МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №6
по дисциплине «Искусственные нейронные сети»
Тема: Прогноз успеха фильмов по обзорам

Студентка гр. 8383 _____ Максимова А.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2021

Цель работы

Прогноз успеха фильмов по обзорам (Predict Sentiment From Movie Reviews).

Задачи

- Ознакомиться с задачей классификации
- Изучить способы представления текста для передачи ИНС
- Достигнуть точность прогноза не менее 95%

Требования

1. Построить и обучить нейронную сеть для обработки текста
2. Исследовать результаты при различном размере вектора представления текста
3. Написать функцию, которая позволяет ввести пользовательский текст (в отчете привести пример работы сети на пользовательском тексте)

Основные теоретические положения

Сентимент-анализ: используется для анализа массивов сообщений и иных данных и определения того, как они эмоционально окрашены - негативно, положительно или нейтрально (последнее не обязательно). То есть ставится задача классификации.

Задача классификации: задача, в которой имеется множество объектов, где каждый объект, исходя из его свойств и параметров, можно отнести к конкретному классу. Таким образом, нейронная сеть, получая на вход объект, возвращает на выход вероятность (дискретную величину) его принадлежности к каждому из классов. В процессе обучения ИНС стремимся достигнуть результата, когда вероятность принадлежности к правильному классу имеет значение единицы, к другим классам - нуля.

Задача бинарной классификации: задача классификации элементов заданного множества в две группы (предсказание, какой из

групп принадлежит каждый элемент множества) на основе правил классификации.

Датасет IMDb:

Датасет, состоящий из 50 000 отзывов на фильмы от пользователей, помеченных как положительные (1) и отрицательные (0).

- Рецензии предварительно обрабатываются, и каждая из них кодируется последовательностью индексов слов в виде целых чисел.
- Слова в обзорах индексируются по их общей частоте появления в датасете. Например, целое число "2" кодирует второе наиболее частое используемое слово.
- Датасет разделен на два набора: 25 000 для обучения и 25 000 на тестирование.

Выполнение работы

1. Были импортированы все необходимые для работы классы и функции.

```
import matplotlib
import matplotlib.pyplot as plt
import numpy as np

from keras.utils import to_categorical
from keras import models
from keras import layers
```

2. Был загружен встроенный в Keras датасет IMDb. Для изменения исходного разбиения (50/50) отношения обучающих и контрольных данных, загруженные данные были объединены с помощью метода concatenate() для последующего разделения в пропорции 80/20.

```
# загрузка данных из IMDb - нужно указать максимальное кол-во слов, исп-ое для анализа
max_words = 10000  # 10к самых популярных слов, исп-х для анализа
(training_data, training_targets), (testing_data, testing_targets) = imdb.load_data(num_words=max_words)
# объединение данных для последующего разделения в других пропорциях 5:5->8:2
# np.concatenate() - соед. массивы вдоль указанной оси
data = np.concatenate((training_data, testing_data), axis=0)
targets = np.concatenate((training_targets, testing_targets), axis=0)
```

3. В результате изучения датасета было выяснено, что:

- датасет содержит 9998 уникальных слов;
- все содержащиеся в датасете данные можно разделить на 2 категории: 0 (отрицательный отзыв) или 1 (положительный отзыв);
- средняя длина отзыва - 234 слова;
- один элемент датасета - обзор: содержит в себе список чисел, каждое из которых представляет одно слово исходного отзыва (токенизация на уровне слов);
- для кодирования текста отзыва используется числовое кодирование: слово заменяется на частоту его появления.

4. С помощью метода imdb.get_word_index() был загружен словарь, используемые при кодировании данных (ключ - слово, значение - частота, с которой слово встречается в обзорах). Был создан реверсивный словарь, определяющий по числу закодированное слово. Был выполнен перевод закодированного текста в исходное состояние, с учетом использования в Keras служебных символов (0 - символ заполнитель, 1 - начало последовательности, 2 - неизвестное слово). Так как используется ограничение максимального количества слов, используемых для анализа данных, то возможна ситуация, когда закодированное слово может не находится в словаре, в таком случае такое слово заменяется символом #. Также можно заметить, что все слова раскодированного отзыва приведены к нижнему регистру, все знаки препинания удалены.

