МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южно-Уральский государственный университет (национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

Разработка модели автоматической оценки заявок некоммерческих организаций на финансирование от Правительства Челябинской области

Автор работы:
студент группы КЭ-229
А.А. Зелениченко

Рецензент:
доцент кафедры ИИТиМОИ
ФГБОУ ВО «ЮУрГГПУ», к.п.н.
Л.Н. Носова

Научный руководитель:
доцент кафедры СП, к.п.н.
О.Н. Иванова

Челябинск, 2024 г.

Актуальность

– Высокая социальная значимость деятельности Фонда поддержки гражданских инициатив Южного Урала

– Трудоемкость процесса проведения оценки с привлечением нескольких экспертов (800 заявок за год)

– Снижение времени обработки заявки

– Проверка корректности входных данных

– Сложность проверки значений атрибутов входных данных

– Необходимость проверок контекстуальных зависимостей

– Отсутствие аналогов проекта в государственных информационных системах

Цель и задачи

Цель:

Разработка модели автоматической оценки заявок некоммерческих организаций на финансирование от Правительства Челябинской области

Задачи:

1. Изучить предметную область, провести обзор научной литературы
2. Подготовить набор данных для обучения, спроектировать архитектуры нейронных сетей и модель оценки
3. Обучить нейронные сети, оценить результаты их работы
4. Реализовать программную систему для оценки заявок на основе нейросетевых технологий
5. Протестировать программную систему

Обзор аналогичных решений

Задача	Количество признаков	Размерность датасета	Векторизация	Точность
Бинарная классификация [12]	4	2903	Мешок слов	66%
Многоклассовая классификация [13]	6	10685	TF-IDF	87%
Бинарная классификация [10]	–	–	–	–
Классификация [15]	–	–	–	–

10. Сироткин А.В. Оценка актуальности заявки на проектное финансирование в автоматизированной системе распределения грантов. / А.В. Сироткин, В.К. Копченко // Современные наукоемкие технологии, 2021. – № 12–1. – С. 109–113.
12. Рогожин Д.К. Разработка и применение модели оценки заявок на предоставление грантов федеральным вузам и федеральным государственным учреждениям из бюджета города Москвы на основе методов машинного обучения для автоматической обработки текстов // Инжиниринг предприятий и управление знаниями: Сборник научных трудов XXV Российской научной конференции. – Москва: РЭУ имени Г.В. Плеханова, 2022. – С. 235–241.
13. Гусев П.Ю. Разработка системы классификации текстов по научным специальностям с применением методов машинного обучения. // Вестник НГУ, 2021. – № 1. – С. 39–47.
15. Introducing AI-Assisted Application Screening: Transform your grant review process with intelligent pre-screening. [Электронный ресурс] URL: https://www.smartsimple.com/blog/introducing-ai-assisted-application-screening (дата обращения: 10.02.2024 г.).

Декомпозиция задачи

Подзадача «Классификация заявок» – многоклассовая классификация текста:

- Количество классов: 12 классов атрибута «Направление» исходного датасета
- Исследуемые архитектуры: RNN, LSTM, GRU, BERT
- Атрибуты: «Краткое написание», «Цель», «Задачи»

Подзадача «Схожесть текстов» – семантическое сходство текста:

- Модель: paraphrase-multilingual-MiniLM-L12-v2
- Мера схожести: Cosine distance
- Атрибуты: «Социальная значимость», «Цель» , «Задачи»

Модель оценки заявки

Сбор данных

№	Наименование	Атрибут	Среднее кол-во токенов
1	Ссылка	Reference	–
2	Направление	Category	3
3	Название	Application Name	7
4	Краткое описание	Short Description Raw	256
5	Дата начала	Start Date	1
6	Дата окончания	Final Date	1
7	Социальная значимость	Social Description Raw	308
8	Целевые группы	Social Groups Raw	13
9	Цель	Aims Raw	25
10	Задачи	Tasks Raw	57
11	Качественные результаты	Quality Result	111
12	Оценка результатов	Evaluation	196

Источник данных:

оценка.гранты.рф [ссылка]

Характеристики датасета:

– формат: «.json»

