

УДК 004.9

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТИ ДЛЯ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА С ЦЕЛЮ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ РАБОТЫ КОМПАНИИ - ПРИМЕНЕНИЕ КЛАССИФИКАЦИИ ЭМОЦИЙ НА ОСНОВЕ МОДЕЛИ BERT В АНАЛИЗЕ ОТЗЫВОВ О ГОСТИНИЦАХ

Забоев Михаил Валерьевич

Доцент Кафедры информационных систем в экономике
Санкт-Петербургский государственный университет
m.zaboev@spbu.ru

Ван Хуэй

Степень магистра
Санкт-Петербургский государственный университет
1165318105@qq.com

Го Хао

Степень магистра
Санкт-Петербургский государственный университет
guohao6200@gmail.com

Аннотация

Данная работа посвящена применению модели глубокого обучения BERT для анализа эмоций в отзывах о гостиницах. Модель позволяет выявлять эмоциональные предпочтения пользователей, улучшать качество обслуживания и предоставлять точные данные для персонализированных рекомендаций. Несмотря на значительные успехи BERT, остаются вызовы, связанные с распознаванием сложных эмоций и качеством аннотирования данных. Исследование демонстрирует, как эффективно использовать BERT для анализа эмоций, включая классификацию отзывов по эмоциональной окраске и присвоение соответствующих оценок. Результаты исследования показывают, что BERT может быть полезен для анализа эмоционального настроения клиентов в реальном времени, что открывает новые возможности для оптимизации услуг и улучшения клиентского опыта.

Ключевые слова: анализ эмоций; отзывы о гостиницах; BERT; глубокое обучение

USING NEURAL NETWORKS FOR NATURAL LANGUAGE PROCESSING TO IMPROVE COMPANY PERFORMANCE--APPLYING EMOTION CLASSIFICATION BASED ON THE BERT MODEL IN ANALYZING HOTEL REVIEWS

Zaboev Mikhail Valerievich,

Associate Professor of the Department of Information Systems in Economics

Saint Petersburg State University
m.zaboev@spbu.ru

Wang Hui,

Master's degree
Saint Petersburg State University
1165318105@qq.com

Guo Hao,

Master's degree
Saint Petersburg State University
guohao6200@gmail.com

ABSTRACT

This work focuses on the application of the BERT deep learning model for sentiment analysis in hotel reviews. The model enables the identification of users' emotional preferences, improvement of service quality, and provision of accurate data for personalized recommendations. Despite the significant success of BERT, challenges remain, such as recognizing complex emotions and ensuring the quality of data annotation. This study demonstrates how to effectively use BERT for sentiment analysis, including classifying reviews by emotional tone and assigning appropriate scores. The results show that BERT can be highly beneficial for real-time analysis of customer sentiment, opening up new opportunities for service optimization and enhancing customer experience.

Keywords: sentiment analysis; hotel reviews; BERT; deep learning

Эмоциональное распознавание играет ключевую роль в анализе социальных медиа, позволяя быстро и точно выявлять эмоциональные настроения пользователей и интерпретировать их [1][3]. Это способствует более глубокому пониманию эмоциональной динамики, улучшению качества решений и эффективности работы в таких областях, как исследование данных и обучение. Однако анализ эмоций сталкивается с вызовами, связанными с разнообразием выражений и сложностью контента, что требует совершенствования алгоритмов и моделей [2].

В данной работе основное внимание уделено использованию модели BERT, основанной на архитектуре Transformer[6], для создания высокоэффективной и точной модели анализа эмоций. Цель исследования – удовлетворение потребностей в глубоком анализе данных и повышение качества обслуживания и удовлетворенности пользователей в бизнес-сценариях [7].

Нейронные сети – это алгоритмы искусственного интеллекта, имитирующие структуру и работу нейронов человеческого мозга. Они состоят из входного, скрытых и выходного слоев, где нейроны связаны весами, регулирующими передачу информации [7]. Во время обучения сеть проходит этапы прямого распространения для получения предсказаний и обратного распространения для корректировки параметров, что позволяет постепенно улучшать точность модели. Такой подход делает нейронные сети важным инструментом в задачах обработки данных и распознавания изображений.

