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 ТОНАЛЬНОСТИ ТЕКСТА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Панченко Алёны Андреевны

Научный руководитель
доцент, к. ф.-м. н. _____ И.Д. Сагаева

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

Саратов 2021

ВВЕДЕНИЕ

В современном мире люди довольно свободны в выражении собственного мнения, которое занимает центральное место практически во всей человеческой деятельности, и оказывают ключевое влияние на наше поведение. Наши собственные убеждения, восприятие и выбор, который мы совершаем, в значительной степени зависят от того, как другие люди видят и оценивают мир. Именно поэтому при принятии того или иного решения, мы часто ищем мнение окружающих по этому поводу. Данное суждение верно не только в отношении людей, но и организаций.

С ростом социальных сетей в Интернете, организации начали придавать первостепенное значение анализу мнений людей, в связи с тем, что это их единственный путь к глубокому пониманию своей клиентской базы и их ожиданий от бренда. Просмотр социальных сетей может помочь организациям понять жалобы и проблемы своих клиентов, что в конечном счете помогает им в будущем улучшать и расширять как сферу своих услуг, так и целевую аудиторию.

Впервые в истории человечества у нас есть огромный объем мнений, записанных в цифровом виде. Без этих данных было бы невозможно провести множество современных исследований. Неудивительно, что зарождение и быстрый рост, такой области знаний как «Анализ тональности» совпадают с таковыми в социальных сетях. Фактически, анализ тональности сейчас находится в центре исследований социальных сетей. Следовательно, исследования в области анализа настроений имеют важное влияние на такие области как политология, экономика, маркетинг, социология, психология, поскольку все они подвержены влиянию мнения людей.

Анализ тональности текста (sentiment analysis) – это область лингвистики, которая занимается выявлением эмоциональной оценки автора по отношению к таким сущностям, как продукты, товары, услуги, организации и прочие [1]. Так анализ тональности можно рассматривать, как метод количественного описания качественных данных, реализуемый путем присваивания некоторых оценок настроения.

Каждый сайт с отзывами на тот или иной объект, содержит огромный объем разнообразных мнений по отношению к нему, поэтому среднестатистический пользователь испытывает трудности с извлечением и обобщени-

ем мнений на подобных сайтах. К тому же по отношению к организациям, мнения о компании могут содержаться не только в социальных сетях (внешние данные), но и внутри компании (внутренние данные), такие как отзывы клиентов, собранные из электронных писем и центров обработки вызовов, или результаты опросов, проведенных организациями. И объем таких данных может быть колоссальным и совершенно не пригодным для обработки их «вручную», из-за больших производственных затрат.

Таким образом, необходимы автоматизированные системы анализа тональности. Для его определения используются разные методы, самым актуальным и часто используемым в наше время является машинное обучение, в том числе нейронные сети.

Целью выпускной квалификационной работы является разработка приложения для анализа тональности текста с использованием нейронных сетей.

Для достижения поставленной цели были сформулированы и решены следующие задачи:

1. Систематизация знаний о нейронных сетях
2. Изучение методов обработки естественного языка
3. Изучение программных возможностей языка python в построении нейронных сетей
4. Построение нескольких видов нейронных сетей для анализа тональности текста и сравнение полученных результатов обучения
5. Построение пользовательского web-приложения для анализа тональности отзывов
6. Проведение тестирования приложения на отзывах из социальных сетей

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

1. Нейронные сети

Нейронная сеть представляет собой сеть искусственных нейронов, которые связаны между собой синапсами.

Формальные нейроны могут быть объединены в сети различными способами. В дипломной работе используются следующие виды нейронных сетей:

— Сверточная нейронная сеть (CNN)
— LSTM-сеть
— GRU-сеть

1.1 Формальный нейрон

Нейрон – это сложная система, математическая модель которого до сих пор не имеет полной реализации [3]. Существует множество моделей, различающихся вычислительной сложностью и сходством с реальным нейроном. Одной из важнейших моделей является формальный нейрон. Формальный нейрон состоит из взвешенного сумматора и нелинейного элемента, параметрами формального нейрона, определяющими его работу, является вид функции активации F.

