Научная статья
УДК 51-76
doi: 10.17586/2713-1874-2021-2-35-48

РАЗРАБОТКА МОДЕЛИ И АЛГОРИТМА АГРЕГАЦИИ И КЛАССИФИКАЦИИ ДАННЫХ ДЛЯ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ ПЕРСОНАЛИЗИРОВАННОГО ПИТАНИЯ

Александр Александрович Рейбандт[1], Алексей Николаевич Арсениев[2]✉, Татьяна Геннадьевна Максимова[3]

[1,2,3]Университет ИТМО, Санкт-Петербург, Россия
[1]reybandt99@mail.ru
[2]arsenievaleksey@gmail.com✉
[3]tgmaximova@itmo.ru
Язык статьи – русский

Аннотация: В статье продемонстрирован проект и реализация алгоритма агрегации данных для будущей рекомендательной системы, ориентированной на поддержку решения по выбору персонализированной схемы питания. Отличительной особенностью классификатора, реализуемого в рамках системы, является тот факт, что на вход он принимает одновременно изображения и текстовые данные для выработки более точных и сбалансированных вариантов решений. Проведен обзор инструментов и подходов на различных этапах агрегации изображений и текстовых данных. Определены метрики оценивания предсказаний реализованной модели для классификации географических меток, а также анализа среднего сентимента отзывов пользователей. Предсказанные географические теги и выявленные тональности комментариев были добавлены в основной датафрейм в качестве дополнительных признаков.

Ключевые слова: разработка программного обеспечения для задач принятия решения, парсинг, обработка информации, тональность текста, персонализированное питание, нейронные сети

Ссылка для цитирования: Рейбандт А.А., Арсениев А.Н., Максимова Т.Г. Разработка модели и алгоритма агрегации и классификации данных для рекомендательной системы персонализированного питания // Экономика. Право. Инновации. 2021. № 2. С. 35–48. http://dx.doi.org/10.17586/2713-1874-2021-2-35-48.

DEVELOPMENT OF A MODEL AND ALGORITHM FOR DATA AGGREGATION AND CLASSIFICATION FOR A PERSONALIZED NUTRITION RECOMMENDATION SYSTEM

Aleksandr A. Reybandt[1], Aleksey N. Areseniev[2]✉, Tatyana G. Maximova[3]

[1,2,3]ITMO University, Saint Petersburg, Russia
[1]reybandt99@mail.ru
[2]arsenievaleksey@gmail.com✉
[3]tgmaximova@itmo.ru
Article in Russian

Abstract: The article demonstrates the design and implementation of a data aggregation algorithm for a future recommendation system in the field of personalized nutrition. It was based on theoretical materials on machine learning methods in natural language processing, as well as tutorials on building classification models using the Keras library. A distinctive feature of the classifier implemented within the framework of this project is the fact that it simultaneously accepts images and text data as input to obtain more accurate and balanced predictions. The implementation of the designed data aggregation algorithm for the recommendation system in the field of personalized nutrition is considered in detail. A review was made of the tools and approaches chosen at various stages of aggregation. The metrics for evaluating the predictions of the implemented model for the classification of geographic labels, as well as the analysis of the average sentiment of user reviews are determined and the results are visualized. Predicted geo tags and revealed comment sentiments were added to the main data frame as additional features.

Keywords: software development for decision-making tasks, parsing, information processing, text tonality, personalized nutrition, neural networks

For citation: Reybandt A.A., Arseniev A.N., Maximova T.G. Development of a Model and Algorithm for Data Aggregation and Classification for a Personalized Nutrition Recommendation System. *Ekonomika. Pravo. Innovacii.* 2021. No. 2. pp. 35–48. (in Russ.). http://dx.doi.org/10.17586/2713-1874-2021-2-35-48.

Введение. Пищевая промышленность всегда ставила перед собой задачу, заключающуюся в том, чтобы продавать продукты, адаптированные к вкусам и предпочтениям конкретных групп потребителей (например, энергетические батончики для спортсменов). Персонализированное питание идет еще дальше в этом отношении и пытается учесть интересы конкретных лиц. Для этой цели используются индивидуальные показатели состояния человека, считываемые смартчасами, фитнес браслетами (которые сегодня можно встретить практически у каждого) или определяемые в процессе личного медицинского осмотра у доктора с использованием анализа крови, микробиома или ДНК.

Сегодня на передний план выходят рекомендательные системы, способные подбирать пользователям персонализированную диету, делать индивидуальные рекомендации по рациону питания на основе уровня активности, а также таких показателей, как рост, вес, возраст, баланса белков, жиров и углеводов в организме, результатов анализа крови и даже ДНК-тестов.

Цель данной работы – спроектировать и реализовать алгоритм агрегации данных для рекомендательной системы персонализированного рациона питания.

Данная работа является частью проектной разработки сервиса «Foodline», целевой аудиторией которого являются диетологи. Основная его идея – поддержка принятия решений диетологов при составлении рационов питания с учетом индивидуальных особенностей пациента. Реализуемая информационная система справляется с этими задачами благодаря автоматическому расчету КБЖУ из рациона, гибкой настройке анкеты, сбора анкетных данных, первичного анамнеза, учета питания и показателей, а также обширной базе рецептов.

Агрегации данных: состояние вопроса. Для детального рассмотрения методов агрегации данных для рекомендательных систем сначала необходимо дать определение непосредственно самому понятию «агрегации данных». Процедура агрегирования данных заключается в том, чтобы с помощью известных методов разбить первоначальный набор данных на поднаборы, которые меньше исходного по объему, при этом сохранив и выявив новые знания и закономерности в данных. В таком контексте агрегация напоминает собой процесс «сжатия» информации.

Но есть и другая точка зрения, когда агрегация рассматривается как процедура приведения детализированного набора данных к наиболее общему виду. Это реализуется с помощью нахождения специальных значений, именуемых агрегатами (отсюда и берется название рассматриваемой процедуры). Эти значения являются результатом применения того или иного преобразования к существующему набору прецедентов. Яркими примерами агрегированных процедур могут считаться нахождение максимального, минимального, медианного значений, избавление от выбросов в данных и т.д.

Среди сильных сторон агрегации данных является тот факт, что агрегированная информация зачастую оказывается очень устойчивой к изменениям по причине того, что случайные факторы оказывают на нее меньшее влияние. Помимо этого, после агрегации набора данных явнее прослеживаются общие тенденции и закономерности процесса исследования. Однако чрезмерное обобщение данных и целенаправленный глубокий уход от деталей может являться причиной, по которой часть важной информации об исследуемом процессе может быть потеряна.

В работе процедура агрегации данных также включает в себя процесс их объединения и создания на разрозненных элементах единой системы знаний.