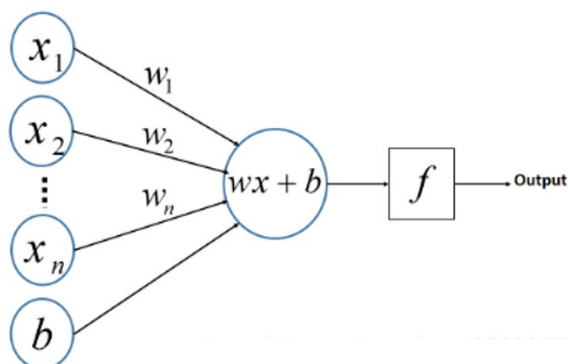


Рис. 2.1. Структура нейронной сети

Принцип работы нейронных сетей заключается в том, что входные данные последовательно проходят через несколько уровней преобразований и нелинейных функций активации, в результате чего получается предсказание. На каждом уровне нейроны обрабатывают выходные данные предыдущего уровня с использованием линейных преобразований и функций активации, передавая их на следующий уровень.

Этот пошаговый процесс позволяет нейронным сетям эффективно обрабатывать сложные нелинейные зависимости в данных.

$$u = \sum_{i=1}^n w_i x_i + b \tag{2.1}$$

Здесь ω обозначает матрицу признаков, а b представляет свободный член.

$$y = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases} \tag{2.2}$$

Алгоритм обратного распространения ошибки широко используется для обучения нейронных сетей, минимизируя разницу между предсказанным и истинным значением путем корректировки весов. Процесс включает инициализацию весов, прямое распространение для вычисления выхода, расчет ошибки и её обратное распространение. В ходе обратного распространения рассчитываются градиенты, обновляются веса и ошибка передается на предыдущие слои. Эти шаги повторяются до минимизации ошибки. Алгоритм улучшает точность предсказаний, используя цепное правило и методы, такие как градиентный спуск, что делает его ключевым инструментом в обучении нейронных сетей.

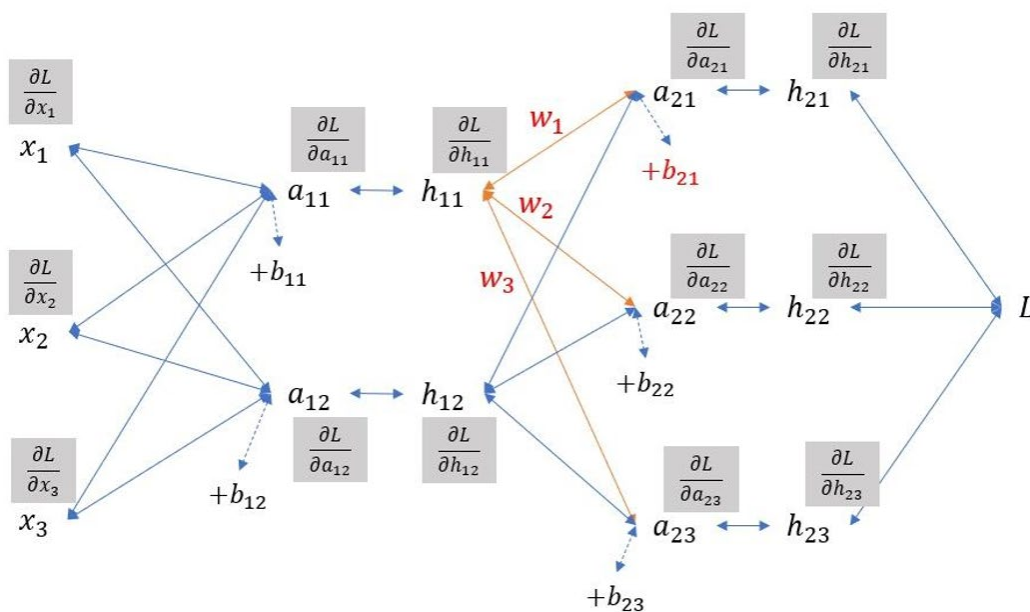


Рис. 2.2. Процесс обратного распространения ошибки

Набор данных Trip Advisor Hotel Reviews, доступный на платформе Kaggle, представляет собой крупный текстовый датасет, предоставленный TripAdvisor[3][4]. Он содержит большое количество отзывов о гостиницах, что делает его ценным ресурсом для исследований и приложений в таких областях, как обработка естественного языка (NLP), текстовая аналитика и анализ эмоций.