5. Так как обзоры на фильмы состоят из разного количества слов, а полносвязная нейронная сеть может работать только с данными одной длины, было необходимо модифицировать исходные данные. Для этого каждый обзор был векторизован и заполнен нулями, так, чтобы его длина была равна 10 000. Также было выполнено преобразование переменных к типу float.

```
# подготовка данных
def vectorize(sequences, dimension=10000):
    results = np.zeros((len(sequences), dimension))  # массив из 50к строк, длина каждой 10к
    for i, sequence in enumerate(sequences):         # индекс и значение
        results[i, sequence] = 1                     # 1 - число с данным индексом встречается, 0 - не встречается
    return results

data = vectorize(data)
targets = np.array(targets).astype("float32")
```

В результате каждый обзор изначально написанный в виде слов, составляющих предложения, был преобразован в числовые последовательности, а мы - выровняли длины каждого обзора до 10 тысяч, и заполнили эти списки 0, если в данном обзоре не было слова из 10000 самых популярных слов, и 1, если было. В итоге обзоры имеют следующий вид, представленный на рисунке ниже.

```
[0. 1. 1. ... 0. 0. 0.]
```

6. После была определена функция для создания модели ИНС прямого распространения, состоящая из 7 слоев: первый (входной) содержит 10000 нейронов; второй, четвертый, шестой (скрытые) - содержат по 50 нейронов, используется полулинейная функция активации Relu: max(0, x); третий, четвертый, пятый (скрытые) - Dropout, используемый для предотвращения переобучения ИНС (вероятность исключения нейронов из сети находится в интервале от 0.2 до 0.5); седьмое слой (выходной) содержит 1 нейрон, так как задача бинарной классификации, функция активации Sigmoid: $\frac{1}{1+e^{-x}}$, выдающая значения из диапазона от 0 до 1.

```
model = models.Sequential() #последовательная

model.add(layers.Dense(50, activation='relu', input_shape=(10000, )))
model.add(layers.Dropout(0.3, noise_shape=None, seed=None))

model.add(layers.Dense(50, activation='relu'))
model.add(layers.Dropout(0.2, noise_shape=None, seed=None))

model.add(layers.Dense(50, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))  # бинарная классификация
```

7. Были определены параметры обучения сети: в качестве функции потерь используется "binary_crossentropy" - функция, которая в основном используется при бинарной классификации, метрика качества работы сети - точность, оптимизатор - "adam".

```
model.compile(
    optimizer="adam",
    loss="binary_crossentropy",
    metrics=["accuracy"]
)
```

8. После было запущено обучение сети с помощью метода fit (адаптирует модель под обучающие данные).

```
history = model.fit(
    train_x, train_y,
    epochs=2,
    batch_size=500,
    validation_data=(test_x, test_y)
)
```

9. Была проведена оценка работы модели: 89%

```
print(np.mean(history.history["val_accuracy"]))
```

10. Для проверки работоспособности программы она была запущена. В процессе обучения нейронной сети отображаются две величины: loss - потери сети и accuracy - точность на обучающих данных. Как видно, сеть уже показывает хорошие результаты.

```
Epoch 1/2
80/80 [==============================] - 4s 40ms/step - loss: 0.5146 - accuracy: 0.7266 - val_loss: 0.2600 - val_accuracy: 0.8958
Epoch 2/2
80/80 [==============================] - 1s 13ms/step - loss: 0.2096 - accuracy: 0.9227 - val_loss: 0.2672 - val_accuracy: 0.8951
0.8954499959945679
```

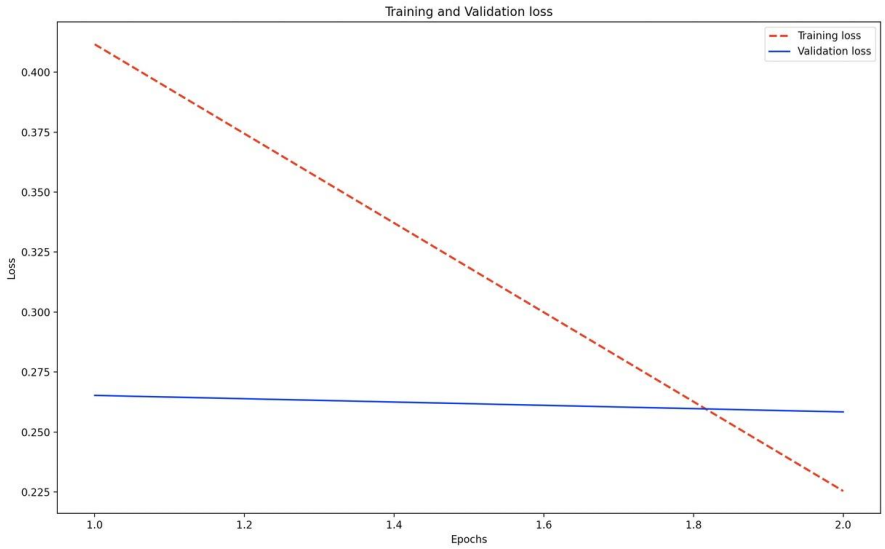
11. Для построения графиков ошибок и точности в ходе обучения сети были написаны следующие функции:

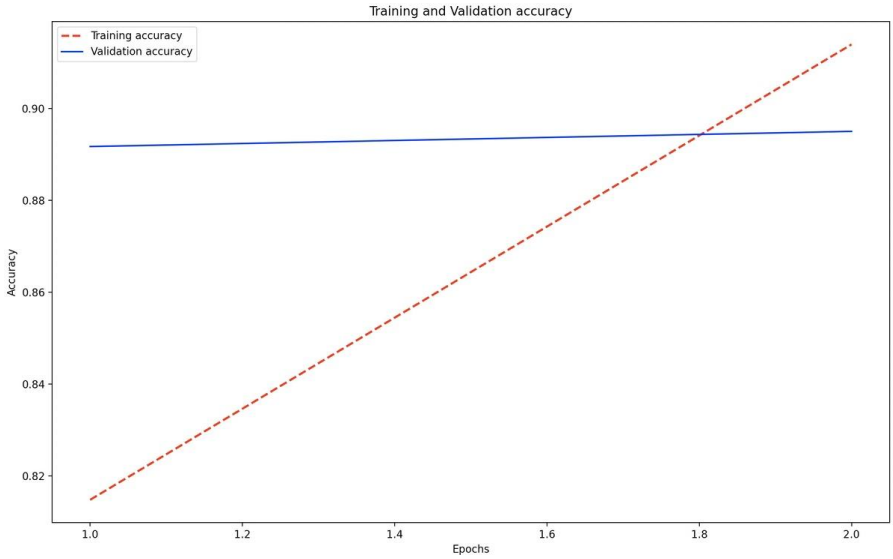
```python
def plot_loss(loss, val_loss, epochs):
    plt.plot(epochs, loss, label='Training loss', linestyle='--', linewidth=2, color="red")
    plt.plot(epochs, val_loss, 'b', label='Validation loss', color="blue")
    plt.title('Training and Validation loss')  # оглавление на рисунке
    plt.xlabel('Epochs')
    plt.ylabel('Loss')
    plt.legend()
    plt.show()

def plot_acc(acc, val_acc, epochs):
    plt.clf()
    plt.plot(epochs, acc, label='Training accuracy', linestyle='--', linewidth=2, color="red")
    plt.plot(epochs, val_acc, 'b', label='Validation accuracy', color="blue")
    plt.title('Training and Validation accuracy')
    plt.xlabel('Epochs')
    plt.ylabel('Accuracy')
    plt.legend()
    plt.show()
```

Модификация исходной модели

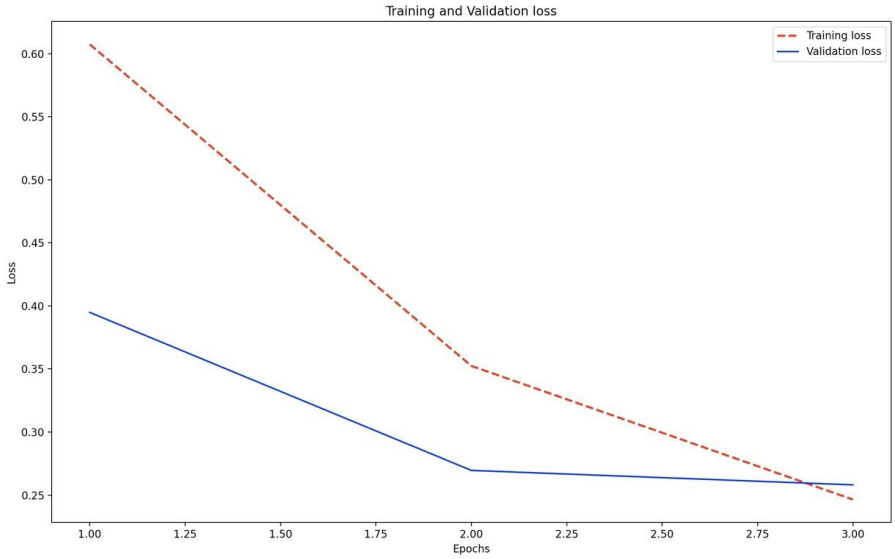
Достигнуть точности прогноза не менее 95% не получилось, вероятно, по причине относительно наивной обработки входных данных и простоте модели нейронной сети, но получилось сделать модель менее переобучаемой, за счет увеличения количества эпох обучения, размера батчей и увеличения вероятности отсеивания нейронов на слоях Dropout. Графики до и после модификации программы точности и потерь нейронной сети представлены ниже.