1.2 Функции активации

Функция активации определяет выходное значение нейрона в зависимости от результата взвешенной суммы входов и порогового значения. Основными функциями активации, используемыми в дипломной работе являются:

— Логическая функция (сигмоида)
— Гиперболический тангенс
— ReLU-функция (rectified linear unit)
— SOFTMAX-функция

1.2 Методы обучения

Перед тем как нейронную сеть можно будет использовать для решения поставленной задачи, ее нужно обучить. Процессом обучения нейронной сети называется подстройка внутренних параметров сети под решаемую задачу. Алгоритмы обучения итеративные, каждый шаг алгоритма называется эпохой или циклом.

Существует два метода обучения нейронных сете: обучение с учителем и без.

В данной работе используется обучение с учителем. Обучением с учите-

лем (supervised learning) — метод, при котором на момент обучения известна сама задача и ее решение, то есть у нас есть известные входные и выходные вектора сети.

Достоинства данного подхода:

— хорошая точность при определении тональности
— на основе обучающей выборки классификатор самостоятельно выделяет признаки, влияющие на тональность. Таким образом, проблема зависимости от предметной области решается с помощью использования обучающей выборки из той же области
— существует множество способов улучшить точность

Недостатки данного подхода:

— требуется размеченная обучающая выборка
— результаты могут сильно зависеть от выбранного алгоритма, его параметров, обучающей выборки

1.3 Подбор гиперпараметров

Перед тем как перейти к построению модели нейронной сети в python необходимо установить ее будущую архитектуру: количество слоев, функции активации, метод оптимизатора и пр., которые носят названия гиперпараметров сети.

Для этой задачи в дипломной работе используется метод байесовской оптимизации для поиска гиперпараметров нейронной сети.

Байесовский подход основан на идее, что для выбора лучшей области пространства гиперпараметров следует учитывать историю уже рассмотренных точек, в которых уже были получены результаты обучения модели [5].

Байесовский алгоритм оптимизации имеет два основных компонента:

— Вероятностная модель функции: строится вероятностная модель функции $f[x]$, используя информацию о точках в которых уже получено значение функции
— Функция получения: выбирается следующая точка для анализа

2 Анализ тональности текста

Анализ тональности - это автоматический анализ мнений и эмоционально окрашенной лексики, имеющиеся в тексте [3].

Задача анализа тональности текста является задачей классификации. Алгоритм классификации тренируется на основе обучающей выборки, состо-

ящей из документов, классы которых заранее известны.

В общем виде задача текстовой классификации определяется следующим образом:

Пусть существует описание документа $d \in X$, где X –– множество всех документов, и фиксированный набор меток $Y = \{y_1, y_2, ..., y_N\}$. Из обучающей выборки (множества документов с заранее известными метками) $D = \{(x, y)|(x, y) \in X \times Y\}$ с помощью метода обучения Γ необходимо получить классифицирующую функцию (или классификатор) $\Gamma(D) = \gamma$, которая отображает документы в классы $\gamma : X \to Y$. В задаче определения тональности множество Y состоит из двух элементов {положительный, отрицательный} [4].

2.1 Обработка естественного языка. Предварительная обработка

Обработка естественного языка (НЛП) - междисциплинарная область, лежащая на стыке лингвистики и информатики (в частности, искусственного интеллекта) [5].

Предварительная обработка текстовых данных является важным шагом, поскольку она делает необработанный текст готовым для анализа, т. е. становится проще извлекать информацию из текста и применять к нему алгоритмы машинного обучения. Если пропустить этот шаг, то нейронная сеть будет работать с большими и несогласованными данными. Цель этого шага состоит в том, чтобы очистить текст от ненужных символов или слов, который не влияют на результаты анализа тональности.