Набор данных включает информацию о гостиницах из многих стран мира. Каждый отзыв содержит оценку пользователя гостиницы (обычно в диапазоне от 1 до 5 звезд), текстовый отзыв, а также сведения о профиле пользователя.

Таблица 3.1. Пример содержания набора данных

Отзыв (предложение)	Метка
great hotel location union square stay great 6 people total 3 rooms, room little different, enjoyed location close union square shopping just cable car away wharf, staff friendly helpful, definately stay returning sf,	3

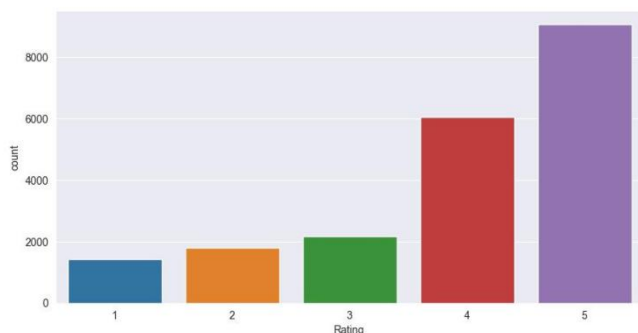


Рис. 3.1. Гистограмма распределения оценок в отзывах о гостиницах

Рисунок 3.1 демонстрирует гистограмму распределения оценок в отзывах. Из рисунка видно, что большинство отзывов имеют высокие оценки, тогда как низкие оценки встречаются реже, что свидетельствует о хорошем уровне репутации гостиницы.

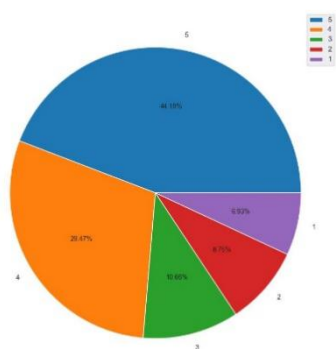


Рис. 3.2. Круговая диаграмма распределения оценок в отзывах о гостиницах

Рисунок 3.2 показывает круговую диаграмму распределения оценок. Из рисунка видно, что высокие оценки составляют значительную часть — 73,66% от общего количества отзывов.

В наборе данных содержится всего 20 491 запись, которые были разделены на обучающую и тестовую выборки. Обучающая выборка включает 16 392 записи и

используется для обучения модели, а тестовая выборка состоит из 4 099 записей и предназначена для оценки её производительности.

В бинарной классификации экземпляры делятся на положительный (positive) и отрицательный (negative) классы, и возможны четыре варианта классификации: истинно положительные (TP) – правильно предсказанные положительные; ложно отрицательные (FN) – положительные, ошибочно предсказанные как отрицательные; ложно положительные (FP) – отрицательные, ошибочно предсказанные как положительные; и истинно отрицательные (TN) – правильно предсказанные отрицательные.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Рис. 3.3 Матрица ошибок

ROC-кривая (Receiver Operating Characteristic) – это метрика для оценки классификационных моделей, отображающая зависимость между истинно положительной долей (TPR) и ложноположительной долей (FPR) при различных порогах классификации. TPR показывает долю правильно классифицированных положительных примеров, а FPR – долю отрицательных примеров, ошибочно классифицированных как положительные. Для количественной оценки модели используется площадь под кривой (AUC): значения ближе к 1 свидетельствуют о высокой производительности модели, а значение 0.5 указывает на её равенство случайному угадыванию.