– количество атрибутов: 12

– размерность: 25000

Предобработка данных

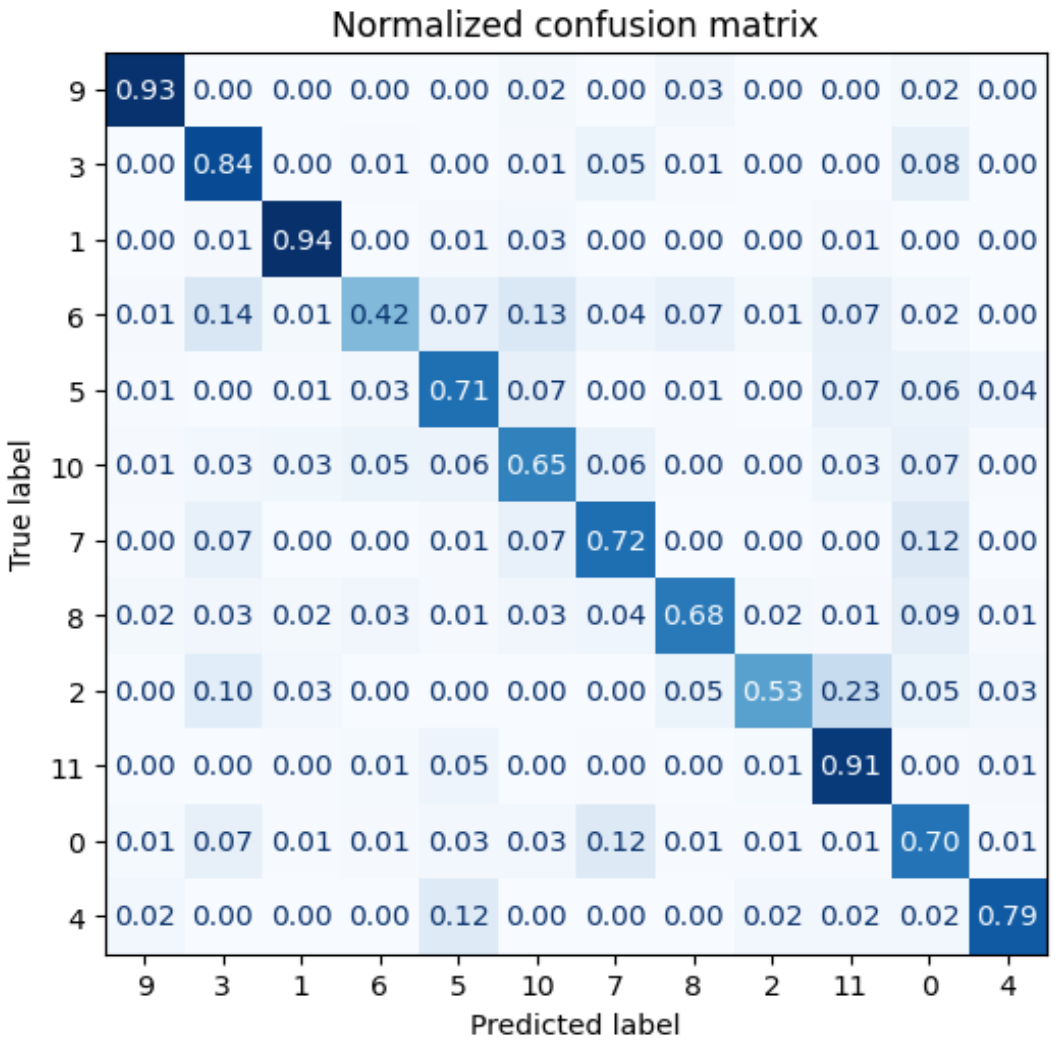
№	Метод	Реализация
1	Исключение экземпляров, которые содержат пустые значения	Функция df.dropna
2	Удаление html атрибутов	Пакет BeautifulSoup
3	Удаление классов, содержащих менее 150 экземпляров	Функция df.drop
4	Удаление числовых значений	Регулярное выражение
5	Перевод слов в нижний регистр	Функция lower
6	Удаление знаков препинания	Регулярное выражение
7	Удаление слов-дубликатов	Регулярное выражение
8	Удаление дублирующих пробелов	Регулярное выражение
9	Удаление стоп-слов	Пакет nltk.stopwords
10.1	Лемматизация	Пакет pymystem3 (время выполнения – 1722 сек)
10.2	Стемминг	Пакет pymorphy3 (время выполнения – 5543 сек)

