МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра дискретной математики и информационных технологий

АНАЛИЗ МНЕНИЙ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Сапаевой Иоанны Викторовны

Научный руководитель
Профессор кафедры ДМиИТ _____ Л. В. Кальянов

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2023

ВВЕДЕНИЕ

Основной целью данной работы является разработка и применение моделей машинного обучения для классификации отзывов на основе их тональности, то есть определение, является ли отзыв положительным, или отрицательным. Для достижения этой цели необходимо решить следующие задачи:

Сбор и предобработка данных: необходимо собрать достаточно большой датасет отзывов с сайта Амазон и провести их предварительную обработку, включающую очистку данных, токенизацию и удаление стоп-слов.

Выбор и обучение моделей: необходимо выбрать подходящие модели машинного обучения для анализа мнений и обучить их на предварительно обработанных данных.

Оценка производительности моделей: необходимо провести эксперименты для оценки производительности различных моделей.

Анализ результатов: необходимо проанализировать полученные результаты и сделать выводы о применимости и эффективности выбранных моделей машинного обучения для анализа мнений на основе датасета отзывов с сайта Амазон.

Ожидается, что результаты данной работы помогут в создании инструмента для автоматического анализа мнений на основе отзывов с сайта Амазон. Этот инструмент может быть полезен для бизнеса в плане понимания общественного мнения о продуктах и услугах, выявления проблем и улучшения качества предлагаемых товаров. Кроме того, результаты исследования могут быть использованы для дальнейших исследований в области анализа мнений и машинного обучения.

Актуальность данной работы заключается в том, что результаты работы могут быть использованы для принятия решений в области управления качеством продуктов и услуг, а также для улучшения взаимодействия с клиентами. К примеру, возможна реализация оценки негативного отзыва и выяснения причины недовольства клиента с дальнейшей рекомендацией ему по исправлению причинённых неприятностей. Допустим, рекомендация купить схожий по цене и предназначению продукт лучшего качества, связь с техническим специалистом или советы по исправлению недочётов уже приобретённого продукта.

Бакалаврская работа состоит из введения, двух разделов, заключения, списка использованных источников и двух приложений, где продемонстрирован

использованный код. Общий объем работы, не включая титульник, 63 страницы. Из них 39 - основное содержание, включая 44 рисунка. Список использованных источников - 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе данной выпускной квалификационной работы будут рассмотрены теоретические сведения, которые понадобятся в дальнейшем в практической части.

Так как основной поставленной задачей является анализ мнений методами машинного обучения, то дадим ему определение.

Анализ мнений или анализ тональности текстов - это спектр задач из области компьютерной лингвистики, решающий проблемы автоматизированного распознавания и изучения эмоциональной оценки авторов с использованием эмоционально окрашенной лексики по отношению к объектам, о которых идёт речь в конкретном тексте.

В анализе мнений методами машинного обучения используются алгоритмы классификации, позволяющие делать предсказания к какому классу, то есть настроению, относится то или иное высказывание.

Задача классификации - это подкатегория методов машинного обучения с учителем, суть которой заключается в идентификации категориальных меток классов для новых экземпляров на основе предыдущих наблюдений. Метка класса представляет собой дискретное, неупорядоченное значение, которое может пониматься как принадлежность группе экземпляров.

Бинарная классификация - это один из типов задач классификации в машинном обучении, когда необходимо классифицировать два взаимоисключающих класса. То есть, это классификация с бинарной переменной класса, или категориальной выходной переменной, которая может принимать только два значения. Бинарные классификационные модели являются более понятными и интерпретируемыми.

Для их построения широко распространены такие методы, как логистическая регрессия и свёрточные нейронные сети.

Логистическая регрессия - это полезный инструмент для решения задач регрессии и классификации, представляет собой разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными, называемыми также предикторами или регрессорами, и зависимой переменной. Бинарная логистическая регрессия применяется в том случае, когда когда зависимая переменная является бинарной, то есть принимающей только два значения, к примеру, 0 или 1. С помощью логи-

стической регрессии можно оценивать вероятность того, что событие наступит для конкретного испытуемого объекта.

Рассмотрим также свёрточные нейронные сети, которые будут использоваться в дальнейшем в практической части работы.