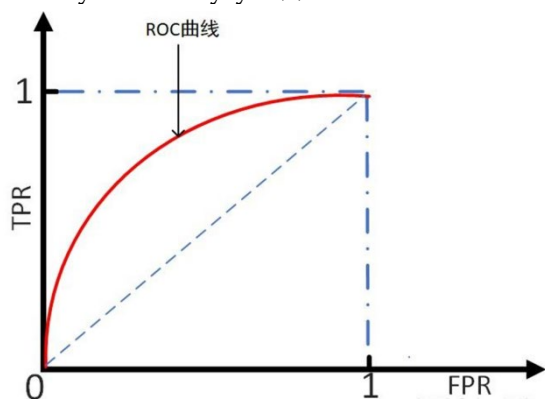


Рис. 3.4. ROC-кривая

PR-кривая (Precision-Recall curve) используется для оценки производительности классификационных моделей, показывая взаимосвязь между точностью (precision) и полнотой (recall) при разных порогах классификации. Каждая точка кривой соответствует конкретному порогу, что позволяет выбирать модель в зависимости от приоритетов: высокой точности при большом пороге или высокой полноты при меньшем пороге. PR-кривая особенно полезна для анализа моделей на несбалансированных данных.

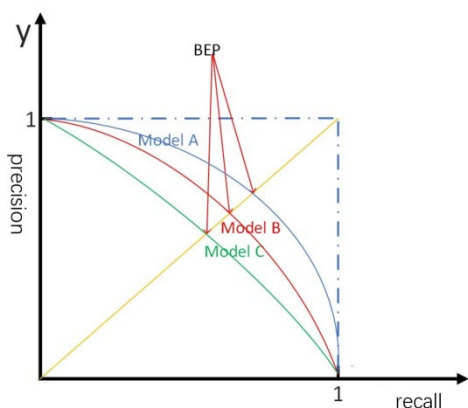


Рис. 3.5. PR-кривая

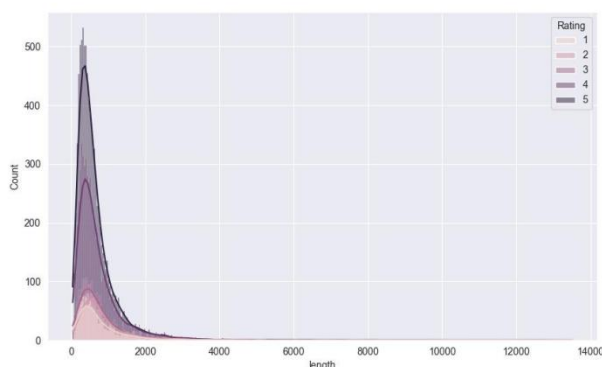


Рис. 3.6. Гистограмма распределения длины отзывов с разными оценками

На рис. 3.6 показана гистограмма распределения длины отзывов в зависимости от оценки. Как видно, отзывы с высокими оценками в среднем короче, чем отзывы с низкими оценками.

Аппаратная среда: CPU: Intel Core i7-12700H, RAM: 32GB, GPU: RTX 3060.

Программная среда: Windows 11, PyCharm 2021, Anaconda 3, Python 3.9.7, Pytorch 1.8.0, torchvision 0.9.0, cuda 11.7, scikit-learn 1.0.2, transformers 4.28.0, PyQt5 5.15.9

Таблица 4.1 Параметры обучения модели BERT

номер	Имя параметра	Значение параметра
1	Количество эпох (Epoch)	15
2	Размер пакета (Batch size)	16
3	Моментум (momentum)	0.9
4	Скорость обучения (Learning rate, Lr)	0.001
5	Стратегия изменения скорости обучения	Линейное убывание
6	Оптимизатор	Adam optimizer
7	Функция потерь	Cross Entropy Loss

Модель Transformer использует алгоритм многоголового внимания для эффективного извлечения семантических признаков, поэтому в данном эксперименте BERT применяется в качестве модуля извлечения семантической информации для построения модели классификации эмоций [6][7].