Известно, что методы агрегирования данных для рекомендательных систем достаточно разнообразны. Но им предшествует непосредственно сбор данных. Этому этапу уделяется большое внимание, так как он является базовым и непосредственно влияет на

всю закладываемую цепочку процессов в любом проекте по машинному обучению. Также остается неоспоримым тот факт, что качество предсказаний реализуемой модели коренным образом зависит от качества собранных данных, на которых тот или иной алгоритм машинного обучения обучается.

Далее рассматриваются техники, с помощью которых формируются наборы данных для обучения прогнозирующих моделей.

В качестве первого варианта выступают предварительно очищенные наборы данных, находящиеся в свободном доступе. Это идеальный вариант, при котором постановка задачи (например, распознавание объектов) имеет под собой базу из уже имеющегося опыта работы.

Второй техникой является парсинг и извлечение данных из веб-ресурсов средствами специальных инструментов и так называемых «пауков», которые предоставляют возможность получения содержимого интересующего исследователя интернет-ресурса.

Далее идут личные данные. Инженеры по машинному обучению могут создавать свои собственные данные. Это полезно в тех случаях, когда объем данных, необходимых для обучения модели, невелик, а изложение проблемы слишком специфично для обобщения по набору данных с открытым исходным кодом.

Основой для обучения моделей машинного обучения являются пользовательские данные. Агентства могут создавать или получать данные посредством краудсорсинга, при этом взимая определенную плату за свои услуги [1].

Вышеописанные способы сбора данных являются неотъемлемым подспорьем для рассматриваемых ниже методов агрегации информационных потоков. Далее рассматриваются способы агрегации данных в области информационных систем.

Первой из рассматриваемых является распределенная агрегация. Это вид агрегирования информации, который основывается на использовании пространственного признака с помощью гистограмм. Преимуществом данного подхода является тот факт, что он уже на начальной стадии подготовки данных к моделированию делает возможным проведение предварительного анализа, целью которого является извлечение из данных полезной информации, избавление от так называемых выбросов, которые в последствии могут негативно повлиять на предсказания, и непосредственное сжатие данных, которое в свою очередь будет способствовать уменьшению временных затрат на обучение модели [2].

Помимо распределенной агрегации интерес представляет временная агрегация, которая связана с учетом изменения частоты наблюдения исследуемой переменной. Служит альтернативой использованию средних значений или интервального анализа, при которых происходят потери информации [2].

Идея способа агрегирования и преобразования данных заключается в формировании древовидной структуры объекта исследования и создании трех типов контейнеров данных, в которые записывается каждое свойство этого объекта. Каждый тип контейнера отвечает за текущее состояние данных: в первом происходит агрегация сырых данных по отдельным свойствам (устройство обработки), во втором – осуществляется консолидированная обработка агрегированных данных, а третий тип служит в качестве устройства хранения консолидированной информации [3].

Также нельзя не упомянуть агрегацию информации с использованием механизма формальных проекций. Указанный способ подразумевает введение дополнительного механизма, целью которого является преобразование данных из произвольного вида в удобный для централизованной обработки, поддержки произвольных запросов и индексирования вид [4].

Перед рекомендательными системами персонализированного питания выдвигаются серьезные требования. Так, например, эксперты считают, что модели по составлению персонализированного питания должны использовать многоуровневую организацию знаний и собирать информацию о биологических концепциях на уровне интервенции (пища, нутриенты, соединения, образ жизни), промежуточном молекулярном / биологическом уровне (возможные маркеры болезни, отслеживание состояния внутренних органов), а также на уровне фенотипа [5].

Материалы и методы исследования. Одной из будущих функций подбора индивидуального рациона питания для пользова-

телей будет являться учет географии кухни. При парсинге информации и составлении датафрейма возникла проблема, связанная с отсутствием разметки части рецептов по географии приготовления. Для решения этой задачи было принято решение построить и обучить модель на размеченных данных, которая на вход принимает изображение рецепта и его название, а на выходе производит его классификацию относительно региональных особенностей кухни.

Алгоритм действий при классификации изображений формируется из следующих этапов:

1) изучить и понять данные;

2) настроить процесс разработки по типу конвейера (от англ. «pipeline») для входных данных;

3) построить модель;

4) обучить модель;

5) протестировать модель;

6) улучшить модель и повторить процесс [6].

Однако в этом случае задача классификации осложнялась тем фактом, что на вход модель машинного обучения также должна была принимать текстовые данные (название рецепта), так как они очень сильно коррелируют с географией кухни (например, рецепт «Американские блины» можно однозначно отнести к американской кухне). Возникшая проблема при такой структуре модели заключалась в том, что было необходимо произвести обработку данных таким образом, чтобы векторы признаков как изображений, так и текстового описания учитывались при получении предсказаний.

В качестве модели для классификации изображений была выбрана сверхточная нейронная сеть, в то время как для предобработки названий рецептов (перед их непосредственной передачей в качестве входных данных в модель машинного обучения) были использованы средства обработки естественного языка, речь о которых велась в пункте выше. В частности, были проделаны такие шаги, как приведение текста к нижнему регистру, удаление стоп-слов, токенизация и векторизация текста.

Несмотря на то, что полученный метод веб-скрапинга датафрейм рецептов оказался не таким уж и маленьким (размеченных данных для обучения получилось около четырех тысяч). Однако он едва ли может сравниться с массивной базой данных аннотированных изображений под названием ImageNet, состоящей из более, чем 1,2 миллиона картинок, размеченных по 1000 классам. Поэтому было принято решение задействовать подход, который носит название «Transfer Learning», суть которого заключается в использовании предобученных моделей для новых задач.

Указанная техника завоевала большую популярность по причине того, что при обучении нейросети с нуля на не очень большом наборе данных часто возникает проблема переобучения (когда модель отлично работает на данных для обучения, но имеет низкую обобщающую способность) [7]. На таком фоне с целью получения более качественной модели, способной давать надежные предсказания как на обучающей, так и на тестовой выборке, исследователи в области компьютерного зрения часто дообучают сильную уже предобученную нейронную сеть на непосредственно своих наборах данных.

Для разложения текстового описания рецептов на векторы использовался TF-IDF векторизатор. В качестве значений он выдает относительную частоту слова в документе, умноженную на обратную частоту документов, в котором есть это слово. Данный подход позволяет присвоить высокую оценку тем словам, которые часто встречаются в одном документе, но в целом корпусе представлены в небольшом количестве документов.

Результаты: проектирование алгоритма агрегации данных. Процесс реализации алгоритма агрегации данных условно разделен на три этапа:

– веб-скрапинг информации о рецептах;

– создание модели для классификации рецептов относительно географии приготовления;

– анализ общего сентимента комментариев пользователей по каждому рецепту.

Диаграмма, отображающая процесс парсинга информации о рецептах, начиная инициализацией проекта и заканчивая формированием датафрейма, представлена на Рисунке 1.