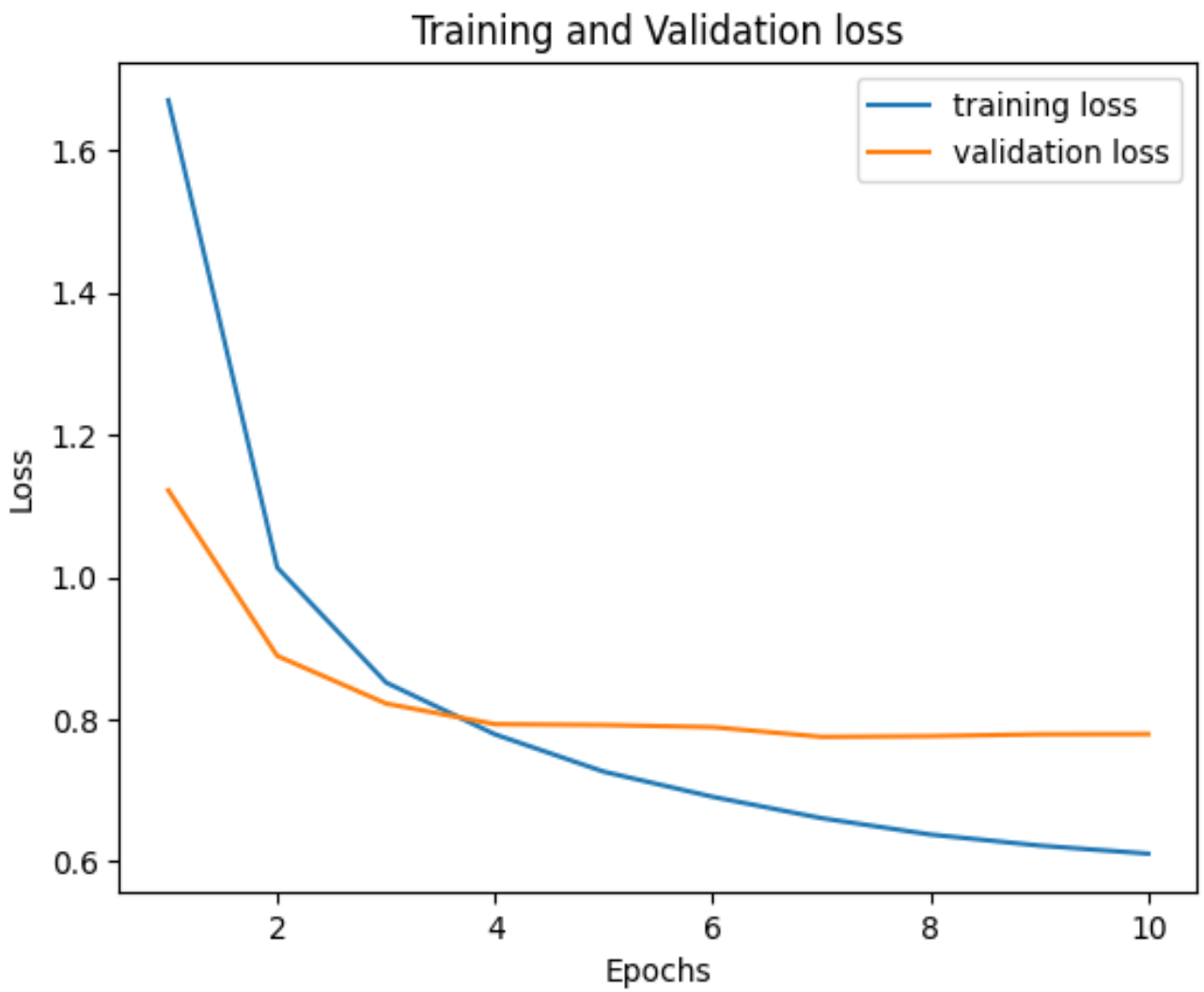
Классификация заявок: модели

№	Токенизатор	Архитектура	Эпохи	Batch size	Loss	Accuracy
1	TfidfVectorizer/ Word2Vec/ AutoTokenizer rubert-tiny2	Random forest	–	32	–	0,6716
2	Tensorflow keras.preprocessing.text	CNN	5	16	0,0166	0,7162
3	Tensorflow keras.preprocessing.text	LSTM	8	16	0,2872	0,6022
4	Tensorflow keras.preprocessing.text	GRU	8	16	0,3111	0,5808
5	**AutoTokenizer rubert-tiny2**	**BERT**	**3**	**8**	**0,7877**	**0,7458**

