Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.032.26

Витковский
Артем Викторович

Применение нейронных сетей для анализа отзывов пользователей

АВТОРЕФЕРАТ
на соискание академической степени
магистра информатики и вычислительной техники

по специальности 1-40 81 04 – Обработка больших объемов информации

Научный руководитель
Жвакина А.В.
кандидат технических наук, доцент

Минск 2019

КРАТКОЕ ВВЕДЕНИЕ

Современное общество все больше и больше использует удобства, предоставляемые всемирной паутиной. В наше время уже сложно представить, как можно обойтись без использования сети Интернет. Все больше привычных и необходимых людям действий, например, общение, поиск информации, совершение финансовых операций, совершается именно там. Так общение по телефону и SMS уступает место социальным сетям и программам для обмена текстовыми сообщениями, большинство банковских услуг можно получить, используя мобильный телефон, подключенный к сети Интернет, а новости, несмотря на то, что газеты и журналы все еще существуют, всё больше людей получает именно из интернет-изданий.

Кроме этого, интернет позволил людям легко делится своим мнением. Многие новостные порталы позволяют пользователям оставлять комментарии к новостям. Или, например, пользователям часто разрешено оставлять отзывов в интернет-магазинах к конкретным товарам. Особенно можно выделить социальные сети, где люди делятся своим мнениям по разным поводам не только со знакомыми им людьми в личной переписке, но также оставляют публичные высказывания.

Эти мнения людей интересны специалистам, занимающимся различным продвижением брендов, идей, или рекламой. Появляются даже специальные профессии такие, как бренд-менеджер, специалист по продвижению в социальных медиа. Этим профессиям необходимо изучать результаты своей деятельности. Это достаточно сложно для одного человека, ведь просто просматривая случайные сообщения сложно получить правильную выборку, которая максимально точно описывает положение дел. Лучше всего иметь некие численные характеристики, которые понятны и отображают результаты работы. Но в таком случае встают закономерные вопросы: как получить такие данные и как их обрабатывать.

Популярным направлением в обработке естественных языков является использование алгоритмов машинного обучения. Машинное обучение (Machine Learning, ML) – это научное исследование алгоритмов и статистических моделей, которые используются компьютерными системами для эффективного выполнения конкретной задачи без использования явных инструкций, вместо этого опираясь на шаблоны и умозаключения.

В рамках данной работы рассмотрена возможность применения искусственных нейронных сетей для решения задачи анализа тональности текста, а также разработана система по типу микросервисного веб-приложения для сбора и анализа отзывов по заданной пользователем теме.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является исследование эффективности различных архитектур нейронных сетей в задаче анализа тональности текста и разработка программного обеспечения, демонстрирующая применимость разработанных моделей для решения практических задач.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Исследовать различные виды архитектур нейронных сетей.
2. Определить применимость и выбрать наиболее подходящие модели НС для решения задачи анализа тональности текста.
3. Разработать конкретные модели нейронных сетей, определить обучающее множество, реализовать разработанные модели, произвести их обучение.
4. Сравнить результаты работы моделей НС и определить наилучшую.
5. Разработать веб-приложение на основе современных подходов и технологий, использующее и демонстрирующее эффективность разработанной нейронной сети.

Объектом исследования является классификатор на основе нейронной сети, определяющий тональность текста.

Предметом исследования является программная реализация моделей нейронных сетей на основе CNN и LSTM для решения задачи анализа тональности текста.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность применения моделей сверточных нейронных сетей и сетей на основе долгой краткосрочной памяти для определения тональности текста. Модели нейронных сетей могут использоваться как часть программного комплекса веб-приложения.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично.

