

## **РАЗРАБОТКА И ИССЛЕДОВАНИЕ КЛАССИФИКАТОРА ВОПРОСОВ ЧАТ-БОТУ В ИГРЕ ДЛЯ АБИТУРИЕНТОВ**

**Аннотация.** В работе рассматривается реализация классификаторов вопросов чат-боту в игре для абитуриентов для решения проблемы низкой вовлеченности игроков. На потенциальных вопросах, размеченных ответами, была обучена глубокая нейронная сеть. С помощью предобработки и аугментации данных удалось достичь точности классификации, равной 92%. Полученные классификаторы могут работать с короткими вопросами в рамках контекста диалога.

**Ключевые слова:** чат-бот, аугментация текстов, предобработка текстов, классификация текстов, методы глубокого обучения, рекуррентная нейронная сеть.

**Введение.** В современных программных продуктах чат-боты являются одним из решений для автоматизации задач и широко применяются в различных сферах, включая образование. Они информируют о курсах, занятиях и расписании, а также помогают абитуриентам, предоставляя необходимую информацию об учебном заведении. Благодаря кроссплатформенности бота можно интегрировать на различные устройства или в собственные информационные системы, что делает их востребованным инструментом для обеспечения доступа к информации и поддержки в образовательной среде.

Чат-боты по функционалу делятся на четыре категории: разговорные, ассистенты, функциональные и q&a (вопрос-ответ), которые подробно рассмотрены в работе О. Трофименко и др. [1]. Разговорные боты имитируют человеческое общение, не имея конкретной цели в разговоре, их задача поддерживать несложный диалог с пользователем. Ассистенты собирают сведения о клиенте, например, данные о заказе, таким образом, заменяя операторов. Боты функционального типа автоматизируют выполнение определенных задач. Q&A боты предназначены для ответов на вопросы пользователя, заменяя привычный поиск информации.

На текущий момент в основе работы многих диалоговых ботов, относящихся к типу “q&a”, лежит механизм, способный интерпретировать вопросы пользователя и подбирать подходящие ответы, что делает их часто используемыми в сфере образования. Такие боты могут служить средством повышения интерактивности [2] или выступать в роли консультанта [3].

В настоящей работе рассматривается реализация игры в жанре визуальной новеллы [4] для абитуриентов, где пользователь выступает в роли абитуриента, сопровождаемого персонажем, который, следуя сценарию, рассказывает о вузе через диалоги с заготовленными репликами игрока. Однако такая игровая механика предопределяет диалог пользователя с персонажем. Поэтому возникла потребность в создании возможности для пользователя самостоятельно задавать вопросы.

Соответственно целью работы является создание серии классификаторов, которые смогут соотносить вопросы с сюжетными линиями. Предположительно, данные классификаторы позволят с достаточной точностью определять подходящие ответы на короткие вопросы (до 20 слов) в рамках контекста диалога. Задачи исследования: определить необходимый набор обучающих данных, провести анализ частоты появления одинаковых слов в разных классах, при которой их можно исключить из фраз без потери точности, выбрать архитектуру классификатора для достижения достаточной точности.

**Проблема исследования.** Построение и анализ модели классификации направлены на решение задачи, которую формально можно представить следующим образом:

Пусть

$X = x_1, x_2 \dots x_m$  — фразы игроков,

$L = l_1, l_2 \dots l_n$  — метки классов,

$X^{labeled} = x_{i_1}, x_{i_2} \dots x_{i_k}$  — размеченные потенциальные фразы игроков,  $X^{labeled} \subset X$ .

Необходимо с учетом закономерностей в размеченных фразах  $X^{labeled}$  построить такую функцию  $F$ , которая для любой фразы игрока из множества  $X$  определяла бы соответствующую метку класса из множества  $L$ :

$$F: X \rightarrow L.$$

## Материалы и методы

### *Описание данных*

Для обучения классификаторов был получен набор размеченных потенциальных фраз игроков из текстовых описаний диалогов между персонажем и игроком, состоящих из фраз на русском языке, длиной до 20 слов. Тексты были структурированы, в результате чего выделены 62 сюжетные точки. Для каждой точки обучается один классификатор на 2-6 классах (всего 151 класс), при этом для некоторых классов потребовалось обучить несколько моделей. В корпусе потенциальных фраз представлено от 60 до 70 экземпляров каждого класса, что оказалось недостаточным для достижения необходимой точности, поэтому набор был расширен с помощью аугментации.