Глубокое обучение может пониматься, как набор алгоритмов, разработанных для наиболее эффективной тренировки искусственных нейронных сетей, состоящих из множества слоёв.

Искусственные нейроны составляют базовые элементы многослойных искусственных нейронных сетей. Ключевая идея, лежащая в основе искусственных нейроных сетей, опирается на гипотезы и модели того, каким образом человеческий мозг работает для решения сложных практических задач.

Одним из видов нейронных сетей являются свёрточные нейронные сети (convolutional neural network, CNN или ConvNet), которые завоевали популярность в компьютерном зрении из-за их экстраординарно хорошего качества работы на задачах классификации изображений, аудиофайлов, в частности речи и анализа тональности текстов. На сегодняшний день CNN являются одной из самых популярных архитектур нейронных сетей в глубоком обучении. Лежащая в основе свёрточных нейронных сетей ключевая идея состоит в создании множества слоёв детекторов признаков (feature detectors), чтобы, к примеру, учитывать пространственное расположение пикселов во входном изображении.

Свёрточная нейронная сеть состоит из двух основных слоёв:

1. Слой свёртки для получения признаков из данных.
2. Слой объединения для уменьшения размера карты признаков.

Слой свёртки и объединяющий слой являются основными составляющими свёрточной нейронной сети.

Также в практической части будут использованы батч-нормализация и эмбеддинги.

Батч-нормализация - это метод, который позволяет повысить производительность и стабилизировать работу искусственных нейронных сетей. Суть данного метода заключается в том, что некоторым слоям нейронной сети на вход подаются данные, предварительно обработанные и имеющие нулевое математическое ожидание и единичную дисперсию.

Эмбеддинги - это возможность уменьшения размерности таких признаков ради повышения производительности модели. В самой примитивной форме

эмбеддинги слов получают простой нумерацией слов в некотором достаточно обширном словаре и установкой значения единицы в длинном векторе размерности, равной числу слов в словаре.

Обучение эмбеддингов может быть выполнено с использованием различных методов, таких как Word2Vec, GloVe, FastText и других. Эти методы обучаются на больших текстовых корпусах и пытаются уловить семантические и синтаксические свойства слов на основе их соседей в тексте.

Применение эмбеддингов в NLP позволяет моделям машинного обучения улучшить качество работы на задачах, таких как классификация текстов, машинный перевод, определение тональности текста и других. Эмбеддинги позволяют моделям лучше понимать семантическую близость и отношения между словами, что помогает уловить более высокий уровень информации в тексте.

Во второй главе рассмотрена практическая реализация поставленной задачи.

Практическая работа была выполнена на языке Python, поскольку он поддерживает большое количество библиотек, удобных для использования при реализации машинного обучения, в частности, библиотеки scikit-learn, TensorFlow, Keras, NLTK. Для реализации практической части данной выпускной квалификационной работы был выбран датасет отзывов к товару Alexa, продающемуся на интернет-площадке Amazon. Сам датасет был выбран на сайте kaggle.com. В датасете представлены около 3000 отзывов клиентов Amazon, звёздный рейтинг и дата оставленного отзыва. Был взят англоязычный датасет, поскольку использованная в работе библиотека NLTK плохо поддерживает русский язык.

Также вся работа выполнялась в бесплатной интерактивной облачной среде для работы с кодом Google Colab.

Были рассмотрены четыре модели обучения с различной точностью. Лучшая точность была получена на модели глубокого обучения с использованием библиотеки Keras, где свёрточные слои были заменены на рекуррентные.

Выбранный датасет был обработан на наличие пустых значений и дубликатов. К нему была применена лемматизация и токенизация, составлен мешок слов. Было проведено разделение данных на обучающую и тестовую выборки с использованием функции train_test_split.

Были программно реализованы четыре модели: логистическая регрессия, модель глубокого обучения на основе библиотеки Keras, свёрточные нейронные

сети и рекуррентная нейронная сеть.

ЗАКЛЮЧЕНИЕ

В данной выпускной квалификационной работе был проведён анализ мнений с использованием методов машинного обучения на основе датасета отзывов с сайта Амазон. В процессе исследования были реализованы несколько моделей машинного обучения с разной точностью.

Вначале был проведён предварительный анализ датасета, включающий обработку текстовых данных, удаление лишних символов, токенизацию и векторизацию текстов. Затем было проведено разделение данных на обучающую и тестовую выборки с использованием функции train_test_split, учитывая стратификацию для сохранения баланса классов.