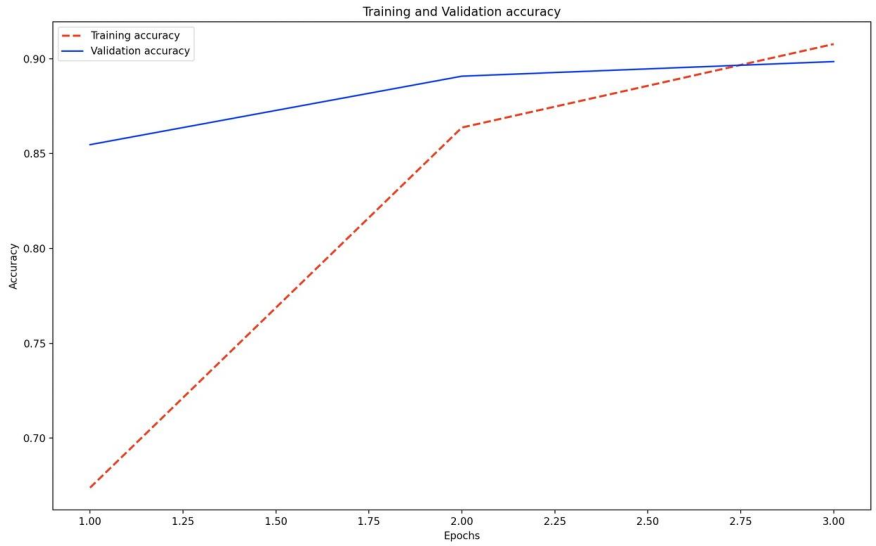
Графики потерь и точности сети на обучающих и контрольных данных до модификации ИНС:





Графики потерь и точности сети на обучающих и контрольных данных после модификации ИНС:





Исследование результатов при различном размере вектора представления текста

Эксперимент 1: *размер вектора представления текста 10000*

Табл. 1 - Результаты запуска при размере 10000

Потери сети на обучающих данных	Потери сети на контрольных данных
0.2505	0.2577
Точность сети на обучающих данных	Точность сети на контрольных данных
91%	90%
Переобучения достигается на 3 эпохе	
Оценка работы модели 88%	

График потерь нейронной сети на обучающих и контрольных данных:

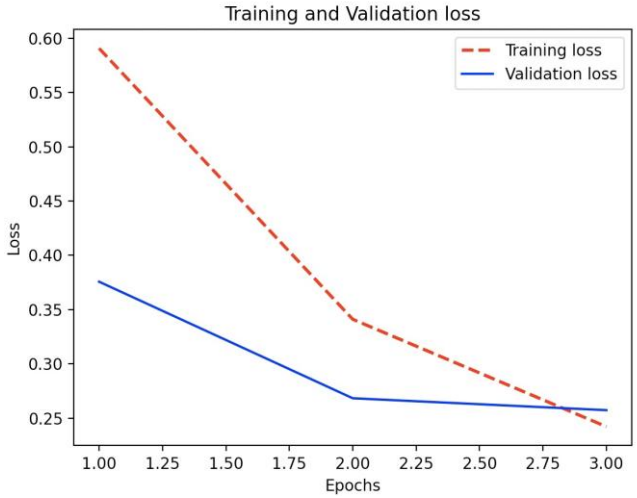
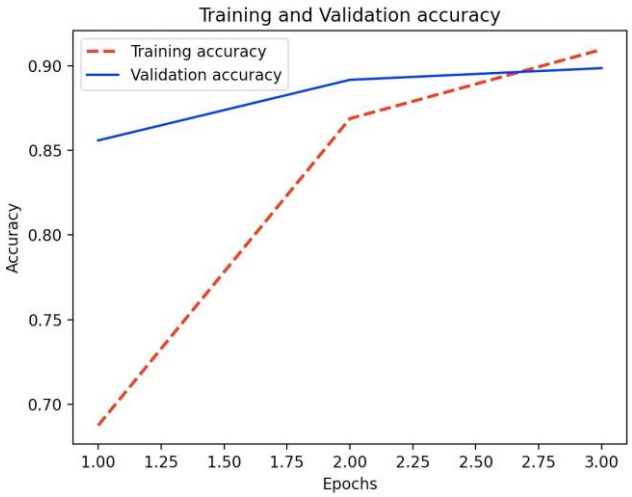


График точности нейронной сети на обучающих и контрольных данных:



Эксперимент 2: *размер вектора представления текста 15000*

Табл. 2 - Результаты запуска при размере *15000*

Потери сети на обучающих данных	Потери сети на контрольных данных
0.2569	0.2693
Точность сети на обучающих данных	Точность сети на контрольных данных
91%	89%
Переобучения достигается между 2 и 3 эпохами	
Оценка работы модели 87%	

График потерь нейронной сети на обучающих и контрольных данных:

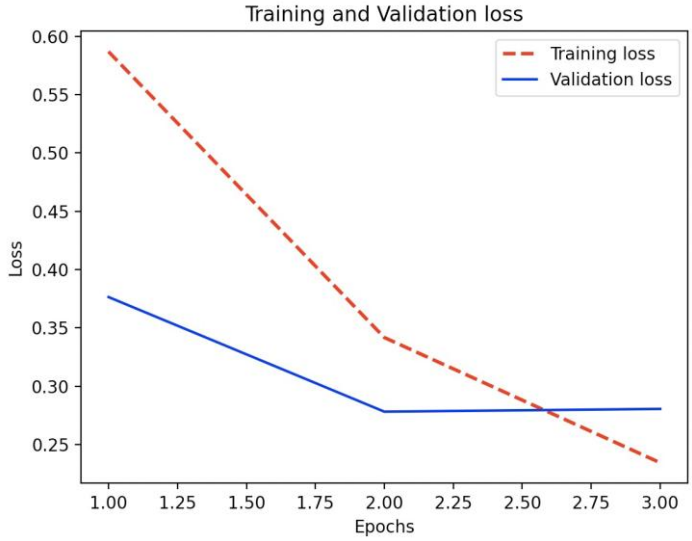
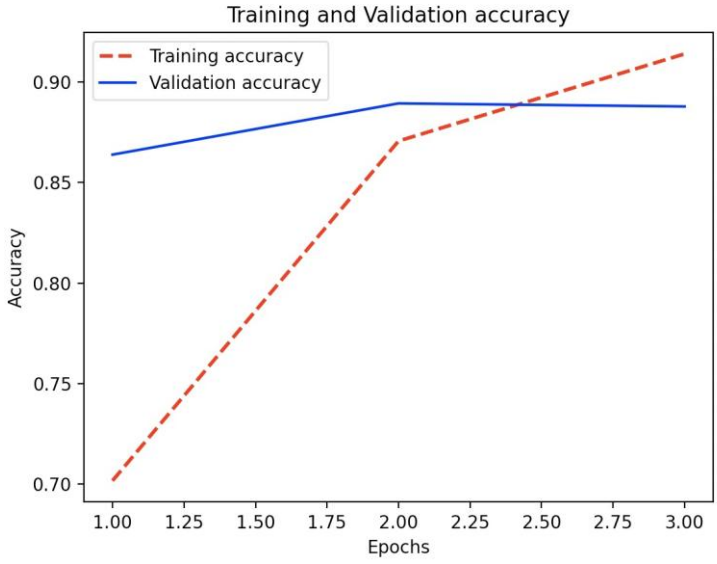


График точности нейронной сети на обучающих и контрольных данных:



Эксперимент 3: *размер вектора представления текста 20000*

Табл. 3 - Результаты запуска при размере *20000*

Потери сети на обучающих данных	Потери сети на контрольных данных
0.2865	0.2766
Точность сети на обучающих данных	Точность сети на контрольных данных
90%	89%
Переобучения достигается на 3 эпохе	
Оценка работы модели 87%	

График потерь нейронной сети на обучающих и контрольных данных:

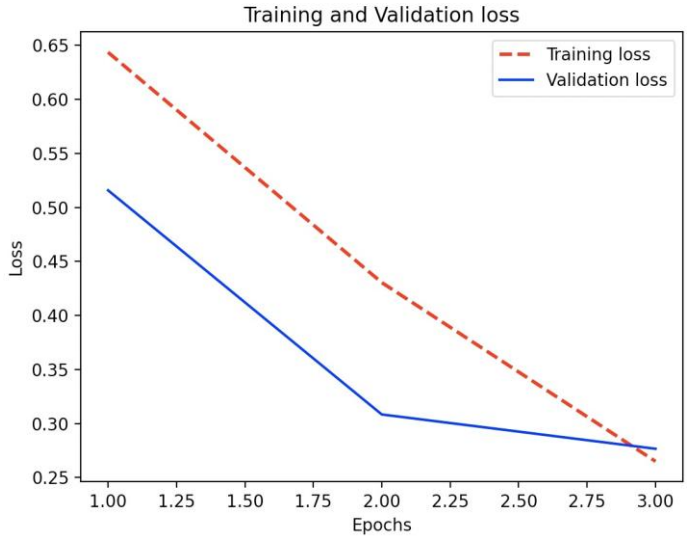
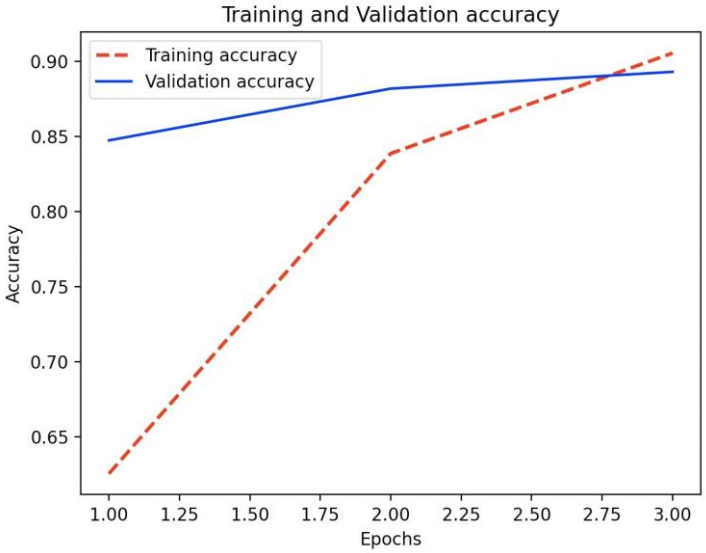


График точности нейронной сети на обучающих и контрольных данных:



Эксперимент 4: *размер вектора представления текста 7500*

Табл. 4 - Результаты запуска при размере *7500*

Потери сети на обучающих данных	Потери сети на контрольных данных
0.2619	0.2568
Точность сети на обучающих данных	Точность сети на контрольных данных
90%	90%
Переобучения достигается на 3 эпохе	
Оценка работы модели 88%	

График потерь нейронной сети на обучающих и контрольных данных:

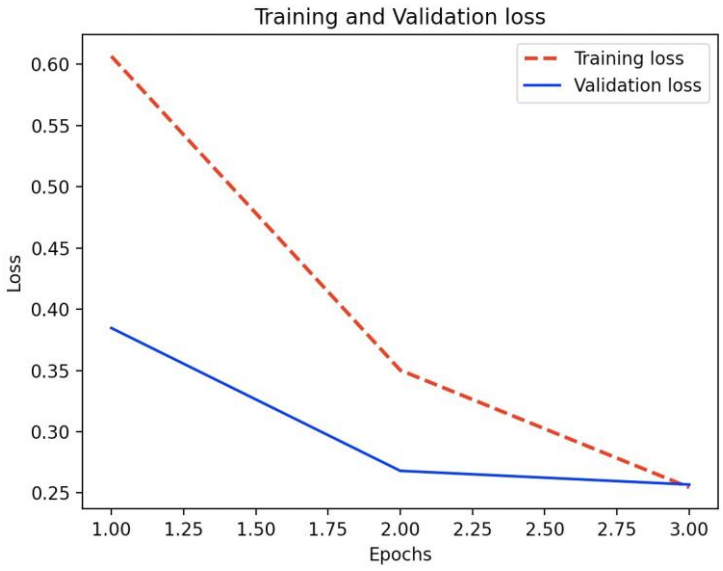
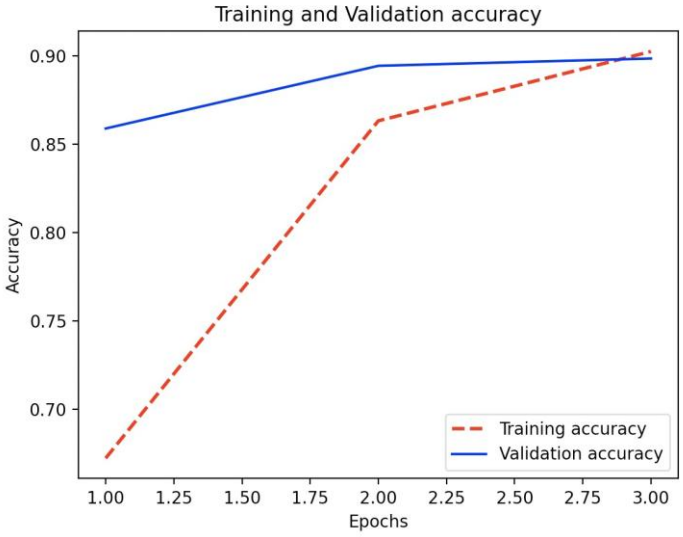


График точности нейронной сети на обучающих и контрольных данных:



Эксперимент 5: *размер вектора представления текста 2500*

Табл. 5 - Результаты запуска при размере *1500*

Потери сети на обучающих данных	Потери сети на контрольных данных
0.3164	0.2863
Точность сети на обучающих данных	Точность сети на контрольных данных
87%	87%
Переобучения достигается на 2 эпохе	
Оценка работы модели 86%	

График потерь нейронной сети на обучающих и контрольных данных:

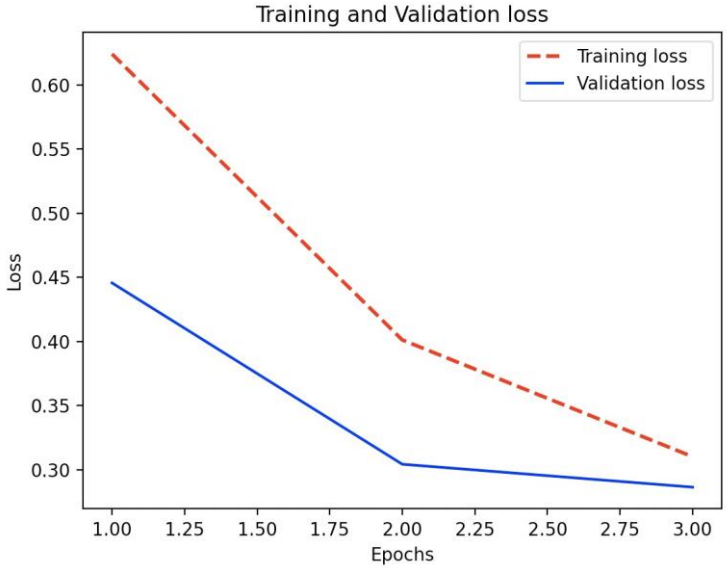
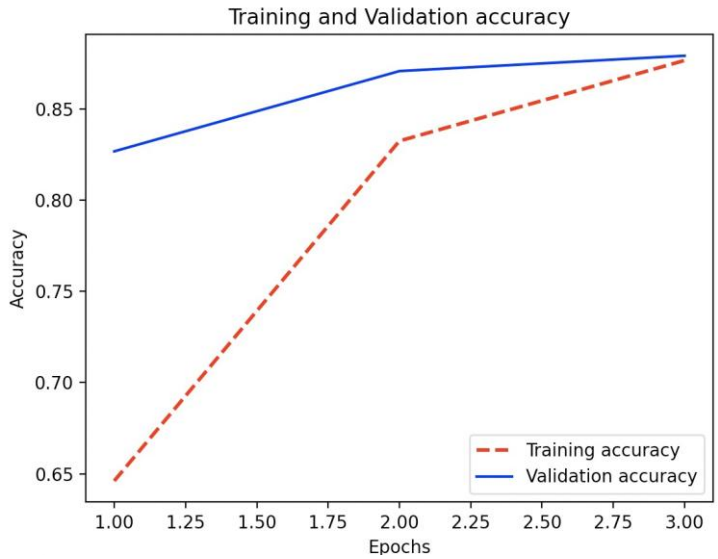


График точности нейронной сети на обучающих и контрольных данных:



Выводы

Как видно из проведенных экспериментов, при увеличении размера вектора представления текста значения потерь и точности сети остаются практически неизменными, при этом потребление ресурсов увеличивается.

При небольшом изменении размера вектора значения ошибки и точности сети остаются постоянными.

При использовании размера до 500, точность сети упала на 10%. При увеличении размера до 35000 точность сети уменьшилась на 2 % от исходной.

Таким образом, можно сделать вывод, что небольшой объем вектора представления текста негативно влияет на работу нейронной сети. Увеличение же этого показателя не приводит к весомым улучшениям, а потому, является излишним.

Оптимальные значения длины вектора представления текста в данной задаче лежат в интервале от 7500 до 15000.

Тестирование функции, позволяющей вводить пользовательский текст

Тестирование производилось на англоязычных отзывах разной длины и настроения.

Тест 1

Исходный текст:

Several days ago I watched a British crime thriller the Legend. Brian Helgeland is the scriptwriter and the director of the film as well. It is adapted from a book The Profession of Violence: The Rise and Fall of the Kray Twins which is based on a real story.

The film tells about the life of twin brothers, Reggie and Ronnie Kray, who were violent and vulgar gangsters. They also were iconic figures in the criminal environment of London in 1960s. They headed the most influential gang of bandits of the East-End.

They strongarmed, attempted assassination and killed several crime bosses. They also owned a nightclub where came even Hollywood stars. However, it is not easy to be a criminal and it is impossible to give up the crime. It destroyed their lives and they both ended in jail and died alone having lost everything they had and loved.