В данной работе проводятся следующие операции с текстом:

— установка всех символов в нижний регистр
— удаление знаков пунктуации
— удаление сокращения слов
— коррекция орфографии
— удаление шума
— лемматизация
— удаление стоп-слов

2.2 Результаты работы нейронных сетей

В результате обучения и тестирования описанных в предыдущих разделах нейронных сетей, были получены следующие результаты (рисунок 1):

	Количество классов	Предварительная обработка	
		Нет	Есть
CNN	2	87,97%	95.87%
CNN	5	41,09%	74.15%
LSTM	2	90,06%	97,64%
LSTM	5	40,67%	78,89%
GRU	2	89,45%	96,97%
GRU	5	42,54%	76,34%

Рисунок 1 – Сравнение работы нейронных сетей

Из полученных результатов можно сделать вывод что предварительная обработка позволяет повысить уровень точности обучения, в более сложной задаче классификации на 5 классов приблизительно на 30%, при классификации на два класса точность увеличивается приблизительно на 7%.

По результатам, полученным ранее, было выявлено, что из рассмотренных сетей: LSTM-сеть лучше всех подходит для качественного анализа тональности текстов, поэтому в основе построенного приложения будет лежать именно она.

3 Построение пользовательского web-приложения

Приложение построено с помощью программных возможностей языка python и Streamlit. Streamlit — это веб-фреймворк, предназначенный для создания интерфейсов приложений и имеющий широкий круг возможностей, таких как работа с нейронными сетями [6].

Алгоритм работы построенного приложения представлен на рисунке 2:

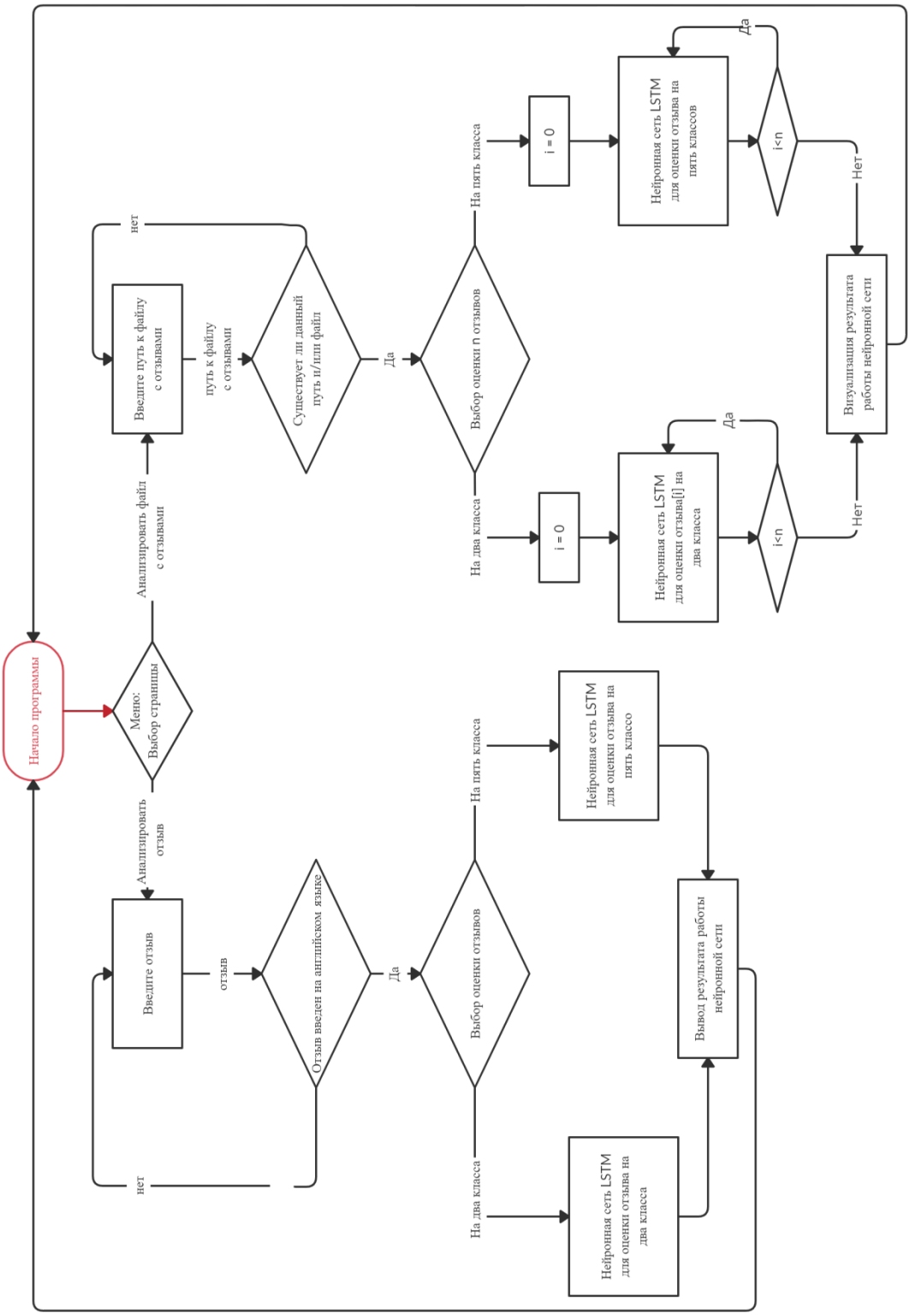

Рисунок 2 – Алгоритм работы приложения

Таким образом приложение может анализировать отзывы на два и пять классов:

— введенные вручную в приложении

— находящиеся в файле

Результаты работы приложения на отзыве, написанном вручную, для двух методов представлены на рисунках 3 и 4.