### *Аугментация данных*

В наборе представлены различные формулировки одних и тех же запросов, однако многие содержат уникальные токены и отличительную структуру, что не позволяет классификатору обнаружить некоторые зависимости. Для решения обозначенной проблемы осуществлена аугментация с помощью модели `rut5-base-paraphraser` [5]. Данная модель позволяет перефразировать текст таким образом, чтобы сохранить наиболее значимые  $n$ -граммы этого текста. Модель принимает на вход исходный текст и длину учитываемых  $n$ -грамм. В зависимости от значения  $n$  возможно сохранение структуры или учет токенов в похожих структурах. В результатах работы представлен сравнительный анализ качества классификации от объема обучающих данных.

### *Предобработка данных*

Перед обучением классификатора тексты вопросов с помощью библиотеки `spacy` (<https://spacy.io/usage/models>) были приведены к нижнему регистру и очищены от знаков пунктуации, значимые слова лемматизированы, а стоп-слова исключены. В данном случае под стоп-словами помимо предлогов подразумеваются слова, для которых соизмерима частота появления в вопросах, относящихся к разным классам. Для всех сюжетных точек были сформированы словари, где ключ — это лемма, а значение — массив с количествами ее появления в текстах каждого класса. За соотношение частот взято отношение двух наибольших значений. Соотношение, при котором слово можно отнести к стоп-словам, определено в результатах работы.

### *Подходы к классификации текстов*

Существует множество подходов к решению задачи классификации текстов. В работе В. А. Яцко [6] представлен способ вычисления классификационных индексов текстов на

основе анализа распределения стоп-слов. Для небольших наборов обучающих данных предложена комбинация механизма сжатия текстов и алгоритма k-ближайших соседей [7]. В настоящее время все активнее используются рекуррентные нейронные сети [8]. На небольших текстах они часто показывают более высокую точность, так как учитывают контекст. Вопросы в имеющемся наборе данных в среднем состоят из 10 слов, поэтому для реализации выбран данный подход.

#### *Архитектура классификатора*

Для реализации классификаторов были использованы методы глубокого обучения, в частности, рекуррентная нейронная сеть типа GRU [11] из библиотеки keras (<https://keras.io/api/>), предоставляющей инструменты для организации многослойной модели. При построении классификаторов был выполнен сравнительный анализ архитектур, в котором варьировались как сами слои, так и их параметры. Архитектура итоговой модели представлена на рис. 1. Результаты сравнительного анализа описаны в следующем разделе данной работы.