The film is very fascinating and it is a pleasure to watch it even if the plot is rather cruel and heavy. Each part of the movie completes the picture. When you watch it, you dig into the atmosphere by means of sounds, music, costumes, and decorations.

Everything is well-orchestrated and the actor that plays the main role is extremely talented. Tom Hardy plays both

brothers which have absolutely different characters and personalities. And Hardy acts fantastically!

The Legend is definitely one of the greatest films I have ever seen. It can win all hearts!

Результат:

```
Enter the name of the file to read: usr_text/positive1.txt
Positive feedback
```

Тест 2

Исходный текст:

The start to the Harry Potter film series is filled with visual splendor, valiant heroes, spectacular special effects, and irresistible characters. It's only fair to say that it's truly magical.

Результат:

```
Enter the name of the file to read: usr_text/positive2.txt
Positive feedback
```

Тест 3

Исходный текст:

As they say, why did I start watching this movie ?! However, the name "Hotel of Psychopaths" ... as they say, he knew what he was doing. It's just awful. By the way, in terms of genre, it is classified as a horror, although this is just a film about a mentally ill man who invited different people to the hotel supposedly for an unforgettable weekend, and he himself "started his game."
Shooting is disgusting, low-budget. Acting, well, it's not even a student circle. Funny, among them was Eric Roberts, brother you know who. But the game ... Dialogues from the category: (everyone is half-bloody on the floor) "- Honey, are you pregnant? - Yes, honey. - This is great, I hope after my death, you will be happy again. - Honey, do not die! - No , I'm dying!" Well, at the end, the theatrical closing of the eyes, apparently means that his hero has outlived his life. In general, there is no sanity in the film at all. You look at the behavior of the heroes and it seems that there are absolutely no normal ones there. Perhaps because of a similar acting. And what are the silent grimaces and strained reduction of eyebrows and wrinkles on the forehead of a former psychotherapist.
We watched and it all seemed, well, it couldn't be that there was such a movie. And the director could. And the actors did it.

Результат:

```
Enter the name of the file to read: usr_text/negative1.txt
Negative feedback
```

Тест 4

Исходный текст:

The zombie theme will probably appear on the screen for a long time. One director found a good solution and now, from time to time, others try to repeat it. Sometimes it turns out pretty well, like in a Korean movie about a train, or in an entire franchise about the sinister Umbrella Corporation. But more often, all zombie films show a cheap craft, which is a pity for the time spent.
This movie belongs to the latter category. Assembled on the knee, by some amateurs, the first big big hands of the camera. The plot is a collection of stupidity and misunderstandings. This time, they decided to blame the aliens for the invasion of the living dead. They flew in to seize the Earth, and as a weapon they used the revival of corpses. Apparently, they also used the "stupefying ray", because even zombies behaved smarter than living people.

The actors' play can be compared to the playing of wooden puppets. The same mediocre, mechanical movements, with minimal emotions barely reflected on the faces. Dialogue with some absurd stupidity and often did not correspond to the moment. For example, an incomprehensible and uterine sound is heard, and then a woman's scream is heard. The heroine declares to her friend: “This is how death itself sounds! That's exactly what it was, but these sounds didn't sound like death at all. It looked more like someone had overeated and was now having digestive problems. Although it was not the most stupid dialogue. They met much worse.
The makeup is disgusting. I don’t remember anywhere else there were such mediocre painted actors. Special effects, if you can call them that proud word, are at the level of eight-bit game consoles. If it's even worse. The shots were simply painted over the top of the frame to simulate a formidable view from the face of the alien invaders, they put a red filter on the camera. Well, and it is absolutely ingenious to place a burning Bengal candle on top of the lens. This is for sparks to fall. Probably this is how they demonstrated a malfunction in the ship's wiring in unearthly hicks.
Let me summarize. The film is very weak. Even though it lasts only seventy minutes, even that time is a pity to waste on it. I rate "Dead Men Again" as one with a minus and categorically do not recommend viewing it.

Результат:

```
Enter the name of the file to read: usr_text/negative2.txt
Negative feedback
```

Выводы

В результате выполнения лабораторной работы был реализован механизм прогнозирования успеха фильмов по отзывам из базы данных IMDb. Была построена модель, решающая задачу бинарной классификации.

Было исследовано влияние размера вектора представления текста на точность работы сети. Были построены графики точности и потерь сети для созданной модели. Получилось достигнуть достаточно высокого значения точности сети - 89%.

Была написана и протестирована программа, реализующая ввод пользовательского текста отзыва о фильме. В результате тестирования на пользовательских данных ИНС дала правильный прогноз в 4 из 4 случаев.