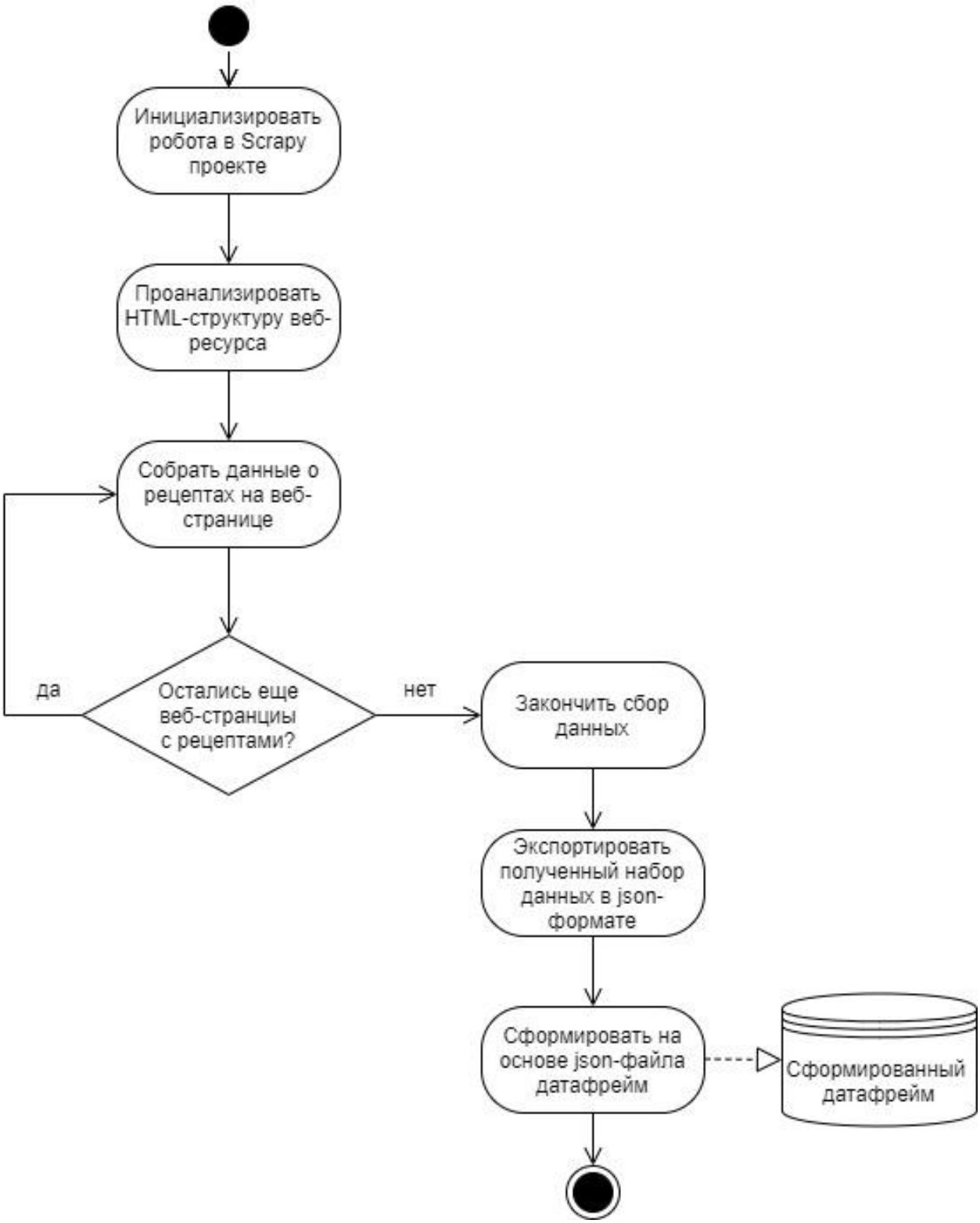

Рисунок 1 – Диаграмма активности парсинга данных

На Рисунке 2 изображена диаграмма активности классификатора, который предсказывает, какую географию имеет тот или иной рецепт. Финальной стадией данного этапа агрегации данных является получение предсказаний для неразмеченных относительно географии кухни данных.