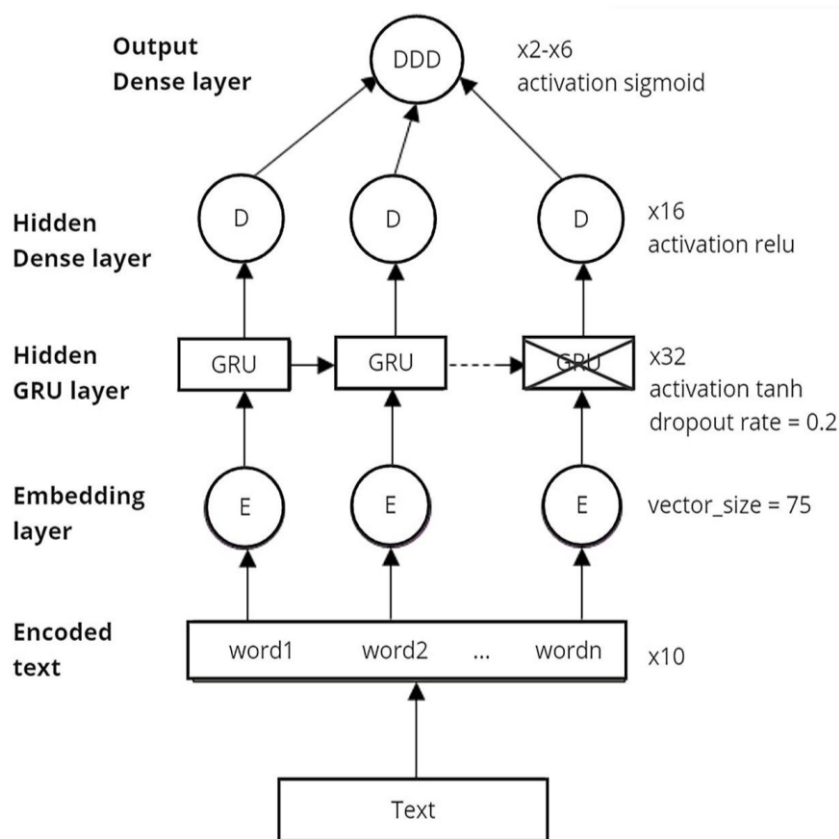


Рис. 1. Архитектура классификаторов

**Результаты.** В результате аугментации для каждого текста из корпуса потенциальных фраз были получены по два перефразы с длинами n-грамм 3 и 5, а также с длиной 4 для случайных фраз из малочисленных классов. В итоге получилось по 210 экземпляров для каждого класса. Из них 180 были взяты в обучающую выборку и 30 в тестовую.

График на рис. 2 показывает изменение точности классификации при увеличении числа экземпляров в одном классе. Достаточная средняя точность была достигнута уже при 140 экземплярах в классе. Однако ее удалось повысить до 0.914 при 180 экземплярах в классе.

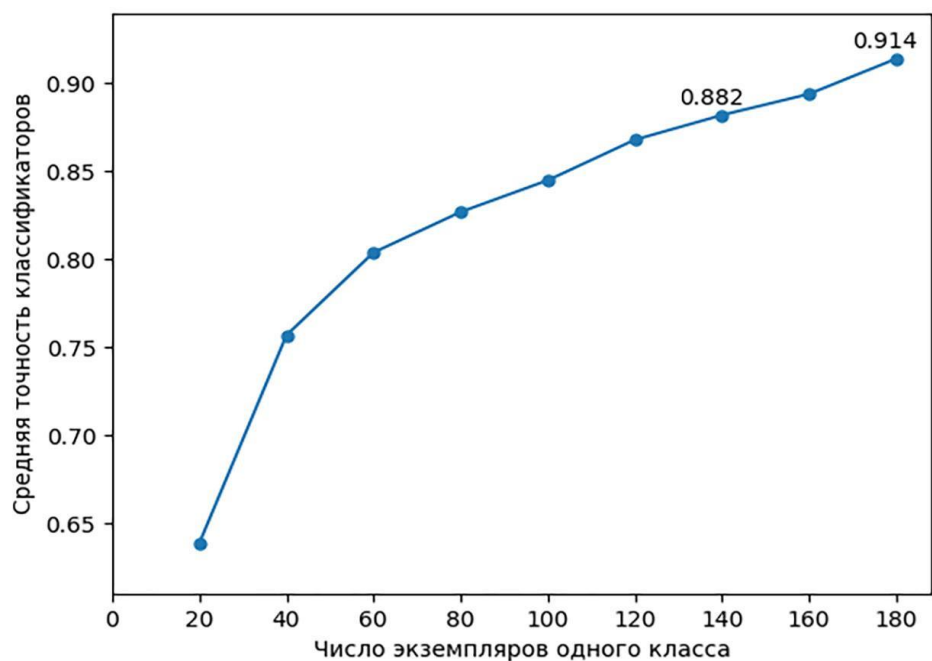


Рис. 2. Зависимость точности классификации от объема обучающих данных

Максимальная точность была достигнута после исключения стоп-слов, являющихся общими для разных классов, например, «расскажи» и «могут» из текстов «Расскажи, кем могут работать выпускники» и «Расскажи, где студенты могут пройти практику».

График на рис. 3 показывает, как меняется точность классификации в зависимости от порога соотношения частот, при котором слово рассматривается как стоп-слово. Было выявлено, что можно получить точность 0.921 при исключении слов с соотношением частот 0.85. В результате от 2 до 5 слов во фразах из каждого класса отнесены к стоп-словам.