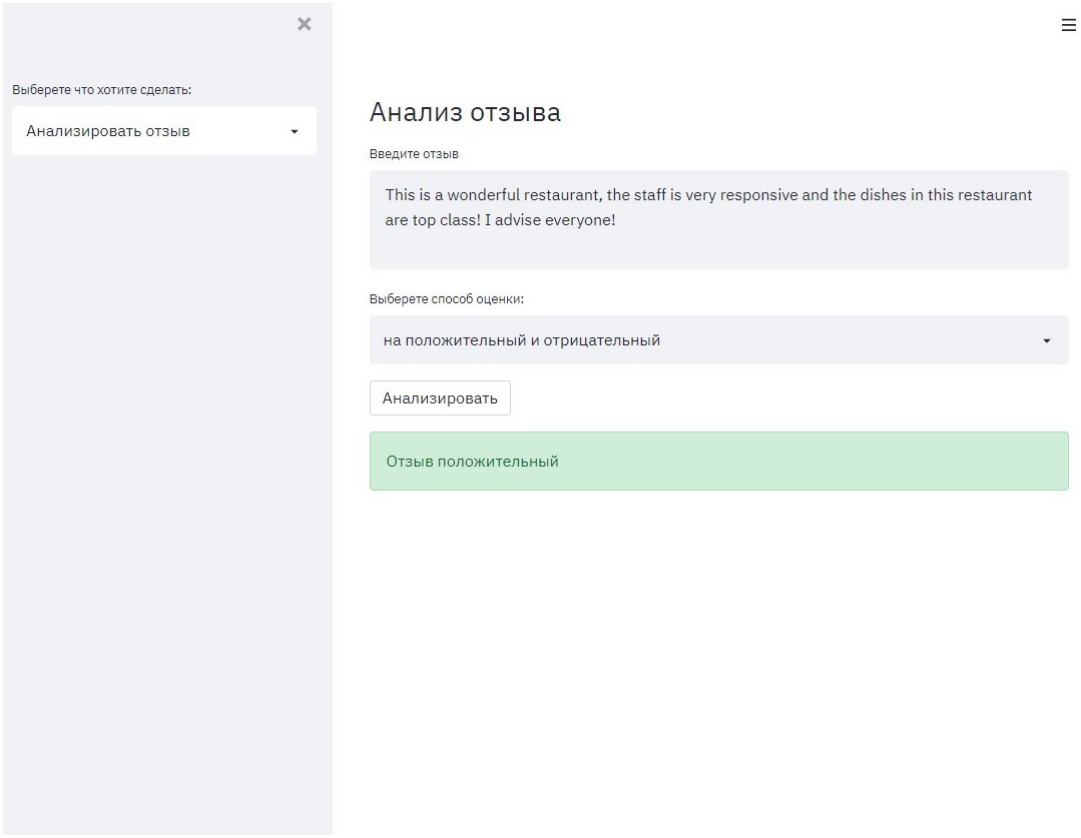

Рисунок 3 – Результат работы приложения при оценке отзыва на два класса

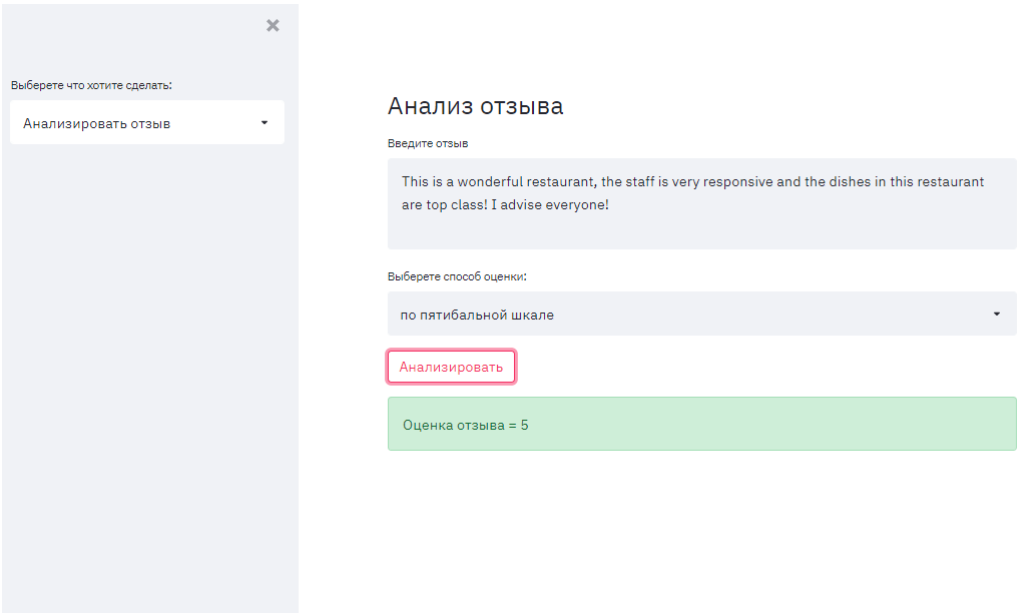

Рисунок 4 – Результат работы приложения при оценке отзыва на пять классов

Набором данных для тестирования отзывов из файла, были взяты 3 выборки с сайта «kaggle» [7], который предоставляет наборы данных из различных источников. Для тестирования возьмем отзывы:

— на магазин музыкальных инструментов – набор данных содержит 10 261 отзыв [8]
— о ресторане Beyond Flavours – набор данных содержит 9819 отзывов [9]
— на компанию Universal Studio - набор данных содержит 50904 отзыв [10]

Количество отзывов в наборах данных разделенные на классы, представлены на рисунке 5:

набор данных	оценка						
	на 2 класса		на 5 классов				
	положительно	отрицательно	1	2	3	4	5
1	9022	1239	217	250	772	2084	6938
2	6210	3609	1732	684	1193	2378	3832
3	41716	9188	1973	1986	5229	13514	28202

Рисунок 5 – Количество отзывов в наборах данных разделенные на классы

Результаты работы программы на этих наборах данных представлены на рисунке 6:

набор данных	оценка						
	на 2 класса		на 5 классов				
	положительно	отрицательно	1	2	3	4	5
1	9016	1245	315	280	864	2045	6757
2	6198	3621	1761	693	1217	2401	3747
3	41693	9211	2136	2067	5317	13998	27386

Рисунок 6 – Результаты работы программы

Из полученных данных можно сделать вывод, что оценка отзывов на 2 класса работает точнее, чем оценка на 5 классов.

В самом приложении можно также увидеть результат работы в виде графика и таблицы (рисунки 7 и 8).

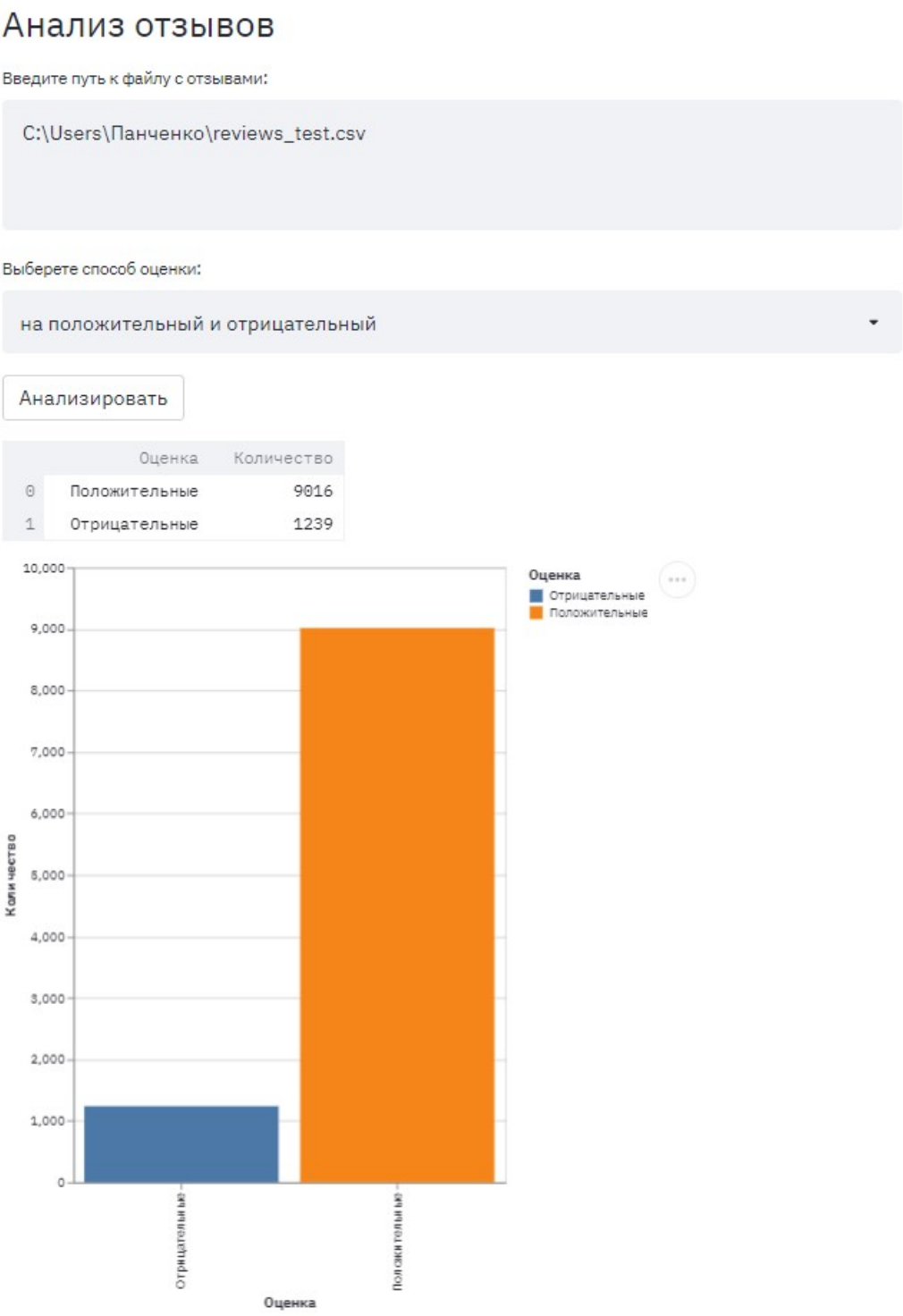

Рисунок 7 – Результат работы приложения при оценке файла с отзывами на два класса