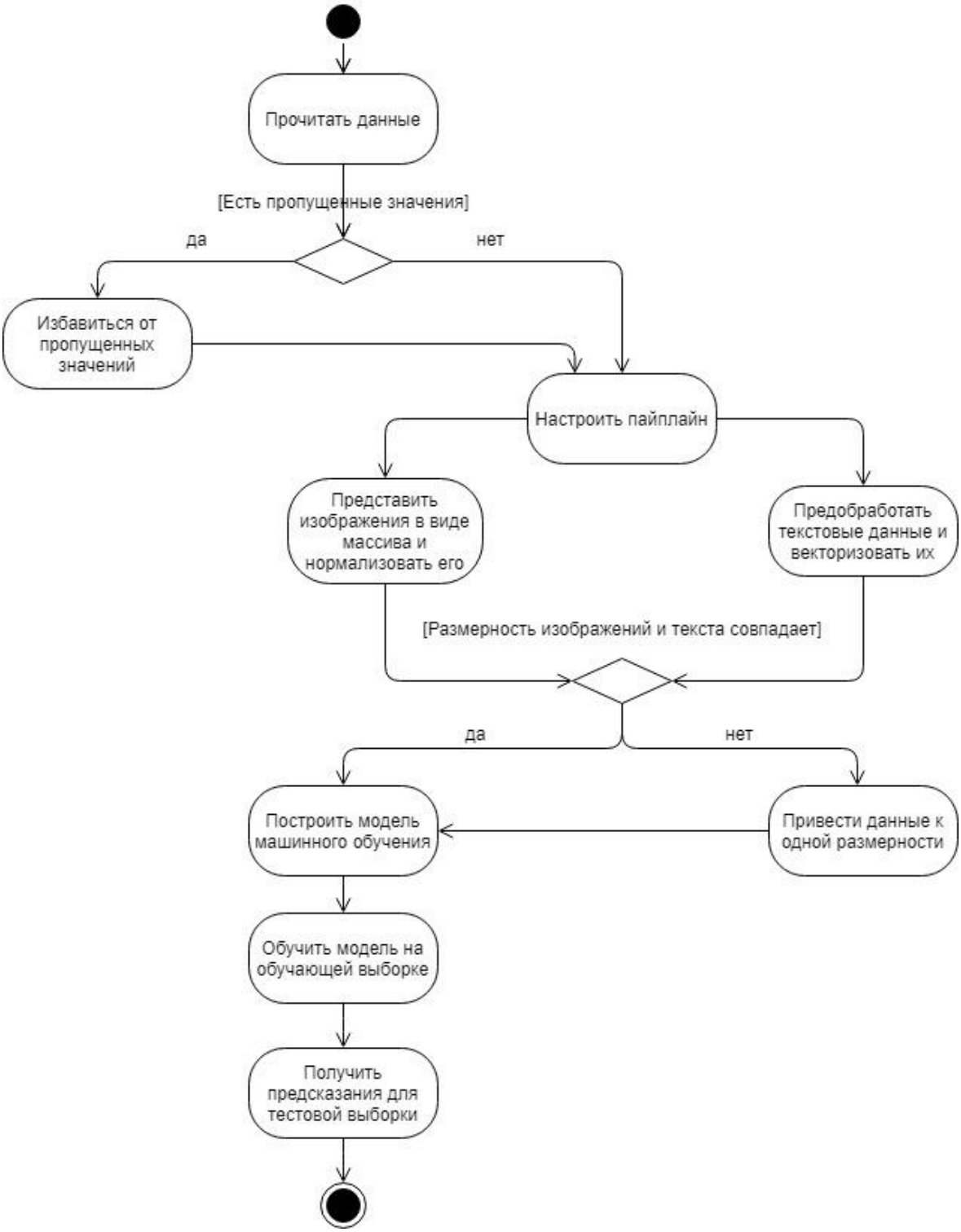

Рисунок 2 – Диаграмма активности классификатора

Заключительным этапом обработки набора данных для будущей рекомендательной системы персонализированного рациона питания является выявление общего сентимента для каждого рецепта, опираясь на комментарии пользователей. Эта стадия проекта также связана с обработкой текстовых данных, построением модели и получением предсказаний относительно того, являются ли пользователи в среднем довольны, нейтральны или разочарованы рецептом. Диаграмма активности процесса анализа общего сентимента комментариев пользователей представлена на Рисунке 3.

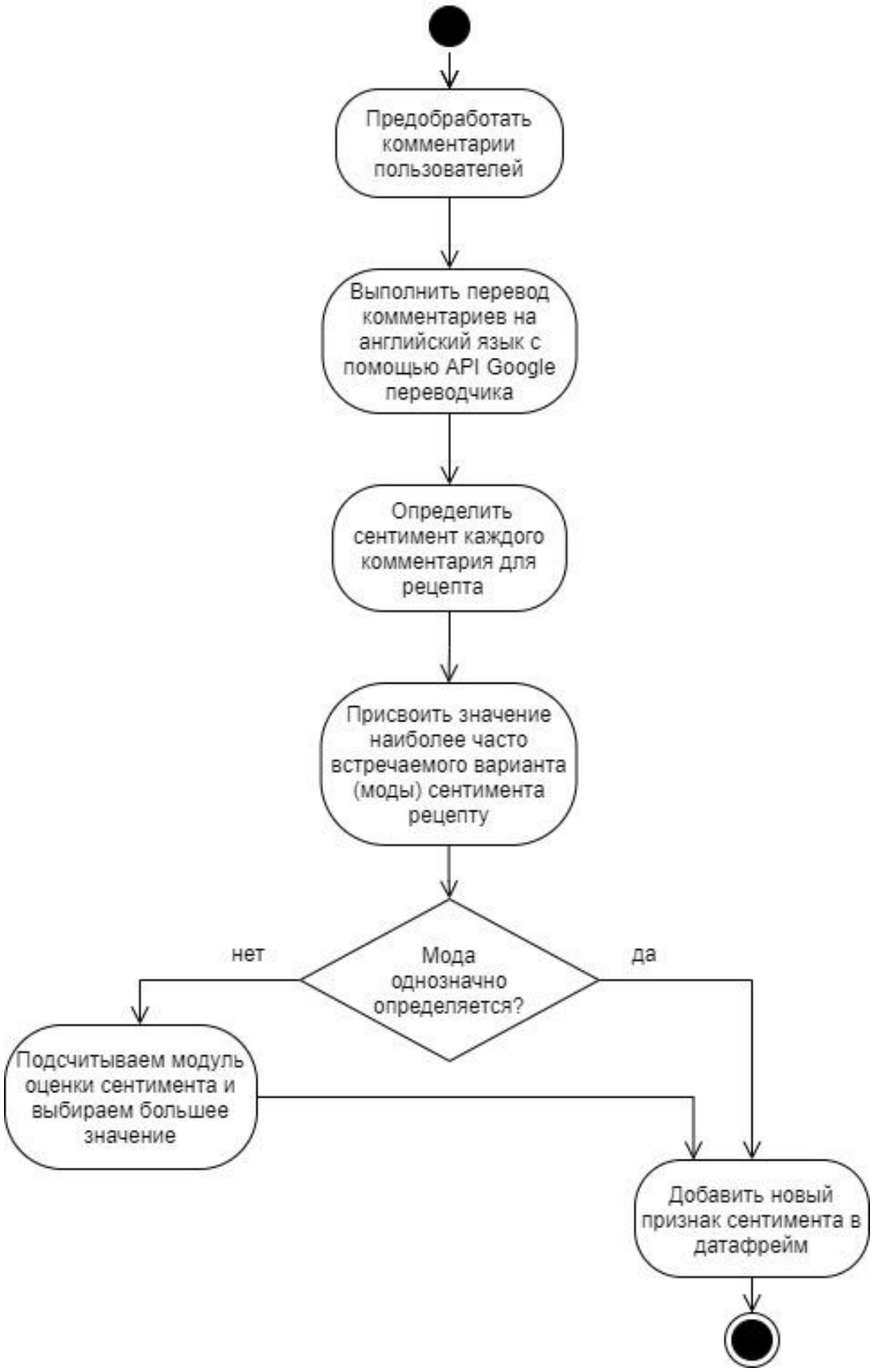

Рисунок 3 – Диаграмма активности процесса анализа общего сентимента

Создание веб-парсера. В качестве инструментов для получения содержимого ресурсов с рецептами (по итогам консультации с диетологами и формирования «рейтинга доверия» к тому или иному источнику были выбраны два веб-сайта с рецептами, а именно «Еда» [eda.ru] и «fatsecret РОССИЯ» [fatsecret.ru]) использовались Scrapy и Selenium.

Scrapy – быстрый веб-парсер высокого уровня, используемый для извлечения структурированных данных с веб-страниц. Он может использоваться для самых различных целей, начиная с анализа данных и заканчивая мониторингом и автоматизированным тестированием [8].

Selenium – комплексный проект, предоставляющий ряд инструментов и библиотек, которые поддерживают автоматизацию веб-браузеров [9].

В рамках реализуемого проекта для каждого рецепта была собрана следующая информация:

1) название рецепта;
2) ингредиенты;
3) количество порций;
4) время приготовления;
5) нутриенты;
6) рейтинг, оставленный пользователями на платформе;
7) ссылка на изображение рецепта;
8) комментарии пользователей.