Разделение обучающей / тестовой / валидационной выборки: 0,7 / 0,2 / 0,1

Классификация заявок: результаты обучения





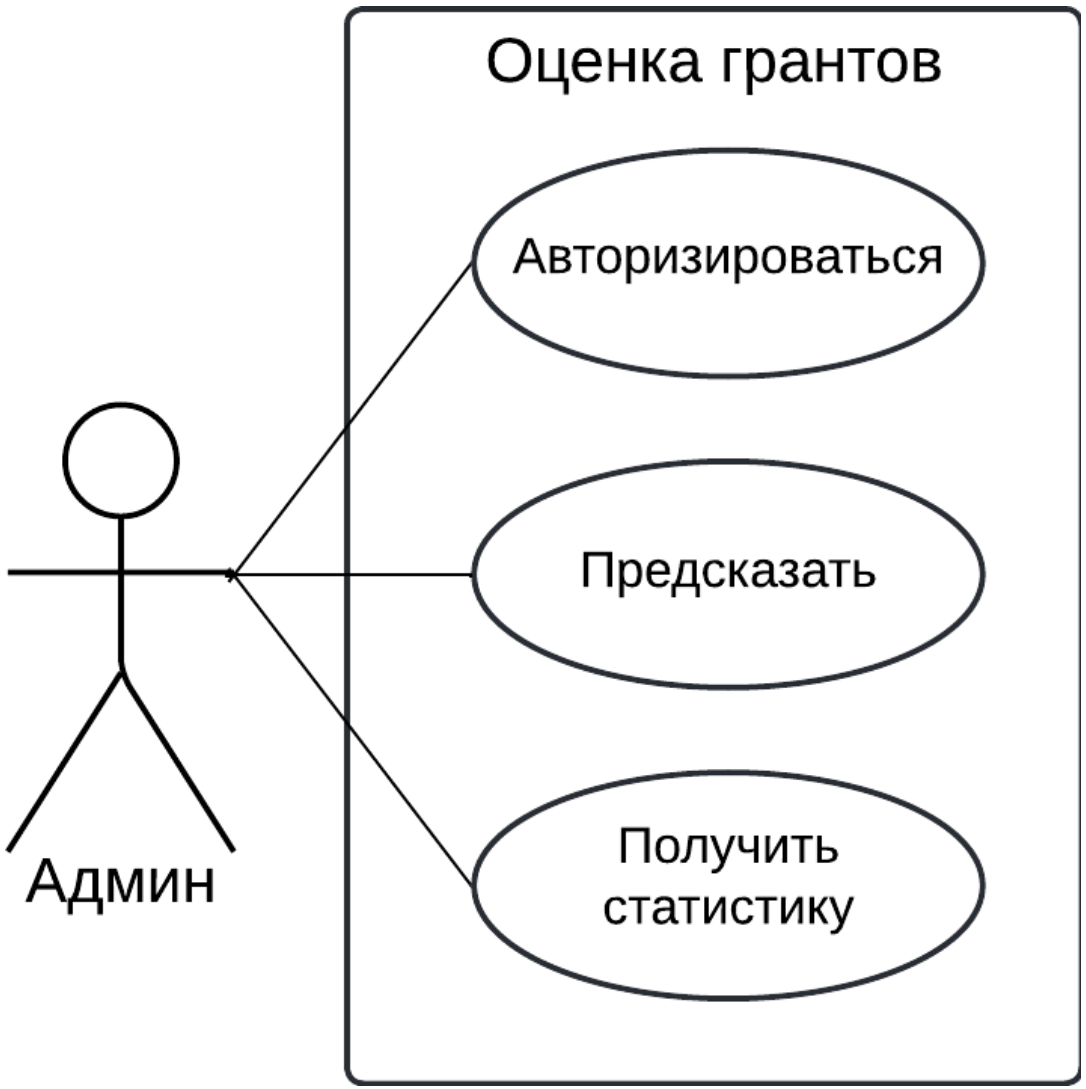
Схожесть текстов: модель

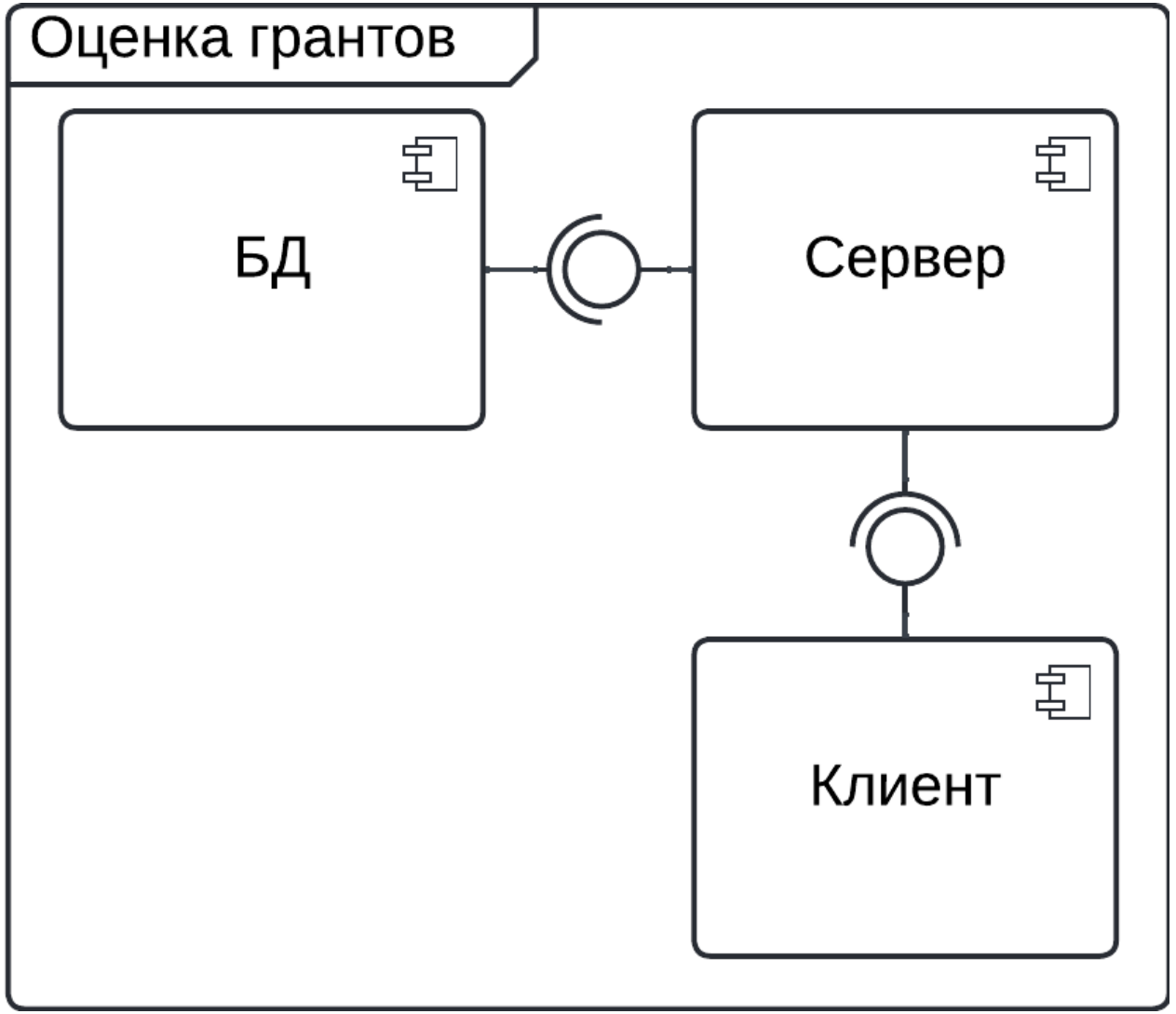
Модель: paraphrase-multilingual-MiniLM-L12-v2 [38]