Таблица 4.2. Показатели модели BERT

Метрика	Accuracy	Precision	Recall	F1-score
---------	----------	-----------	--------	----------

BERT	78.80	78.80	78.80	78.80
------	-------	-------	-------	-------

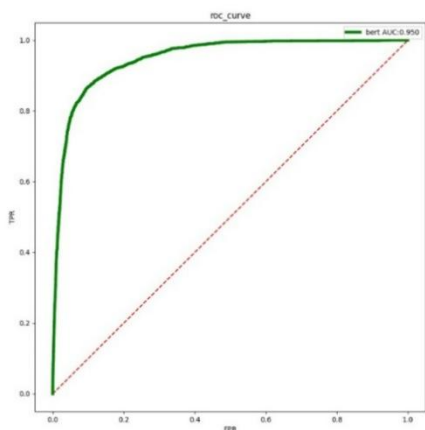


Рис. 4.1. ROC-кривая для модели BERT

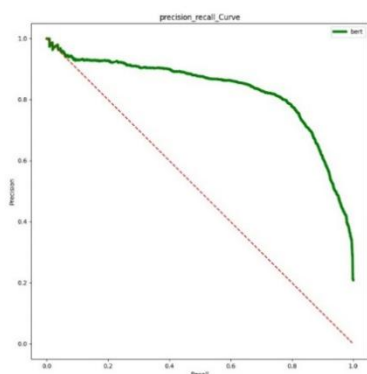


Рис. 4.2. PR-кривая для модели BERT

В таблице 4.2 приведены показатели модели BERT. На рисунках 4.1–4.3 представлены ROC-кривая, PR-кривая и матрица ошибок для тестовой выборки. Результаты показывают, что точность модели BERT составляет 83.50%, а значение AUC – 93.30, что подтверждает высокую эффективность данной модели.

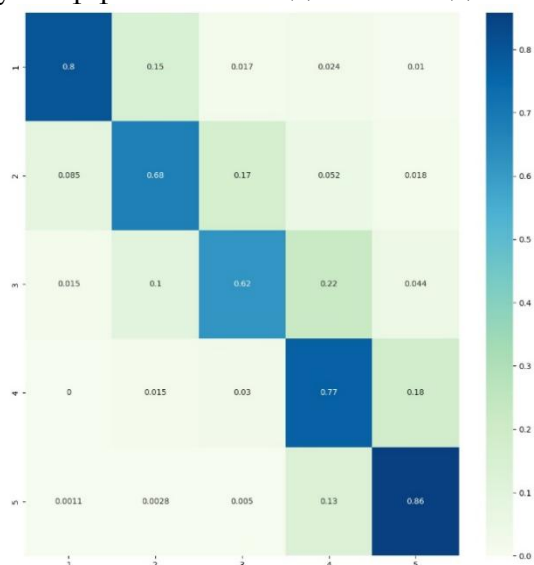
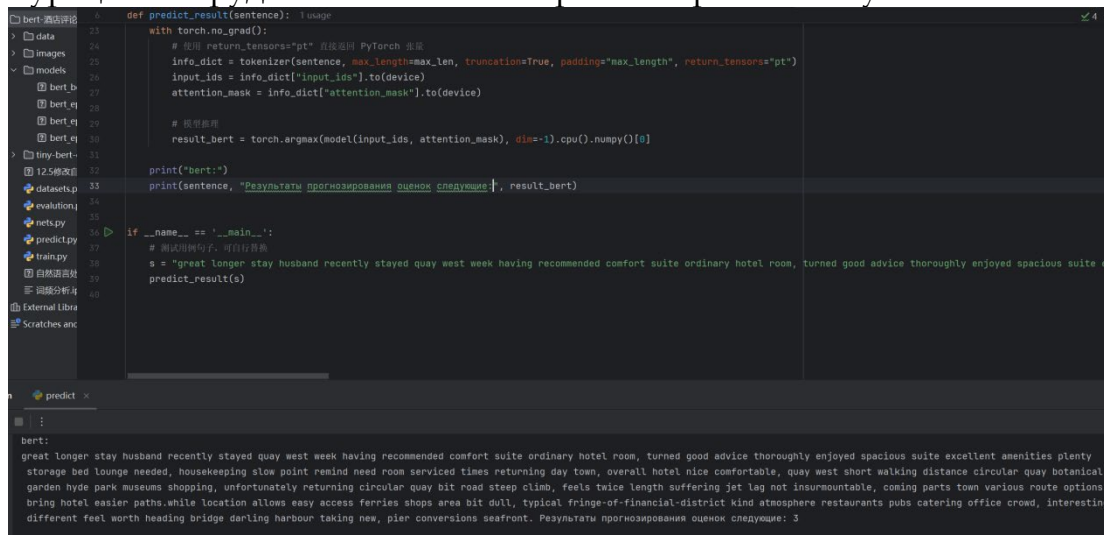


Рис. 4.3. Матрица ошибок модели BERT

BERT – это предобученная модель представления языка на основе архитектуры Transformer[8]. Она способна изучать универсальные языковые представления и демонстрирует высокие результаты в различных задачах обработки естественного языка.