Использование не только библиотеки Scrapy для получения информации о рецептах может быть обосновано тем фактом, что на ресурсе «Еда» для получения полного списка комментариев пользователей необходимо нажать на кнопку «Показать все комментарии». Это означает, что необходимо каким-то образом взаимодействовать напрямую с браузером. Для этой цели был задействован Selenium и с его помощью запущен веб-драйвер (в качестве браузера использовался Google Chrome). С помощью селекторов в HTML-структуре сайта был найден фрагмент кода, отвечающий за отрисовку вышеописанной кнопки и в дальнейшем написан скрипт для ее нажатия в автоматическом режиме.

Результатом работы парсера является JSON-файл. Для проведения дальнейших манипуляций с данными было принято решение представить его в виде датафрейма в среде JupyterNotebook.

При получении содержимого с двух источников с рецептами возникла проблема, связанная с тем фактом, что на ресурсе «fatsecret РОССИЯ» рейтинг рецепта представлен в виде картинки со звездами (от одной до пяти), в то время как на «Еда» приводится информация только о количестве лайков и дизлайков.

Решением данной проблемы стало представление рейтинга в унифицированном виде для обоих источников. Для этого, во-первых, был проведен семантический анализ ссылки на изображение с рейтингом на «fatsecret Россия» и было получено числовое значение на основе этой ссылки. Например, блюда, имеющие четыре звезды в качестве рейтинга, обладают ссылкой https://a.ftscrt.com/static /images/stars/big-four-star.gif, и посредством ее разбиения по символам и анализу последнего элемента этого разбиения для блюда определяется число звезд. Во-вторых, для рецептов с ресурса «Еда» был произведен подсчет доли лайков относительно общего количества лайков/дизлайков и присвоен рейтинг в зависимости от полученного значения.

Зависимость определяемого рейтинга от доли лайков представлена в Таблице 1.

Таблица 1

Сопоставление доли лайков рейтингу рецепта

Доля лайков (X)	$X < 0,3$	$0,3 \leq X < 0,5$	$0,5 \leq X < 0,7$	$0,7 \leq X < 0,9$	$X \geq 0,9$
Рейтинг	1	2	3	4	5

Первые пять строк получившегося набора данных представлены на Рисунке 4:

	recipe_name	measured_ingredients	number_of_servings	cooking_time	nutrients	tags	likes_and_dislikes	img_url	reviews
0	Американские блины	[2 штуки Куриное яйцо, 1 чайная ложка Соль, 3 ...	10	20 минут	[242, 6,5, 7,8, 36,7]	[Главная, Завтраки, Американская кухня, Панкей...	[3118, 297]	https://eda.ru/img/eda/c88x88i/s2.eda.ru/Stati...	[\n Быстро, вкусно и выглядит как на картин...
1	Блины тонкие на кипятке и молоке	[1 стакан Пшеничная мука, 2 штуки Куриное яйцо...	4	15 минут	[331, 10, 14,7, 40,1]	[Главная, Выпечка и десерты, Русская кухня, То...	[1238, 145]	https://eda.ru/img/eda/c88x88i/s1.eda.ru/Stati...	[\n отличный рецепт. Спасибо\n , \n рг...
2	Пирог «Зебра»	[2 стакана Сахар, 5 штук Куриное яйцо, 200 г С...	4	30 минут	[1329, 21,5, 59, 178,6]	[Главная, Пошаговые рецепты, Выпечка и десерты...	[1291, 155]	https://eda.ru/img/eda/c88x88i/s2.eda.ru/Stati...	[\n Этот торт ещё «Тигровым называю», очень...
3	Американский тыквенный пирог с корицей	[400 г Пшеничная мука, 250 г Сливочное масло, ...	8	2 часа	[652, 9,6, 37,4, 72,4]	[Главная, Пошаговые рецепты, Выпечка и десерты...	[1339, 171]	https://eda.ru/img/eda/c88x88i/s2.eda.ru/Stati...	[\n К сожалению, начинка так и не затвердел...
4	Курица «Пикассо»	[4 штуки Куриная грудка, 2 штуки Лук, 3 штуки ...	4	45 минут	[637, 69,6, 30,7, 21,2]	[Главная, Пошаговые рецепты, Основные блюда]	[1355, 159]	https://eda.ru/img/eda/c88x88i/s2.eda.ru/Stati...	[\n Очень вкусно! Большое спасибо за рецепт!...

Рисунок 4 – Фрагмент из собранных парсером данных

После обработки пропущенных значений в итоговом датафрейме оказалось около четырех тысяч вхождений (3985 размеченных по географии кухни рецепта с ресурса «Еда» и 864 неразмеченных рецепта с «fatsecret РОССИЯ»).

Классификация рецептов по географии кухни. Первым этапом в рассматриваемом пункте стал этап выделения целевой переменной. Непосредственно с меткой о географии рецепта на «Еда» в навигационной панели также представлены такие общие теги, как «Главная», «Пошаговые рецепты», «Завтраки» и т.д. После выделения необходимой информации и помещения ее в отдельный столбец под названием «geo_tag» следующей стадией стал анализ целевой переменной. Всего различных географических тегов оказалось 62. Было принято решение распределить их в подгруппы, объединяющие в себя географии сразу нескольких стран. Всего получилось 6 подгрупп, среди них кухни Восточной, Западной, Южной и Северной Европы, Азии, а также Америки. Визуализация целевой переменной приведена на Рисунке 5.

Данный подход, включающий в себя кластеризацию стран, помог справиться с сильным дисбалансом в данных (наблюдался огромный перевес в сторону русской, итальянской и французской кухонь). Однако даже с учетом этого факта нельзя назвать данное распределение достаточно сбалансированным. Возможным решением в сложившейся ситуации может быть учет количества вхождений каждого класса при построении моделей путем присвоения особых весов.

Следующей стадией в данном пункте стало разбиение датафрейма на обучающую, валидационную и тестовую выборки и загрузка изображений в соответствующие директории.

На вход данная функция принимает ссылку на изображение (использовались ссылки, полученные путем веб-скрапинга и сохраненные в датафрейме), а также путь, по которому необходимо сохранить картинки. Инструментарий функции get_img() был реализован с помощью библиотеки requests, которая отвечает за отправку HTTP-запросов.

После того, как изображения были загружены, их нужно прочитать в виде массива. Здесь нужно было учитывать тот факт, что несмотря на то, что большинство изображений имело размерность 86×86×3, у некоторых, наряду с 3 RGB каналами, присутствовал и четвертый альфа-канал (используется для создания эффекта прозрачности). В процессе предобработки изображений этот фактор был учтен, и картинки приводились к размерности с тремя RGB-каналами.

Далее была проведена работа с текстом, которая включала в себя приведение названий рецептов к нижнему регистру, удалению стоп-слов, токенизацию, лемматизацию и векторизацию текстовых. В качестве инструментов для вышеописанных операций с данными выступили средства, предоставляемые библиотекой nltk.