№	Параметр	Значение
1	Base Model	Teacher: paraphrase-MiniLM-L12-v2; Student: microsoft/Multilingual-MiniLM-L12-H384
2	Max Sequence Length	256
3	Dimensions	384
4	Normalized Embeddings	false
5	Suitable Score Functions	cosine-similarity (util.cos_sim)
6	Size	420 MB
7	Pooling	Mean Pooling
8	Training Data	Multi-lingual model of paraphrase-multilingual-MiniLM-L12-v2, extended to 50+ languages.

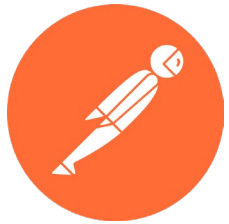
38. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Электронный ресурс] // arXiv.org, 2019. Дата обновления: 27.08.2019 г. URL: https://arxiv.org/abs/1908.10084 (дата обращения: 10.03.2024 г.).

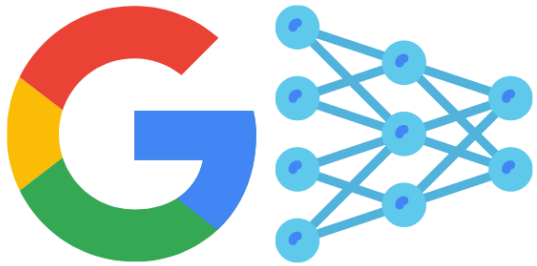
Проектирование



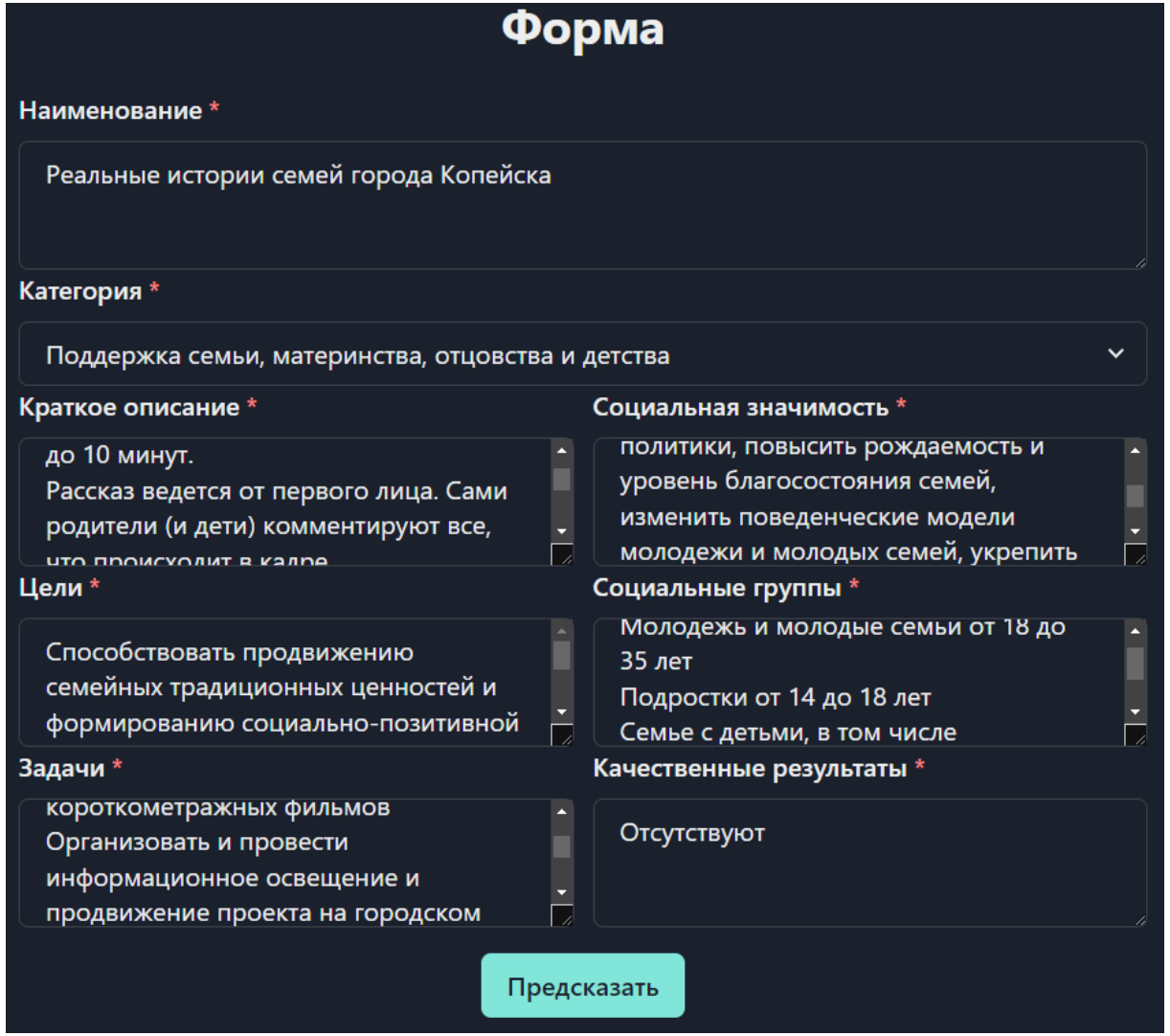


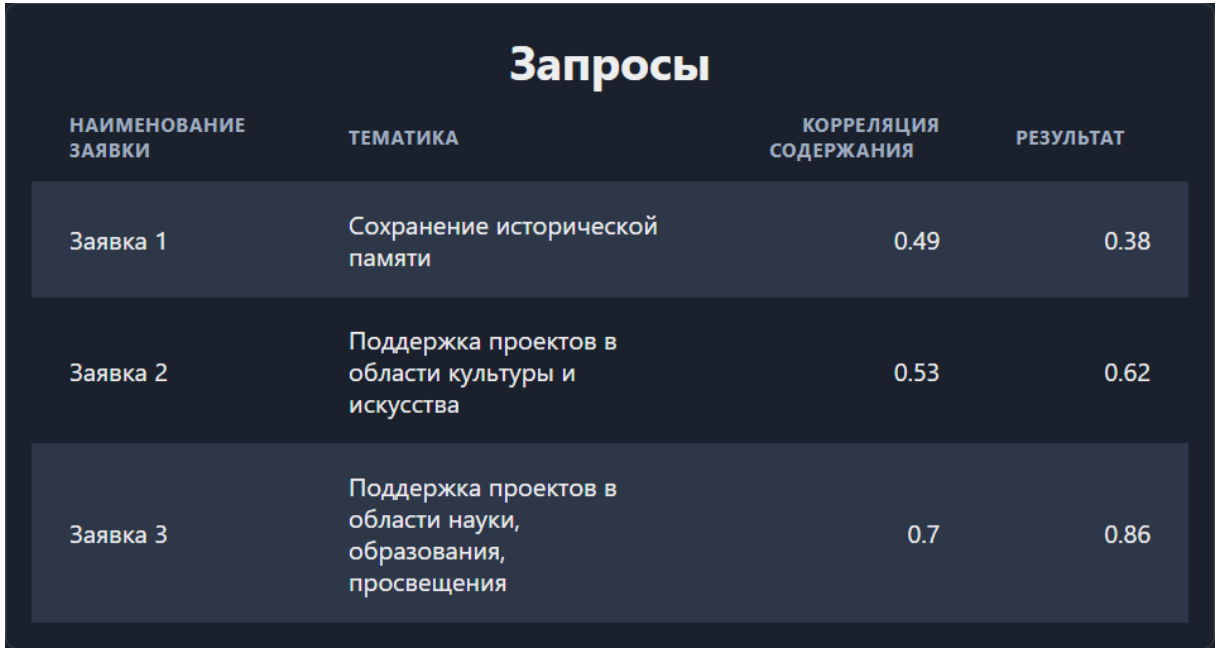
Используемые технологии и ПО

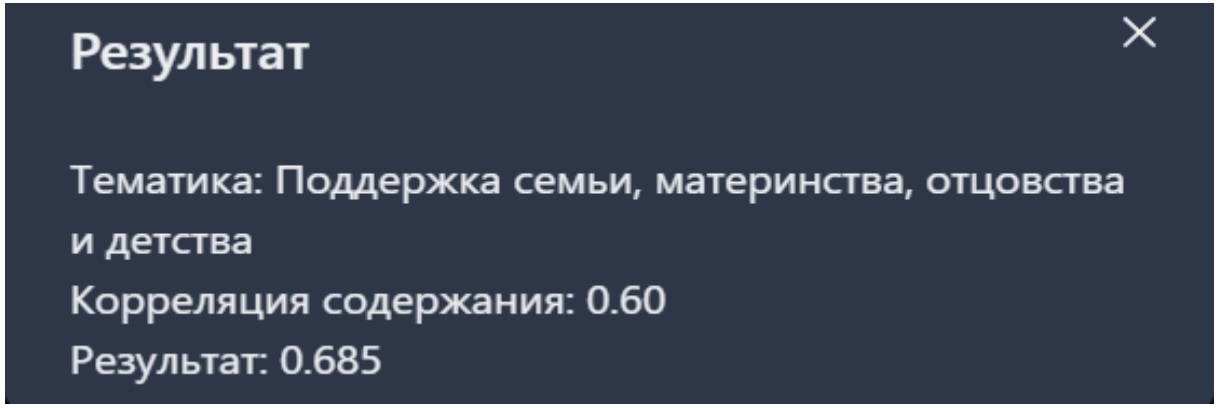




Интерфейс


Форма
Наименование *
Реальные истории семей города Копейска
Категория *
Поддержка семьи, материнства, отцовства и детства
Краткое описание *
до 10 минут.