В задачах классификации эмоций BERT обычно достигает передовых результатов благодаря способности захватывать контекстную информацию и благодаря использованию предобученных весов[9], что позволяет обойтись без дополнительной настройки для конкретной задачи.

Тем не менее, крупные размеры модели BERT требуют значительных вычислительных и хранилищных ресурсов, что становится важным фактором при её применении. Для её эффективного использования важно обеспечить соответствующую конфигурацию оборудования и оптимизировать стратегию обучения.



```

def predict_result(sentence):
    with torch.no_grad():
        # 使用 return_tensors="pt" 直接返回 PyTorch 张量
        info_dict = tokenizer(sentence, max_length=max_len, truncation=True, padding="max_length", return_tensors="pt")
        input_ids = info_dict["input_ids"].to(device)
        attention_mask = info_dict["attention_mask"].to(device)

        # 模型推理
        result_bert = torch.argmax(model(input_ids, attention_mask), dim=-1).cpu().numpy()[0]

    print("bert:")
    print(sentence, "Результаты прогнозирования оценок следующие:", result_bert)

if __name__ == '__main__':
    # 测试用例句子, 可自行替换
    s = "great longer stay husband recently stayed quay west week having recommended comfort suite ordinary hotel room, turned good advice thoroughly enjoyed spacious suite excellent amenities plenty storage bed lounge needed, housekeeping slow point remind need room serviced times returning day town, overall hotel nice comfortable, quay west short walking distance circular quay botanical garden Hyde Park museums shopping, unfortunately returning circular quay bit road steep climb, feels twice length suffering jet lag not insurmountable, coming parts town various route options bring hotel easier paths.while location allows easy access ferries shops area bit dull, typical fringe-of-financial-district kind atmosphere restaurants pubs catering office crowd, interesting different feel worth heading bridge darling harbour taking new, pier conversions seafront. Результаты прогнозирования оценок следующие: 3
  
```

Рис 4.4 Пример применения модели BERT

Модель BERT, обученная на корпусе отзывов о гостиницах, позволяет точно классифицировать эмоции, определяя их положительную или отрицательную окраску с учетом контекста. Она помогает компаниям в реальном времени анализировать клиентские отзывы, выявлять удовлетворенность, жалобы и рекомендации. Основные преимущества BERT — это её предварительное обучение и глубокое понимание контекста, что делает её предпочтительным выбором для задач с эмоциональными нюансами. Однако модель требует значительных вычислительных ресурсов и оптимизации стратегии обучения. BERT предоставляет компаниям детальный анализ отзывов, помогая понять потребности клиентов, улучшить качество услуг и повысить лояльность.

Список литературы:

1. Zhou X, Wang Y and Ma P (2022) Survey on the tourist satisfaction of rural homestay based on the theory of importance–performance analysis. *Front. Psychol.* 13:940345. doi: 10.3389/fpsyg.2022.940345 2022 № 13. С. 01-11
2. Liu, H. C., Jeng, B. C., Mai, Y. T., et al. (2014). Design of online survey system with an advanced IPA discrimination index for customer satisfaction assessment. *Electronic Commerce Research*, 14(3), 223-243. 2024 № 14 С. 12-20.
3. Li, D., et al. (2024). Sentiment analysis method for online reviews based on deep learning. *Modeling and Simulation*, 13(5), 5375-5380. 2024 № 13 С. 1-10.
4. CN113704459A - An online text sentiment analysis method based on neural networks. 2020 С. № 1.1-12.
5. Sentiment analysis of Twitter text based on convolutional neural networks - Data collection and processing. 2018 № 33 С. 921-927.
6. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. 2018 № 2 С. 01-16.

7. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 № 1 С. 01-15.
8. Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. 2019 № 1 С. 01-13.
9. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 649-657. 2015 № 3 С. 01-09.