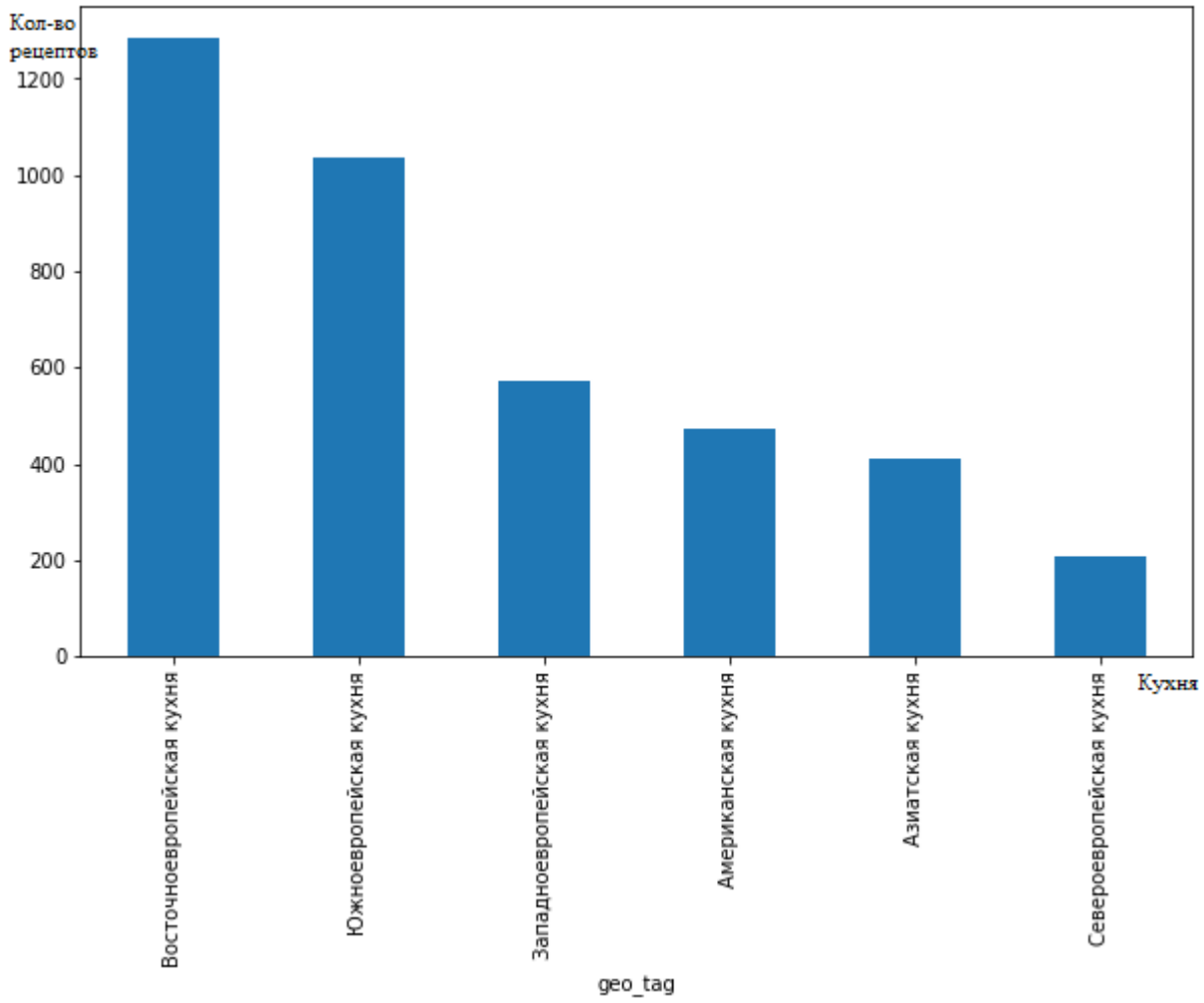

Рисунок 5 – Визуализация целевой переменной

Следующим этапом стало создание модели, которая принимает на вход как изображения рецептов, так и предобработанные их названия. Она создавалась с помощью открытой нейросетевой библиотеки keras. С одной стороны, была реализована модель компьютерного зрения, в основу которой легла модель VGG16 с инициализированными весами ImageNet, с другой – полносвязная нейронная сеть из несколько слоев для классификации текста. В процессе тюнинга гиперпараметров части модели, принимающей на вход изображения, в нее были добавлены два слоя – GlobalAveragePooling2D() и BatchNormalization() – для мультиклассовой классификации. Первый из добавленных слоев является альтернативой Flatten() и вместо того, чтобы просто превращать выходные данные в вектор слой GlobalAveragePooling2D(), подсчитывает среднее число для каждого канала свернутого изображения и передает полученный вектор полносвязным слоям. В дополнение слой BatchNormalization() позволяет проводить нормализацию внутри модели. Часть архитектуры реализованного классификатора (не включая многочисленные начальные слои VGG16) представлена на Рисунке 6.

Описанная выше модель была скомпилирована со следующими параметрами:

– функция потерь: категориальная кросс-сэнтропия;

– оптимизатор: Адам;

– метрика оценивания: точность (accuracy).

Обучение проходило на пяти эпохах. Кривые функции потерь и точности приведены на Рисунке 7.