Для обучения моделей были применены различные алгоритмы машинного обучения, включая логистическую регрессию, модель глубокого обучения с использованием библиотеки Keras и нейронные сети. Каждая модель была обучена на обучающей выборке и оценена на тестовой выборке с использованием метрик точности.

Результаты экспериментов показали, что различные модели машинного обучения демонстрируют разную точность в анализе мнений. Логистическая регрессия показала хорошую точность, достигая практически 93%, в то время как модель на основе Keras достигла 92% точности. Свёрточные нейронные сети также показали неплохие результаты, достигая 93% точности.

На второй модели с использованием Keras, где свёрточные слои были заменены на рекуррентные слои, была достигнута наивысшая в данной работе точность в 94%.

Таким образом, на основе проведённого исследования можно сделать вывод, что анализ мнений методами машинного обучения на основе датасета отзывов с сайта Амазон является эффективным инструментом для определения тональности отзывов. Однако, точность классификации может различаться в зависимости от выбранной модели машинного обучения. Результаты работы могут быть использованы для принятия решений в области управления качеством продуктов и услуг, а также для улучшения взаимодействия с клиентами. К примеру, возможна реализация оценки негативного отзыва и выяснения причины недовольства клиента с дальнейшей рекомендацией ему по исправлению причинённых неприятностей. Допустим, рекомендация купить схожий по цене и предназначению продукт лучшего качества, связь с техническим специалистом

или советы по исправлению недочётов уже приобретённого продукта.

Дальнейшие исследования могут быть направлены на улучшение точности моделей, включая оптимизацию параметров, использование более сложных архитектур нейронных сетей и обработку текста с использованием методов обработки естественного языка. Также можно провести анализ на других датасетах и сравнить результаты с данной работой.

Основные источники информации:

1 Рашка С. Python и машинное обучение / пер. с англ. А. В. Логунова. - М.: ДМК Пресс, 2017. - 418 с.: ил.

2 Логистическая регрессия и ROC-анализ - математический аппарат [Электронный ресурс] URL: https://loginom.ru/blog/logistic-regression-roc-auc (дата обращения: 04.05.2023) - Яз. рус.

3 Derrick Mwiti. Convolutional Neural Network for Sentence Classification [Электронный ресурс] URL: https://cnvrg.io/cnn-sentence-classification/ (дата обращения: 21.05.2023) - Яз. англ.

4 S. Ioffe, C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Электронный ресурс] URL: https://arxiv.org/abs/1502.03167 (дата обращения: 22.05.2023) - Яз. англ.

5 Amazon Alexa Reviews [Электронный ресурс] URL: https://www.kaggle.com/datasets/sid321axn/amazon-alexa-reviews?select=amazon_alexa.tsv (дата обращения: 10.05.2023) - Яз. англ.

6 Алгоритмы бинарной классификации в машинном обучении. [Электронный ресурс] URL: https://biconsult.ru/products/algoritmy-binarnoy-klassifikacii-v-mashinnom-obuchenii (дата обращения: 06.05.2023) - Яз. рус.

7 Самигулин Т.Р., Джурабаев А.Э.У. Анализ тональности текста методами машинного обучения [Электронный ресурс] URL: https://cyberleninka.ru/article/n/a tonalnosti-teksta-metodami-mashinnogo-obucheniya/viewer (дата обращения: 04.05.2023) - Яз. рус.

8 Воронцов К. В. Метрические методы классификации и регрессии [Электронный ресурс] URL: https://www.youtube.com/watch?v=GyOxB2itxnc&ab_chann (дата обращения: 05.03.2023) - Яз. рус.

9 ML: Embedding слов [Электронный ресурс] URL: https://qudata.com/ml/ru/NN_Em (дата обращения: 23.05.2023) - Яз. рус.

10 Keras: описание и особенности [Электронный ресурс] URL: https://learn.microsoft.
ru/azure/machine-learning/component-reference/convert-word-to-vector (дата
обращения: 12.05.2023) - Яз. рус.
11 Batch Normalization. Основы. [Электронный ресурс] URL: http://vbystricky.ru/2020
(дата обращения: 22.05.2023) - Яз. рус.