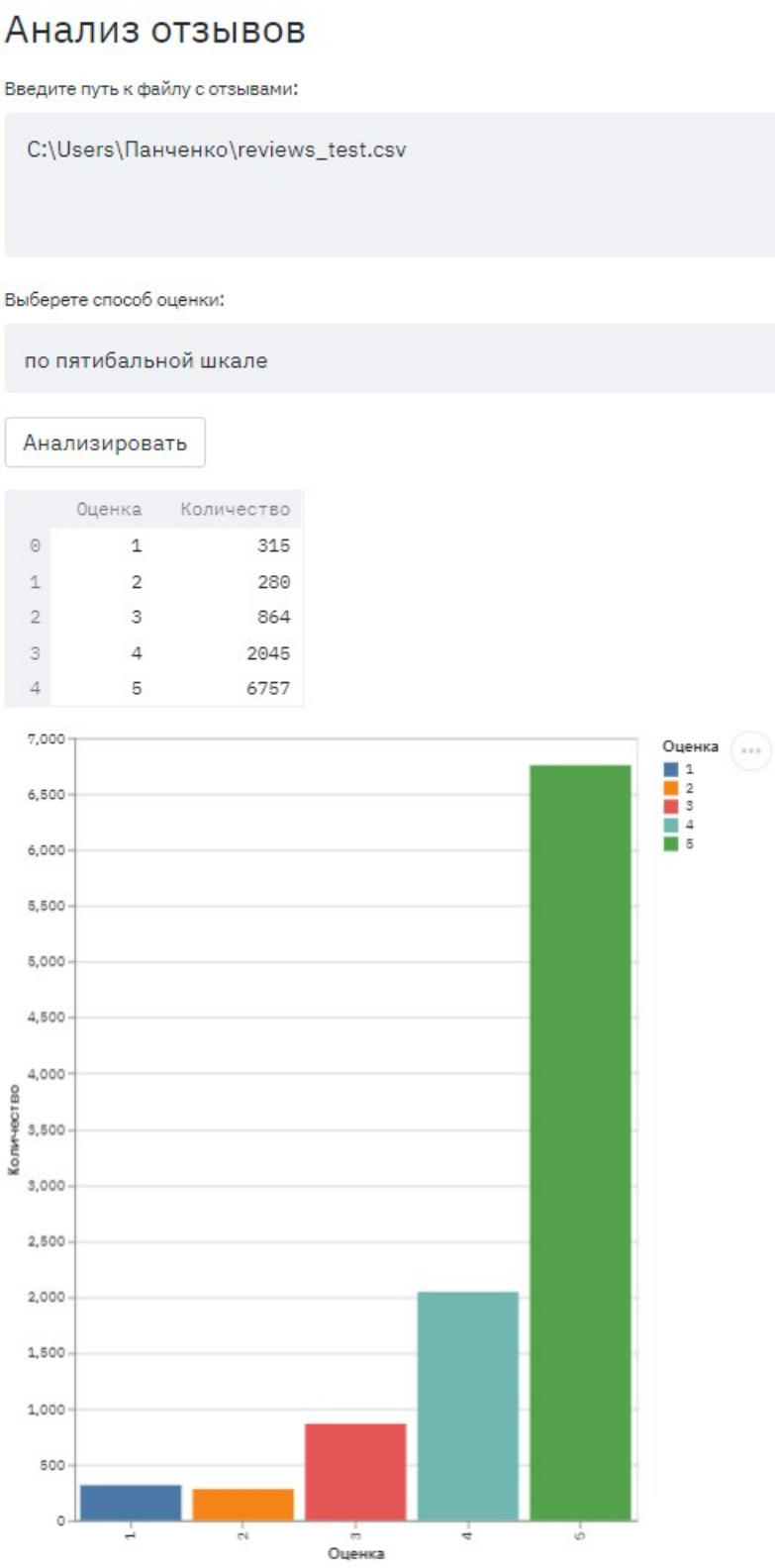

Рисунок 8 – Результат работы приложения при оценке файла с отзывами на пять классов

ЗАКЛЮЧЕНИЕ

В рамках выпускной квалификационной работы были решены следующие задачи:

— изучены методы обработки естественного языка и систематизированы знания о работе нейронных сетей. В результате применения этих методов к построенным в работе нейронным сетям было выявлено улучшение качества работы свыше чем на 30% при разделении отзывов на пять классов, и на 7% процентов при разделении на два класса.

— построены и протестированы LSTM, GRU и CNN-сети. Для разработки сетей использовались библиотеки языка Python В результате тестирования было показано что LSTM-сети показывают лучший результат при решении задачи анализа тональности.

— разработано и протестировано пользовательское web-приложение, которое позволяет проводить анализ тональности введенных пользователем отзывы в автоматическом режиме.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Pang B. & Lee L. Opinion Mining and Sentiment Analysis / Foundations and Trends in Information Retrieval, v.2 n.1-2, January, 2008 - 135 с.

2 Хайкин, С. Нейронные сети: полный курс / С. Хайкин. – М.: Вильямс, 2006. – 1104 с.