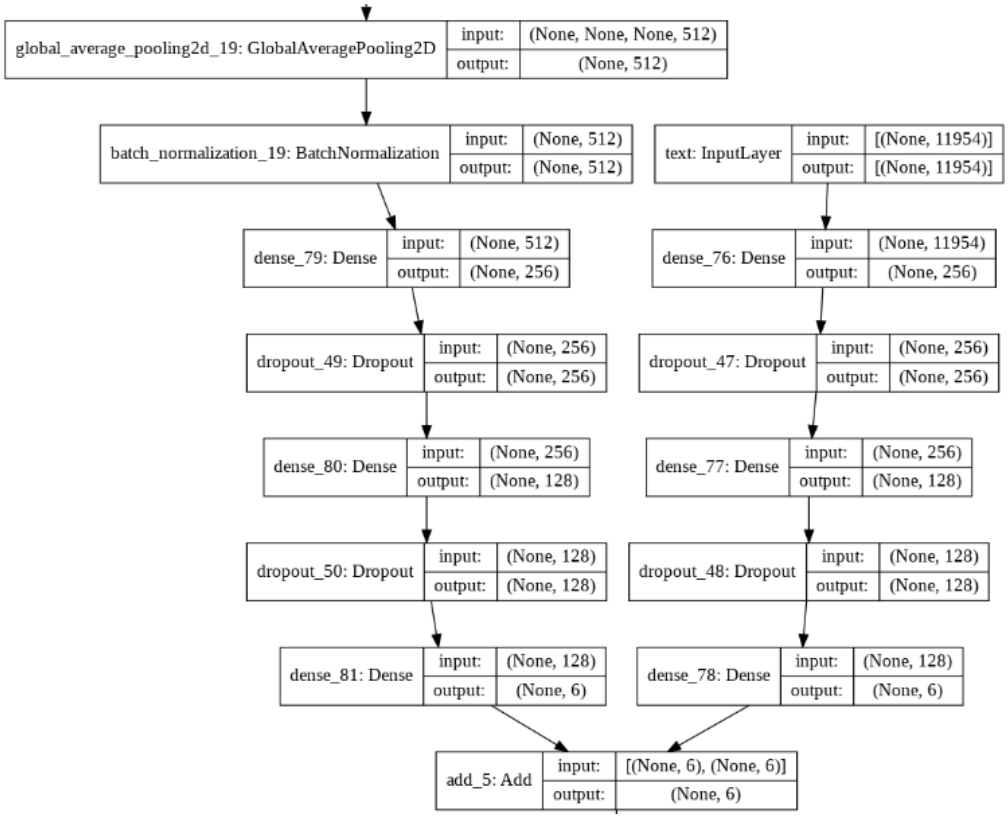

Рисунок 6 – Часть архитектуры реализованного классификатора

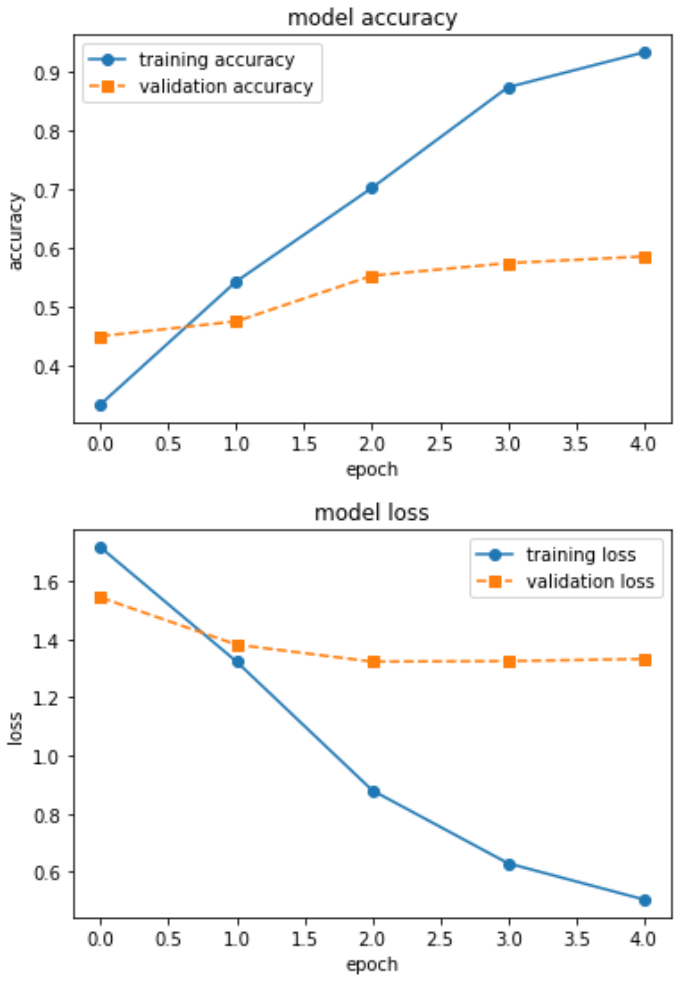

Рисунок 7 – Кривые функции потерь и точности модели

По графикам выше можно заметить, что несмотря на многочисленные слои Dropout(), отвечающие за исключение случайных нейронов на разных итерациях, полностью избежать переобучения не получилось. Модель гораздо лучше подстраивается под обучающую выборку (точность более 90 %) и чувствует себя не очень уверенно на валидационной (точность около 60 %). Однако учитывая специфику данных, такую точность можно считать приемлемой.

В Таблице 2 сравниваются три подхода к классификации рецептов относительно их географии: когда обучение происходит только на картинках, только на тексте и подробно рассмотренный выше способ, включающий в себя первые два подхода.

Таблица 2

Сравнение результатов трех подходов к классификации на валидационной выборке

Используемые данные / Метрика	Только названия рецептов	Только изображения	Названия и изображения
Значение функции потерь	1.75	1.61	1.33
Точность (accuracy)	0.57	0.43	0.59

Анализ общей тональности комментариев. Заключительным этапом агрегации данных для рекомендательной системы персонализированного рациона питания стала оценка общей тональности (или же среднего сентимента) комментариев пользователей для каждого рецепта.

Актуальность добавления нового признака сентимента в датафрейм связана с тем, что, как упоминалось ранее, на ресурсе «Еда» нет как такого рейтинга блюда, а приведена только статистика лайков и дизлайков. Такой подход к оцениванию подразумевает большую лояльность аудитории к тому или иному рецепту, которая в некоторых случаях не совсем оправдана. Отзывы, оставленные пользователями на платформе, говорят о том, что их авторы попробовали приготовить оцениваемое блюдо или уже имеют опыт его приготовления, и готовы поделиться им, подробно изложив возможные плюсы и минусы в своих комментариях.

В результате анализа общей тональности отзывов в датафрейм добавлен новый столбец, содержащий информацию о том, был ли общий сентимент отзывов к конкретному рецепту положительным, нейтральным или отрицательным.

Для реализации данного этапа работы был выбран подход к классификации текста, основанный на правилах. Это обусловлено тем, что методология, при которой используются признаки и обучение моделей на этих признаках, требует наличия тестовых данных для минимизации функции потерь при обучении. Стоит также отметить, что выбранный подход является менее затратным в отношении вычислительных мощностей и его легче интерпретировать и, следовательно, распространить на другие области при необходимости [10].

В качестве модели для анализа среднего сентимента была выбрана модель под названием VADER, реализованная в библиотеке nltk. Аббревиатура VADER может быть расшифрована как словарь валентности для обоснования сентимента (от англ. «Valence Aware Dictionary for Sentiment Reasoning»). Данная модель использует словарь тональностей, который содержит численные показатели позитивности / негативности / нейтральности для каждого слова на основе аннотированных человеком данных. При этом ее отличительной особенностью от других моделей, основанных на правилах (например, от TextBlob), является тот факт, что алгоритм VADER был специально разработан с упором на классификацию текстов из социальных сетей. Примечательно, что данная модель даже превзошла человека в вопросе точности при классификации тональностей записей в Твиттере [10].

В данной работе применение описанного выше алгоритма осложнялось тем фактом, что он поддерживает только английский язык. В связи с этим было принято решение осуществить перевод комментариев на английский язык с помощью библиотеки Ру-

thon, взаимодействующей с APIGoogle переводчика.

После получения предсказаний относительно тональности текста для каждого комментария, оставленного под рецептами, был осуществлен процесс их интерпретации и наиболее часто встречаемый сентимент был выбран в качестве нового признака для каждого рецепта и добавлен в датафрейм.