Опубликованность результатов диссертации

По теме диссертации опубликовано 2 печатные работы в сборниках трудов и материалов научных конференций.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трёх глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлен анализ предметной области, выявлены основные существующие проблемы в рамках тематики исследования, показаны направления их решения. Вторая глава посвящена исследованию конкретных моделей нейронных сетей для решения поставленной задачи. В третьей главе описан эксперимент по реализации и обучению сверточной нейронной сети и нейронной сети с долгой краткосрочной памятью, произведен анализ и сравнение результатов обучения, описана реализация web-приложения, использующего разработанную нейронную сеть.

Общий объем работы составляет 78 страниц, из которых основного текста – 54 страниц, 30 рисунков на 16 страницах, список использованных источников из 32 наименований на 2 страницах и 7 приложений на 11 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** был проведен анализ предметной области, определены виды социальных сетей Подходящей для проведения исследования является социальная сеть Twitter. Twitter – это онлайн-сервис новостей и социальная сеть, в которой пользователи публикуют и общаются с сообщениями, известными как «твиты». Первоначально твиты были ограничены 140 символами, но позже этот предел был удвоен до 280 для всех языков, кроме китайского, японского и корейского. Зарегистрированные пользователи могут публиковать, одобрять и пересылать публичные сообщения, но незарегистрированные пользователи могут только читать их. Пользователи получают доступ к Twitter через интерфейс своего веб-сайта или прикладное программное обеспечение для мобильных устройств.

Рассмотрена направление обработки естественного языка, выделены основные задачи, решаемые в рамках этой области. Рассмотрена задача определения тональности текста и методы ее решения.

Изучены и описаны общая модель формального нейрона, модель искусственной нейронной сети и ее виды. Особенности искусственных нейронных сетей позволяют применять их в различных задачах обработки естественного языка, определения тональности текста в частности.

Среди рассмотренных архитектур нейронных сетей наилучшим потенциалом обладают сверточные нейронные сети и сети с долгой краткосрочной памятью. Первый тип отлично работает в решении задачи выделения признаков. Второй тип используется в задачах из области обработки естественного языка, архитектура позволяет организовать работы с текстом, а также данный тип значительно упрощает всю модель.

Во **второй главе** были рассмотрены различные наборы данных использующиеся для обучения нейронных сетей определять тональность текста. Популярный набор данных IMDB Movies Review подходит для реализации эксперимента по обучению классификаторов на основе нейронной сети и сравнения эффективности их работы.

Рассмотрены различные варианты представления слов естественного языка в задачах машинного обучения: поиск по словарю, мешок слов, векторные представления. Было выявлено что векторные представления слов являются наиболее интересным и перспективным вариантом представления текстовых документов в большинстве задач обработки естественного языка, и определения

тональности текста в частности. Для получения самих векторов используются статистические модель CBOW и Skip-gram, а также алгоритм GloVe.

Среди рассмотренных видов моделей были определены две наиболее перспективные. Для проведения эксперимента выбраны архитектуры сверточной нейронной сети и сети с долгой краткосрочной памятью. Было определено, что для сравнения моделей необходим единые входной и выходной форматы данных. Обе модели нуждаются в едином виде преобразования входных данных, которым наиболее целесообразно выбрать векторные представления слов. Так как архитектуры универсальны, то в рассматриваемой задачи определения тональности текста необходимо использовать полносвязные слои нейронов для получения выхода нейронной сети в одинаковом формате.

В **третьей главе** были реализованы и обучены модели сверточной нейронной сети и сети с долгой краткосрочной памятью. Было рассмотрено несколько наиболее универсальных и популярных библиотек для программной реализации моделей НС и их обучения. Для реализации моделей используется библиотека Keras, так как она обладает развитым функционалом, проста в использовании, позволяет использовать графический процессор для ускорения вычислений. Кроме этого модель нейронной сети может быть сохранена и использована при развертывании.

Для проведения эксперимента были одинаковые параметры для слоя векторных представлений. Кроме этого обучение обеих моделей производилось на одном и том же наборе данных одинаковое количество эпох. Параметры, такие как алгоритм оптимизации, параметры словаря, задавались одинаковыми.