3 Sentiment Analysis: What's with the Tone? [Электронный ресурс] : [сайт]. — URL:https://www.infoq.com/articles/sentiment-analysis-whats-with-the-tone/ — Загл. с экрана. (дата обращения 13.03.2021 г.)

4 Simon Tong и Daphne Koller. "Support vector machine active learning with applications to text classification". В: The Journal of Machine Learning Research (2002), с. 45—66

5 Bird, S., Klein, E., and Loper, E. (2009). Natural Language Processing With Python. California, CA: O'Reilly Media.

6 Streamlit [Электронный ресурс] : [сайт]. — URL: https://streamlit.io/ — Загл. с экрана. (дата обращения 18.05.2021 г.)

7 Kaggle [Электронный ресурс] : [сайт]. — URL: https://www.kaggle.com/ — Загл. с экрана. (дата обращения 29.05.2021 г.)

8 Musical_instruments_reviews.csv [Электронный ресурс] : [сайт]. — URL: https://www.kaggle.com/eswarchandt/amazon-music-reviews?select=Musical_instruments_reviews.csv — Загл. с экрана. (дата обращения 29.05.2021 г.)

9 Restaurant reviews.csv [Электронный ресурс] : [сайт]. — URL: https://www.kaggle.com/batjoker/zomato-restaurants-hyderabad?select=Restaurant+reviews.csv — Загл. с экрана. (дата обращения 29.05.2021 г.)

10 universal_studio_branches.csv [Электронный ресурс] : [сайт]. — URL: https://www.kaggle.com/dwiknrd/reviewuniversalstudio — Загл. с экрана. (дата обращения 29.05.2021 г.)