Рассказ ведется от первого лица. Сами родители (и дети) комментируют все,
Социальная значимость *
политики, повысить рождаемость и уровень благосостояния семей, изменить поведенческие модели молодежи и молодых семей, укрепить
Цели *
Способствовать продвижению семейных традиционных ценностей и формированию социально-позитивной
Социальные группы *
Молодежь и молодые семьи от 18 до 35 лет
Подростки от 14 до 18 лет
Семье с детьми, в том числе
Задачи *
короткометражных фильмов
Организовать и провести информационное освещение и продвижение проекта на городском
Качественные результаты *
Отсутствуют
Предсказать


Запросы
НАИМЕНОВАНИЕ ЗАЯВКИ
ТЕМАТИКА
КОРРЕЛЯЦИЯ СОДЕРЖАНИЯ
РЕЗУЛЬТАТ
Заявка 1
Сохранение исторической памяти
0.49
0.38
Заявка 2
Поддержка проектов в области культуры и искусства
0.53
0.62
Заявка 3
Поддержка проектов в области науки, образования, просвещения
0.7
0.86


Результат
Тематика: Поддержка семьи, материнства, отцовства и детства
Корреляция содержания: 0.60
Результат: 0.685

Тестирование

№	Входные данные	Ожидаемый результат	Тест пройден
1	Аутентификация пользователя в системе	Пользователь переводится на главную страницу	Да
2	Переход пользователя в раздел «Предсказание»	Пользователь перешел в раздел «Предсказание»	Да
3	Пользователь ввел данные в форму и нажал кнопку «предсказать»	Пользователь получил предсказание	Да
4	Переход пользователя в раздел «Статистика»	Пользователь перешел в раздел «Статистика» и получил таблицу со статистикой	Да
5	Пользователь нажал кнопку «Выйти»	Пользователь вышел из системы	Да
6	Пользователь обновил страницу через 60 минут	Пользователь попал на страницу аутентификации	Да
7	Администратор разворачивает систему на машине с иной ОС	Система разворачивается и готова к работе	Да

Основные результаты

В результате выполнения выпускной квалификационной работы были решены следующие задачи:

– Изучена предметная область, проведен обзор литературы

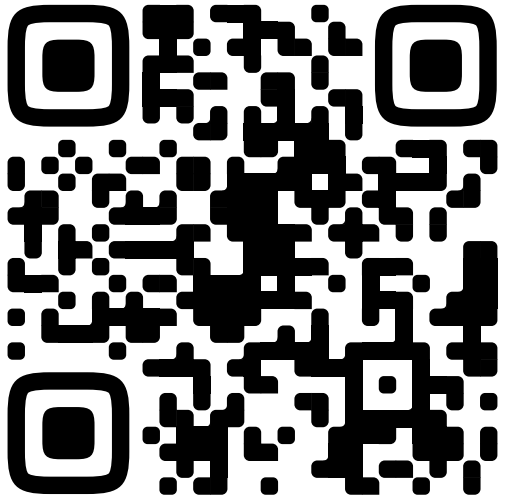
– Подготовлен набор данных для обучения, спроектированы архитектуры нейронных сетей и математическая модель