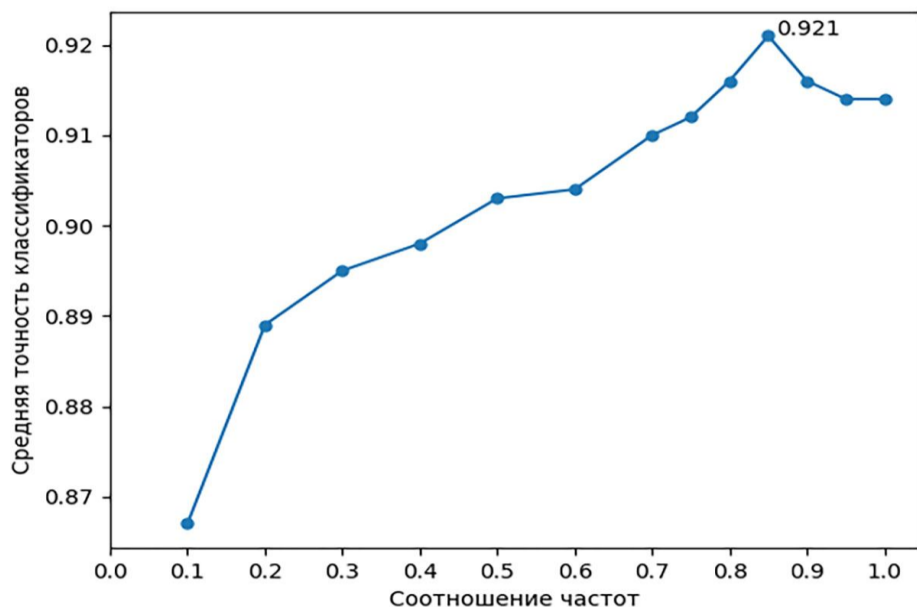


Рис. 3. Зависимость точности классификации от соотношения частот, при котором слово считается стоп-словом

Описанные исследования проводились с итоговыми классификаторами, однако на полном объеме обучающей выборки, очищенной от стоп-слов были рассмотрены и другие архитектуры, сравнительный анализ которых приведен в табл. 1.

Во всех вариациях общими являются Embedding слой и последний классифицирующий Dense слой. Из рекуррентных сетей наиболее точной (0.899) оказалась GRU с размером вектора эмбедингов 75. Добавление Dense слоя с 16 нейронами и слоя Dropout, отбрасывающего 20% единиц, позволило повысить точность до 0.921. Добавление еще одного Dense слоя и BatchNormalization слоя только снизило точность. Лучшую точность 0.921 показал оптимизатор AdamW при обучении в течение 30 эпох с батчем из 25% экземпляров обучающей выборки.

Таблица 1

Точности классификации при различных параметрах сети

Слой			Нейроны рекуррентного слоя			
			8	16	32	64
Embedding (100)	RNN	Dense	0.857	0.849	0.863	0.875
	LSTM		0.875	0.888	0.888	0.883
	GRU		0.875	0.891	<b>0.898</b>	<b>0.898</b>
Слой			Размер вектора эмбедингов			
			25	50	75	100
Embedding	GRU(32)	Dense	0.885	0.891	<b>0.899</b>	0.898
Слой			Нейроны Dense слоя			
			8	16	32	64
Embedding (75)	GRU(32) Dense	Dense	0.895	<b>0.914</b>	0.906	0.906
Слой			Нейроны Dense слоя			
			8	16	32	64
Embedding (75)	GRU(32) Dense(16) Dense	Dense	0.877	0.892	0.897	0.902
Слой			Импульс скользящего среднего			
			0.1	0.4	0.7	0.99
Embedding (75)	GRU(32) Dense(16) Batch Normalization	Dense	0.894	0.886	0.905	0.883
Слой			Доля отбрасываемых единиц			
			0.15	0.2	0.25	0.3
Embedding (75)	GRU(32) Dropout Dense(16)	Dense	0.917	<b>0.921</b>	0.884	0.901
			Оптимизатор			
			Adam	Nadam	AdamW	RMSprop
			0.916	0.912	<b>0.921</b>	0.914
			Число эпох			
			10	20	30	40
			0.828	0.885	<b>0.921</b>	0.913
			Размер батча			
			15%	20%	25%	30%
			0.914	0.916	<b>0.921</b>	0.911

**Обсуждение.** Несмотря на достижение достаточной средней точности, минимальная для отдельного классификатора получилась 0.84, это связано с близостью тематики классов. Можно попытаться увеличить точность в таких ситуациях, дополнив набор данных фразами, в которых сделан больший акцент на их различия. Также для повышения точности можно попробовать изменить логику определения стоп-слов, чтобы помимо соотношения частот учитывать количества появления слов во фразах.