Статистика по распределению общей тональности в масштабах всего набора данных следующая: нейтральная - 81%, позитивная – 18%, негативная – 1%.

После добавления общего сентимента для каждого рецепта в набор данных он стал насчитывать девять столбцов. Фрагмент получившегося в итоге датафрейма изображен на Рисунке 8.

	recipe_name	measured_ingredients	number_of_servings	cooking_time	nutrients	rating	img_url	geo_tag	sentiment
0	Американские блины	['2 штуки Куриное яйцо', '1 чайная ложка Соль'...	10	20 минут	['242', '6,5', '7,8', '36,7']	5	https://eda.ru/img/eda/c88x88i/s2.eda.ru/Stati...	Американская кухня	pos
1	Блины тонкие на кипятке и молоке	['1 стакан Пшеничная мука', '2 штуки Куриное я...	4	15 минут	['331', '10', '14,7', '40,1']	4	https://eda.ru/img/eda/c88x88i/s1.eda.ru/Stati...	Восточноевропейская кухня	pos
2	Американский тыквенный пирог с корицей	['400 г Пшеничная мука', '250 г Сливочное масл...	8	2 часа	['652', '9,6', '37,4', '72,4']	4	https://eda.ru/img/eda/c88x88i/s2.eda.ru/Stati...	Американская кухня	pos
3	Пирог «Зебра»	['2 стакана Сахар', '5 штук Куриное яйцо', '20...	4	30 минут	['1329', '21,5', '59', '178,6']	4	https://eda.ru/img/eda/c88x88i/s2.eda.ru/Stati...	Западноевропейская кухня	neu
4	Лазанья классическая с мясом	['600 г Мясной фарш', '600 г Соус болоньезе', ...	6	40 минут	['965', '50', '72,6', '23,9']	5	https://eda.ru/img/eda/c88x88i/s2.eda.ru/Stati...	Южноевропейская кухня	neu

Рисунок 8 – Фрагмент датафрейма, получившегося в результате агрегации

Заключение. В результате работы был спроектирован и реализован алгоритм агрегации данных для будущей рекомендательной системы персонализированного рациона питания, которая может быть использована для поддержки принятия решений в диетологии. В ходе его создания были использованы инструменты парсинга для получения содержимого веб-страниц, различные способы обработки текста, а также методы машинного обучения, включая нейронные сети, алгоритмы определения тональности текста.

Реализация алгоритма проходила поэтапно в следующем порядке:

1) создание парсера для получения информации о рецептах с двух веб-ресурсов: «Еда» и «fatsecret РОССИЯ»;

2) формирование датафрейма на полученных путем веб-скрапинга данных;

3) предобработка текстовых данных и представление их в виде векторов;

4) создание обучающей, валидационной и тестовой директорий и размещение в них изображений рецептов;

5) инициализация и обучение модели для классификации рецептов относительно географии приготовления с дальнейшим получением предсказаний для тестовых данных (рецептов с «fatsecret РОССИЯ»);

6) анализ общего сентимента отзывов пользователей для каждого рецепта и формирование в датафрейме нового признака на основе этого анализа.

В результате агрегации был создан датафрейм, содержащий необходимую информацию о рецептах и готовый для передачи в нужном формате JSON на серверную часть проекта по разработке системы поддержки принятия решений в области персонализации рациона питания.

Список источников

1. Data Collection and Pre-processing Techniques [Электронный ресурс]. – Режим доступа: https:// developer.qualcomm.com/software/qualcomm-neural-processing-sdk/learning-resources/ai-ml-android-neural-processing/data-collection-pre-processing (in Eng.).

2. Добронец Б.С., Попова О.А. Численный вероятностный анализ неопределенных данных. – М.:

References

1. Data Collection and Pre-processing Techniques [Electronic resource]. Available at: https:// developer.qualcomm.com/software/qualcomm-neural-processing-sdk/learning-resources/ai-ml-android-neural-processing/data-collection-pre-processing

2. Dobronets B.S., Popova O.A. Numerical Probabilistic Analysis of Uncertain Data. *Moscow. Institut*

Институт космических и информационных технологий, 2014. – 166 с.
3. Патент РФ № RU 2688229 C1, 21.05.2019. Способ агрегирования и преобразования данных и устройство для его реализации / Шведенко В.В., Шведенко В.Н., Грачев В.А., Терская Н.А.

4. Курочкин А.В., Садов В.С. Агрегация и индексирование данных нескольких источников на основе графовой модели в базах данных медицинских экспертных систем // Информатика. 2020. Т. 17. № 3. С. 25–35.
5. Ommen B., Broek T., Hoogh I., Erk M. Systems Biology of Personalized Nutrition // Nutrition Reviews. 2017. С. 579–599. (in Eng.).
6. Image Classification // Tensor Flow [Электронный ресурс]. – Режим доступа: https:// www.tensorflow.org/tutorials/images/classification (in Eng.).
7. Geron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow. – O'Reilly Media, 2017. – 272 с. (In Eng.).
8. Scrapy 2.5 Documentation [Электронный ресурс]. – Режим доступа: https://docs.scrapy. org/en/latest/ (in Eng.).
9. The Selenium Browser Automation Project [Электронный ресурс]. – Режим доступа: https: //www.selenium.dev/documentation/en/ (in Eng.).
10. Hutto C., Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text // Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. 2015. С. 1–3. (in Eng.).

kosmicheskih i informatsionnych technology. 2014. 166 p. (in Russ.).
3. RF patent No. RU 2688229 C1, 05.21.2019. A Method of Aggregating and Transforming Data and a Device for its Implementation / Shvedenko V.V., Shvedenko V.N., Grachev V.A., Terskaya N.A. (in Russ.).
4. Kurochkin A.V., Sadov V.S. Aggregation and Indexing of Data from Several Sources Based on a Graph Model in Databases of Medical Expert Systems. *Informatica*. 2020. Vol. 17. No. 3. pp. 25–35. (in Russ.).
5. Ommen B., Broek T., Hoogh I., Erk M. Systems Biology of Personalized Nutrition. *Nutrition Reviews*. 2017. pp. 579–599.
6. Image Classification. *Tensor Flow* [Electronic resource]. Available at: https://www.tensorflow. org/tutorials/images/classification

7. Geron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow. *O'Reilly Media*. 2017. 272 p.
8. Scrapy 2.5 Documentation [Electronic resource]. Available at: https://docs.scrapy. org/en/latest/

9. The Selenium Browser Automation Project [Electronic resource]. Available at: https: //www.selenium.dev/documentation/en/
10. Hutto C., Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media.* 2015. pp. 1–3.

Статья поступила в редакцию 16.06.2021; одобрена после рецензирования 23.06.2021; принята к публикации 30.06.2021.
The article was submitted 16.06.2021; approved after reviewing 23.06.2021; accepted for publication 30.06.2021.