Проведен эксперимент по обучению обеих моделей. Выявлено, что большей точностью на одинаковых наборах данных и при сходных условиях обладает сверточная нейронная сеть. СНС послужила основой для разработанного приложения.

Реализовано веб-приложение на основе микросервисной архитектуры, использующее современные технологии управления контейнерами, в виде которых разработаны модули. Система состоит из трех модулей: пользовательский интерфейс, модуль сбора данных, модуль анализа текста.

Модуль пользовательского интерфейса разработан с помощью библиотеки Angular и предназначен для отображения результатов сбора и анализа данных. Он взаимодействует с модулем сбора данных. Все модули являются контейнерами, предназначенными для работы на кластере Kubernetes.

Модуль сбора данных был разработан на языке Java при помощи технологий Spring Framework и Spring Boot. Он обеспечивает взаимодействие с Twitter API и с модулем анализа текста.

Модуль анализа данных разработан на языке Python и библиотеке Flask. Этот модуль использует ранее обученную модель сверточной нейронной сети для анализа текста.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Рассмотрена возможность использования нейронных сетей для анализа отзывов пользователей. Такие особенности нейронных сетей, как обобщение и абстрагирование позволяют их применять в задачах обработки естественного языка.

2. Изучена возможность применения архитектуры сверточных нейронных сетей и сетей с долгой краткосрочной памятью для решения задачи анализа тональности текста. Обе модели используются в задачах обработки естественного языка. Имеют оптимизированную структуру по сравнению с персептроном, что положительно влияет как на скорость обучения.

3. Разработаны модели нейронных сетей. В случае сверточной нейронной сети использовалась многоканальная архитектура для идентификации признаков среди 3, 4 и 5 соседних слов. LSTM-сеть способна работать сразу со всем входным текстом, обрабатывая его пословно. Показана их эффективность в решении поставленной задачи.

4. Проведено сравнение результатов обучения и применения обеих моделей. Выявлено, что несмотря на то, что LSTM-сеть обучается быстрее (более высокая точность достигается раньше), лучшую максимальную точность показывает модель сверточной нейронной сети с несколькими каналами[2-А.].

5. Реализовано web-приложение на основе модульной системы с применением микросервисной архитектуры в купе с современными технологиями разработки ПО. Приложение использует разработанную и обученною модель СНС и способно применять ее к реальным данным, получаемым самой системой.

Проведена апробация результатов на двух научных конференциях. Результаты опубликованы в двух печатных работах. Реализованный программный продукт внедрен в учебный процесс в качестве материалов для лекционного курса «Нейросетевое моделирование и обработка данных» магистрантов специальности 1-40 81 04 «Обработка больших объемов информации», что подтверждается актом внедрения (Приложение Ж).

Рекомендации по практическому использованию результатов

1. Реализовано применение модели нейронной сети для реальных данных, полученных автоматическим сбором в социальной сети Twitter по заданной пользователем теме. Для реализации применялись современные технологии разработки ПО. Приложение состоит из нескольких модулей, работа которых происходит в контейнерах.

2. Можно улучшить классификатор путем изменения обучающего множества, что может дать более релевантные результаты. Также можно изменить количество классов в классификаторе добавив распознавание нейтральной окраски, а также различных степеней позитивной и негативной.

3. Одним из путей развития разработанной системы может быть развертывание ее в облачных сервисах для общей доступности.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Витковский, А.В. Применение рекурсивных нейронных сетей для анализа тональности текста / А. В. Витковский // Компьютерные системы и сети: материалы 54-й научной конференции аспирантов, магистрантов и студентов – Минск, 2018. – С. 152 – 153.

2-А. Витковский, А.В. Сравнение сверточных и рекуррентных нейронных сетей в задаче анализа тональности текста / А. В. Витковский // 55-я юбилейная конференция аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» – Минск, 2019. – С. 287.