**Заключение.** По итогу данной работы была разработана серия классификаторов с архитектурой глубокой нейронной сети, позволяющих в рамках контекста диалога со средней точностью 92.1% на тестовых выборках соотносить краткие вопросы чат-боту с подходящими сюжетными линиями. Данная точность является достаточной для данной задачи. Она была достигнута при 180 экземплярах каждого класса в наборе обучающих данных и при исключении стоп-слов с порогом соотношения частот их появления в разных классах более 0.85. Дальнейшие исследования могут быть связаны с совершенствованием логики определения стоп-слов и учетом ключевых различиях между классами при аугментации.

## СПИСОК ЛИТЕРАТУРЫ

1. Trofymenko O. Taxonomy of Chatbots / O. Trofymenko, Y. Prokop, N. Loginova, A. Zadereyko — Текст: электронный // 2nd International Conference On Intellectual Systems And Information Technologies, Isit 2021. — Odessa: National University “Odessa Law Academy”, 2021. — С. 165-169. — URL: <https://ceur-ws.org/Vol-3126/paper24.pdf> (дата обращения: 01.04.2024).
2. Веряев А.А. Чат-боты как средство повышения интерактивности учебных занятий / А.А. Веряев, Ю.Э. Лозыченко — Текст: электронный // Информационно-коммуникационные технологии в педагогическом образовании. — 2022. — № 5 (80). — С. 10-18. — URL: <https://infed.ru/articles/1306/> (дата обращения: 01.04.2024).
3. Суханова Н.Т. Использование чат-ботов для автоматизации предоставления справочной информации абитуриентам и студентам вузов / Н.Т. Суханова, Т.М. Вежелис. — Текст: электронный // Проблемы современного педагогического образования. — 2022. — № 76-2. — С. 178-181. — URL: <https://www.elibrary.ru/item.asp?id=49809364> (дата обращения: 01.04.2024). — Режим доступа: Научная электронная библиотека eLIBRARY.RU.
4. Логинова И.А. Визуальная новелла как новое представление литературы / И.А. Логинова, Е.Д. Волегова — Текст: электронный // Инновации в профессиональном и профессионально-педагогическом образовании. — 2023. — С. 94-96. — URL: <https://www.elibrary.ru/item.asp?edn= aqekzi> (дата обращения: 01.04.2024). — Режим доступа: Научная электронная библиотека eLIBRARY.RU.
5. Cointegrated/rut5-base-paraphraser [сайт]. — URL: <https://huggingface.co/cointegrated/rut5-base-paraphraser> (дата обращения: 01.04.2024) — Текст: электронный.
6. Яцко В.А. Y-метод классификации текстов / В.А. Яцко — Текст: электронный // Грани познания. — 2021. — № 3. — С. 52-56. URL: <https://www.elibrary.ru/item.asp?edn=tppwbe> (дата обращения: 01.04.2024).
7. Jiang Z. “Low-Resource” Text Classification: A Parameter-Free Classification Method with Compressors / Z. Jiang, M. Y.R. Yang, M. Tsirlin, R. Tang, Y. Dai, J. Lin — Текст: электронный // Findings of the Association for Computational Linguistics: ACL 2023. — 2023. — С. 6810-6828. URL: <https://aclanthology.org/2023.findings-acl.426.pdf> (дата обращения: 01.04.2024).
8. Jie Du. Novel efficient RNN and LSTM-like architectures: Recurrent and gated broad learning systems and their applications for text classification / Jie Du, Chi-Man Vong, C. L. Philip Chen // IEEE Trans. Cybern. 51, 3 (2021). — С. 1586-1597. URL: <http://dx.doi.org/10.1109/TCYB.2020.2969705> (дата обращения: 01.04.2024).
9. Chung J. Empirical evaluation of gated recurrent neural networks on sequence modeling / J. Chung, C. Gulcehre, K. Cho, Y. Bengio — Текст: электронный // arXiv preprint. arXiv:1412.3555. — 2014. URL: <https://arxiv.org/pdf/1412.3555.pdf> (дата обращения: 01.04.2024).