р., Охеда, Т. Прикладной анализ текстовых данных на Python / Б. Бенгфорт и др.; пер. с англ. А. Киселева; под ред. К. Тульцева. — СПб.: Питер, 2019. — 368 с.

2. Гудфеллоу, Я., Бенджио, И., Курвилль, А. Глубокое обучение / Я. Гудфеллоу и др.; пер. с англ. А.А. Слинкина; под ред. Д.А. Мовчана. — 2-е изд., испр. — М.: ДМК Пресс, 2018. — 652 с.
3. Жерон, О. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow / О. Жерон; пер. с англ. и ред. Ю.Н. Артеменко. — СПб.: Альфа-Книга, 2018. — 688 с.
4. Митчелл, Р. Скрапинг веб-сайтов с помощью Python / Р. Митчелл; пер. с англ. А.В. Груздева; под ред. А.Н. Киселева. — М.: ДМК Пресс, 2016. — 280 с.
5. Николенко, С., Кадури, А., Архангельская, Е. Глубокое обучение: погружение в мир нейронных сетей / С. Николенко и др.; под ред. Н. Гринчика. — СПб.: Питер, 2019. — 480 с.
6. Плас, Дж. В. Python для сложных задач: наука о данных и машинное обучение / Дж. В. Плас; пер. с англ. И. Пальти; под ред. Н. Гринчика. — СПб.: Питер, 2018. — 576 с.
7. Рашка, С., Мирджалили, В. Python и машинное обучение / С. Рашка, В. Мирджалили; пер. с англ. Ю.Н. Артеменко; под ред. С.Н. Тригуба. — 2-е изд., перераб. и доп. — СПб.: Диалектика, 2019. — 656 с.
8. Шарден, Б., Массарон, Л., Боскетти, А. Крупномасштабное машинное обучение вместе с Python / Б. Шарден и др.; пер. с англ. А.В. Логунова; под ред. Д.А. Мовчана. — М.: ДМК Пресс, 2018. — 358 с.
9. Шолле, Ф. Глубокое обучение на Python / Ф. Шолле; пер. с англ. А. Киселева; под ред. К. Тульцева. — СПб.: Питер, 2018. — 400 с.