– Обучена нейронная сеть, оценены результаты обучения

– Реализована программная система для оценки заявок на основе нейросетевых технологий

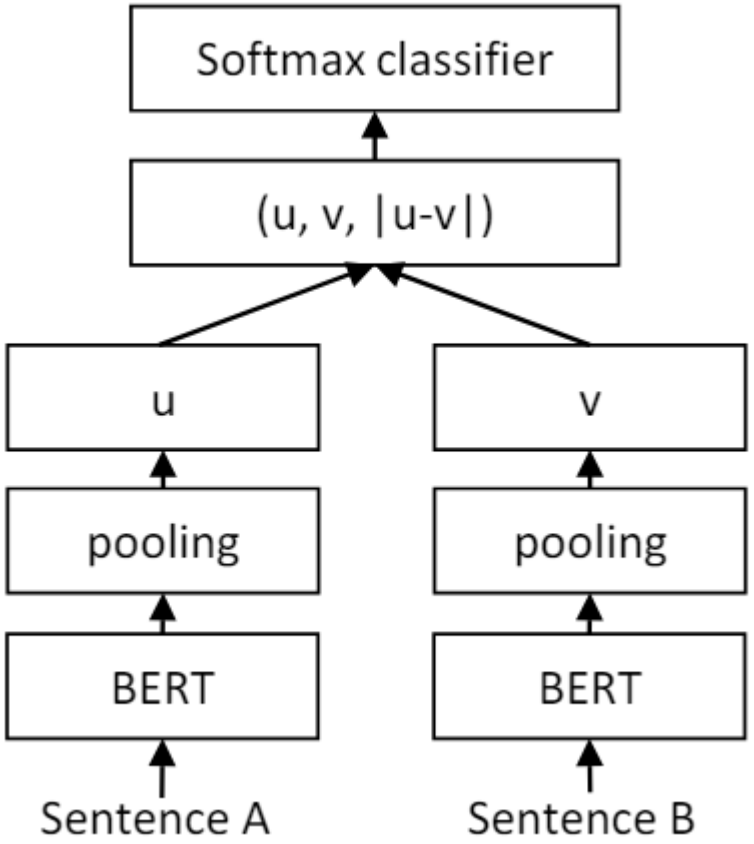
– Проведено тестирование программной системы

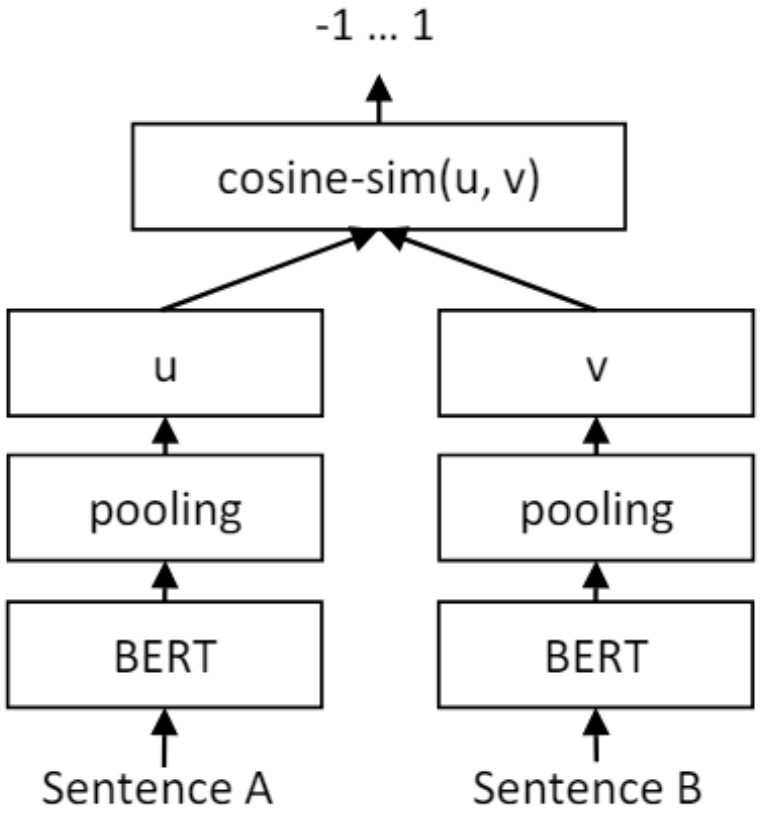
В дальнейшей разработке планируется обучить нейронную сеть для сравнения семантического сходства на основе размеченных данных заказчика и улучшить пользовательский интерфейс.



Семантическое сходство текста

Модель paraphrase-multilingual-MiniLM-L12-v2 [38]





38. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Электронный ресурс] // arXiv.org, 2019. Дата обновления: 27.08.2019 г. URL: https://arxiv.org/abs/1908.10084 (дата обращения: 10.03.2024 г.).